

**USING SEMANTIC SIMILARITY MEASURES IN THE BIOMEDICAL DOMAIN
FOR COMPUTING FUNCTIONAL SIMILARITY BETWEEN GENES BASED ON
GENE ONTOLOGY**

by

Elham Khabiri, B.Eng.

THESIS

**Presented to the Faculty of
The University of Houston Clear Lake
In Partial Fulfillment
of the Requirements
for the Degree**

MASTER OF SCIENCE

THE UNIVERSITY OF HOUSTON-CLEAR LAKE

November, 2007

USING SEMANTIC SIMILARITY MEASURES IN THE BIOMEDICAL DOMAIN
FOR COMPUTING FUNCTIONAL SIMILARITY BETWEEN GENES BASED ON
GENE ONTOLOGY

by

Elham Khabiri

APPROVED BY



Hisham Al Mubaid, Ph.D., Chair



Gary D. Boettcher, Ph.D., Committee Member



M. Bazlur Rashid, Ph.D., Committee Member



Sadegh Davari, Ph.D., Dean

ACKNOWLEDGEMENT

I would like to express my thanks and gratitude to all who supported and helped me through writing this thesis.

I am deeply indebted to my advisor Dr. Hisham Al-Mubaid for his unlimited help and supports throughout my research; I cannot describe how much I learned from him in bioinformatics, genomics, ontologies, and more. His motivation and suggestions were always with me through the completion of my thesis. Without his continuous help and support this work would not have been completed.

I wish to thank Dr. Gary Boetticher for being a great instructor during two semesters, which I was honored to learn a lot from his classes and beyond that. I wish to thank him also as my committee member for reading my thesis thoroughly and for giving me valuable advises during my research. I would like to thank Dr. Bazlur Rashid, as my committee member for his effort on reviewing my thesis.

My sincere thanks are due to Dr. Sadegh Davari for his guidance and support from the time I started my studies in the University of Houston - Clear Lake. He has been one of the greatest and the most helpful people I have ever had in my academic life.

I wish to thank my parents for their love, support and confidence throughout the twenty-five years. I owe them much of what I have become now. I would like to thank my Mom for all her continuous prayers and my Dad for his words which has always inspired me with hope and courage.

Many thanks to my patient and loving husband, Roozbeh, who has been always there for me. This work was impossible without his love, understanding, care and support.

I dedicate this work to my parents and my husband, to gratitude their love, patience and support during these years.

ABSTRACT

USING SEMANTIC SIMILARITY MEASURES IN THE BIOMEDICAL DOMAIN FOR COMPUTING FUNCTIONAL SIMILARITY BETWEEN GENES BASED ON GENE ONTOLOGY

**Elham Khabiri, M.S.
The University of Houston Clear Lake, 2007**

Thesis Chair: Hisham Al-Mubaid

The size and volumes of genomic data resulting from the various genome projects are extremely huge and continuously increasing in very high rates. Finding gene groups with similar functions is one of the most important tasks in bioinformatics. More specifically, computing the similarities between genes as numeric figures will have many benefits and applications in biomedical domain. We present novel techniques for measuring the functional similarity of genes using Gene Ontology (GO) annotations. GO is considered the most comprehensive resource of functional information on genes and gene products. The proposed methods are considered ontology-structure-based methods and rely strictly on ontology-structure features like depth and path length (PL) between GO nodes. We evaluated the proposed measures based on the correlation with gene sequence similarity

using Blast e-values. We conducted experiments with several genome annotation databases. The experimental results proved that the proposed similarity methods are fairly efficient in estimating the functional similarity between genes, gene products, and protein. Hence, ontology structure features can be used as good tools for determining the genes with similar functions within a genome.

TABLE OF CONTENTS

1. Introduction.....	1
1.1. Gene Similarity	2
1.1.1. Sequence Similarity	3
1.1.2. Semantic similarity	5
1.2. How this thesis is organized	9
2. Background and Related Work.....	12
2.1. Gene Ontology	12
2.2. GO Tools and Browsers.....	16
2.3. Distance between terms in GO	19
2.4. Similarity Measures	22
3. A Path Length Method for Gene Similarity Using GO Annotations.....	30
3.1. Path Length Calculation.....	31
3.2. Algorithm for Distance Measure	32
3.2.1. Distance between GO terms.....	32
3.2.2. Distance between genes	37
3.3. Comparing the results with Sequence Similarity.....	40
3.3.1. E-value	40
3.4. Experiments and Results.....	42
3.4.1. Distribution of Path Length	43

3.4.2.	Evaluation based on Correlation with Sequence Similarity	45
3.4.3.	Compare Average and Maxima methods.....	51
3.4.4.	Compare terms in Biological Process and Molecular Function ontologies	
	59	
3.5.	Conclusion	65
4.	A New GO structure Based Measure with Evaluation Using SGD Pathways	67
4.1.	Distance between GO terms.....	67
4.2.	Distance between genes	72
4.3.	Similarity between Genes	74
4.4.	Experimental Results and Evaluation	75
4.5.	Discussion and Conclusion.....	82
5.	Correlation between Depth and Path Length of GO Nodes with Gene Sequence Similarity.....	84
5.1.	Semantic Similarity between GO terms.....	84
5.2.	The Semantic Similarity of Genes	85
5.3.	Experiments and Results.....	86
5.3.1.	Dataset.....	86
5.3.2.	Distribution of $Simp_{LD}$	87
5.4.	Discussion and Conclusion.....	99
6.	Conclusion and Future Work.....	101
6.1.	Future Work.....	103
7.	References.....	105
Appendix A	111

Appendix B	117
-------------------------	------------

LIST OF TABLES

Table 1.1. Compare Sequential and Semantic Measures in High Sequentially Related Genes.....	8
Table 1.2. LCA for genes with multiple annotated GO terms	8
Table 1.3. Semantic Measures in Low Sequentially Related Genes.....	9
Table 3.1. The similarity matrix between two genes	38
Table 3.2. SGD genes with high sequence similarity with AAD10.....	41
Table 3.3. Comparing Group1 with Group2 genes	41
Table 3.4. SGD genes with no similarities with AAD10.....	42
Table 3.5. Example from SGD-LSS gene pairs	46
Table 4.1. Path Length (PL) and number of minimum path (nmp) between the GO-terms for InR and Ror genes from FlyBase organism	69
Table 4.2. PL and nmp values between GO terms of two SGD genes (ABF1 and IFH1).	74
Table 4.3. Comparison of our result with Resnik's result in two pathways from SGD...	77
Table 4.4. Similarity values among genes in tryptophan degradation pathway based on our algorithm.....	79
Table 4.5. Similarity values among genes in tryptophan degradation pathway based Wang et al.'s measure [61].	79
Table 4.6. Three SGD genes with their annotation by GO terms.	82

Table 5.1. Sample of the output of application	98
Table 5.2. Path Length between Ror and InR GO-terms	99
Table 5.3. Depth and PL between Ror and InR GO-terms	99
Table 0.1. Human-Yeast-IO dataset.....	118
Table 0.2. SGD HSS dataset	119
Table 0.3. FlyBase NSS dataset	120

LIST OF FIGURES

Figure 1.1. Sequence Alignment.....	4
Figure 2.1. Overview of Gene Ontology.....	14
Figure 2.2. True path rule: The two children are more specified and have smaller association value than their parent.....	15
Figure 2.3. A tree view of some GO terms with is_a relationships between them (Picture is from Amigo browser [7]).....	16
Figure 2.4. Each node in GO could have more than one parent. The picture is from GOLEM software [50].....	17
Figure 2.5. Genes associated with term GO:0008188 in Amigo Browser.....	18
Figure 2.6. Sample of Amigo Browser output	19
Figure 2.7. XML format of Gene Ontology.....	20
Figure 2.8. Example of a tree structure	24
Figure 3.1. GO is a kind of DAG.....	33
Figure 3.2. Stage 1 of the algorithm.....	34
Figure 3.3. Stage 2 of the algorithm.....	34
Figure 3.4. Stage 3 of the algorithm.....	34
Figure 3.5. Stage 4 of the algorithm.....	35
Figure 3.6. Stage 5 of the algorithm.....	35
Figure 3.7. Stage 6 of the algorithm.....	35

Figure 3.8. Reach the first common ancestor from two target nodes.....	36
Figure 3.9. Source node(target node) of each node in the link list	36
Figure 3.10. The Path Length Calculator application snapshot	37
Figure 3.11. Relationship between path length and bit score	42
Figure 3.12. Distribution of path length among 1000 gene pairs randomly selected from SGD.....	44
Figure 3.13. Distribution of path length among 500 gene pairs randomly selected from FlyBase	45
Figure 3.14. Distribution of path length between gene pairs in Dataset 1	47
Figure 3.15. Distribution of path length between gene pairs in Dataset 2 from SGD	48
Figure 3.16. Distribution of path length between gene pairs in Dataset 3	49
Figure 3.17. Distribution of path length between gene pairs in Dataset 4 from FlyBase	50
Figure 3.18. Comparison between PL and Maxima measure in HSS FlyBase dataset....	51
Figure 3.19. Comparison between PL and Maxima measure in NSS FlyBase dataset....	52
Figure 3.20. Comparison between PL and Maxima measure in HSS SGD dataset.....	53
Figure 3.21. Comparison between PL and Maxima measure in LSS SGD dataset	54
Figure 3.22. Comparison between PL and Maxima measure in NSS SGD dataset.....	55
Figure 3.23. Comparison between PL and Maxima measure in IO Human-Yeast dataset	56
Figure 3.24. Comparison between PL and Maxima measure in HSS Human-Yeast dataset	57
Figure 3.25. Comparison between PL and Maxima measure in LSS Human-Yeast dataset	58

Figure 3.26. Comparison between PL and Maxima measure in NSS Human-Yeast dataset	59
Figure 3.27. Comparison PL between BP and MF in HSS FlyBase dataset.....	60
Figure 3.28. Comparison PL between BP and MF in NSS FlyBase dataset.....	61
Figure 3.29. Distribution of PL in Human-Yeast dataset using BP terms	62
Figure 3.30. Distribution of PL in Human-Yeast dataset using MF terms	62
Figure 3.31. MF vs. BP in Human-Yeast IO dataset	63
Figure 3.32. MF vs. BP in Human-Yeast HSS dataset	64
Figure 3.33. MF vs. BP in Human-Yeast LSS dataset.....	64
Figure 3.34. MF vs. BP in Human-Yeast NSS dataset	65
Figure 4.1. A graph to represent multiple paths in GO.....	68
Figure 4.2. Part of the GO to illustrate the paths between two GO terms 0042626 and 0004129.....	71
Figure 4.3. Clustering genes in tryptophan degradation pathway based on our algorithm	80
Figure 4.4. Clustering genes in tryptophan degradation pathway based on [61].....	81
Figure 5.1. Distribution of Sim_{PLD} value between gene pairs in FlyBase dataset	88
Figure 5.2. Distribution of Sim_{PLD} value between gene pairs in SGD dataset	89
Figure 5.3. Distribution of Sim_{PLD} value between gene pairs in Human-Yeast dataset ..	90
Figure 5.4. Sim_{PLD} in FlyBase HSS dataset	91
Figure 5.5. Sim_{PLD} in FlyBase NSS dataset	92
Figure 5.6. Sim_{PLD} in SGD HSS dataset	93
Figure 5.7. Sim_{PLD} in SGD LSS dataset	93

Figure 5.8. Sim_{PLD} in SGD NSS dataset	94
Figure 5.9. Sim_{PLD} in Human-Yeast IO dataset	95
Figure 5.10. Sim_{PLD} in Human-Yeast HSS dataset	95
Figure 5.11. Sim_{PLD} in Human-Yeast LSS dataset	96
Figure 5.12. Sim_{PLD} in Human-Yeast NSS dataset	96
Figure 5.13. Sample of running of the program.....	97

1. INTRODUCTION

Computing the functional similarity between genes and proteins is an important and necessary task in the bioinformatics and biomedical fields. By comparing similarities between genes and proteins with known functions to those with unknown functions, the functions of the unknown genes and proteins can be determined to certain accuracy [54]. Also it is useful to measure the differences between genes and proteins in different organisms. As an example, one can compare the proteins in yeast with the proteins in human and find those proteins in yeast that have the least biological and functional similarities with those in human. This is an approach for finding drugs and drug targets for human [54]. Thus, those proteins with biological processes or molecular functions, that are absent in human proteins, are considered as potential drug targets in biomedical domain [54].

In general, genes and gene products are functionally similar if they have comparable molecular functions and are involved in similar biological processes [54]. These gene products are not necessarily evolved from a common ancestor, and therefore, do not necessarily show sequence similarity. In this research we explore a number of techniques for measuring the similarity between terms in Gene Ontology (GO). Gene ontology [9] is

a comprehensive and controlled ontology to describe the functional and biological features of genes independent of the organism. We also propose new measures of functional similarity between genes using GO. The proposed measures have been implemented and evaluated with a large number of experiments using multiple sets of annotation databases. We have evaluated our data using three datasets that are:

- Dataset from SGD (Saccharomyces Genome Database)
- Dataset from FlyBase (Database for Fruit Fly)
- Dataset of gene pairs from Human and Yeast

Fruit Fly and Saccharomyces are considered as model organisms. A model organism is a species that is appropriate to understand particular biological events in more complicated organisms, by providing the insight for workings of them [21]. For example, they are widely used to explore potential causes and treatments for human disease when experimentation on humans would be unfeasible or unethical [21]. Some of the model organisms are used for human like mice and fruit fly and some are used for studying plant sciences like *Arabidopsis thaliana* [21].

1.1. Gene Similarity

Finding the similarity between genes and proteins can be done by several computational methods and from different data sources. For example, gene expression data, statistical computation on biological literature, sequential similarity, and semantic similarity are different information sources for measuring the similarity between genes and proteins [10, 32, 51, 54, 66, 69, 70]. For example, in [4], Al-Mubaid and Nguyen investigated the

effectiveness of using *Medline* corpus as the information source for measuring the semantic similarity in the biomedical domain [4]. In this thesis we focus on (1) the semantic similarity and (2) the sequence similarity between genes. In general, we compute the similarity between genes based on the similarity of their GO annotation terms. The general problem of measuring gene functional similarity using GO annotations with semantic similarity measures can be defined as follows: Define a genome annotation set (e.g. SGD, FlyBase) to be a set of genes of one species/organism with GO functional annotations for each gene in the set. That is, every gene in the set is associated with one or more GO terms.

Let $G = \{G_1, G_2, \dots, G_n\}$ be the set of all genome annotations {in BLAST, UniProt, geneontology,...etc.}.

Our goal is to define a general semantic similarity function $S(g_1, g_2, G)$ such that if g_1 is (per blast-sequence-similarity, for example) closer to g_2 than to g'_2 then $S(g_1, g_2) > S(g_1, g'_2)$. Since such a similarity function is defined on all genes having GO annotations, it provides us a unified semantic similarity measure between genes regardless of the organism.

1.1.1. Sequence Similarity

DNA and proteins sequences can be considered as identifiers for genes and proteins. To look at them from the computer science side, they are sequences of alphabets that may have similarities in regions. They can be compared globally means all the sequence is

considered for similarity score or locally means that only specific regions of them are compared to each other. We call the first one global alignment and the latter local alignment. Here is a sample of aligning the two sequences.

```
GAATTCAG
| | | |
GGA-TC-G
```

Figure 1.1. Sequence Alignment

They are some score functions that give positive score to the letters that match and negative scores to those who do not. For example one function score may give the sequence score of +1 to the matched letters and -1 to mismatched ones. And -2 could be given to the gaps (indels) which are inserted to the sequence for maximizing the alignment score [68]. They are different methods of calculating the similarity score for two or more sequences. One of them is BLAST [5]. The BLAST algorithm has the best method that keeps a balance between speed of calculation and sensitiveness in sequence relationships [68]. Instead of relying on global alignments that is commonly used in multiple sequence alignment programs, BLAST emphasizes regions of local alignment to detect relationships among sequences that have regions of similarity (Altschul et al., 1990). The input of BLAST tool is *FASTA format* of the sequences of the genes or proteins. FASTA format is a text-based format for representing either nucleic acid sequences or protein sequences, in which base pairs or protein residues are represented using single-letter codes.

Since most of the bioinformatics data is in the form of sequences, the most accurate way of comparing the genes and proteins is by sequence similarities. The homologous relationship between proteins could be found by sequence comparisons, but not all of the similarities are based on homologies [54]. Based on sequence comparison, proteins of unknown function are assigned to characterized protein families, generating testable hypotheses of their molecular function. However, this established annotation approach has several limitations such as; up to 30% of the function annotations made through sequence similarity searches might be erroneous [16] [17]. 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 [54].

The other problem is that there is no simple relationship between sequence similarity and function, but some general trends have been observed [54]. One other drawback for the sequence notation is that, it is not readable and understandable by human. Semantic measures on the other hand uses the resource data in scientific natural language as text which is human readable and understandable [4, 32].

1.1.2. Semantic similarity

One of the common ways of finding the similarities among genes is by computing the semantic similarities between GO functional annotations of the genes [26, 31, 32, 47, 51, 54, 61]. The resource data used in these kinds of measures are in scientific natural

language format which makes it human readable and understandable. The problem with it is they are not easy to interpret computationally [32]. These approaches use ontology (e.g. Gene Ontology) as the primary information source, and can be divided into two categories: Ontology-Structured-Based and Information-Based measures.

Ontology-Structure-Based Measures

The ontology-structured based measures use the ontology structure features such as path length between nodes (in the ontology), depth of nodes in the ontology tree, and the number of minimum paths between nodes, for computing the semantic similarity between two terms in a given ontology. For example, the shortest path length between two terms (two nodes) in a given ontology can be considered as a good indicator (or metric) of the (relative) similarity between these two terms. Suppose that $PL(t_1, t_2)$ is the shortest path length between the two terms t_1 and t_2 in a given ontology O_x then $PL(t_1, t_2) > PL(t_3, t_4)$ implies that the terms (t_1, t_2) have more similarity than the pair (t_3, t_4) according to ontology O_x . In this thesis we have investigated the semantic similarity that is based on the structure of the Gene Ontology.

Information-Content Based Measures

The information-content-based measures use the information content (IC) of gene terms in computing the semantic similarity. Information Content can be defined as the

frequency of use of a term that can be computed from text corpora or estimated from the ontology (i.e. Gene Ontology) [48].

As an example here we compare the two information based measures Resnik [48] and Lin [30] for 30 random gene pairs selected from SGD [53].

In Resnik measure [48] the similarity between the two terms is calculated by the information content (frequency of use) of the common ancestors. Thus, the semantic similarity between two terms in an ontology is:

$$\text{sim}_{\text{Resnik}}(c_1, c_2) = -\log P(c), c \in S(c_1, c_2)$$

$S(c_1, c_2)$ is the set of common ancestors of terms c_1 and c_2 .

Lin [30] defines the similarity between two terms as the ratio of the LCA occurrence probability of two terms to the information needed to fully describe the two terms individually. The following equation reflects this idea.

$$\text{sim}_{\text{Lin}}(c_1, c_2) = \max \left(\frac{2 \cdot \log P(c)}{\log P(c_1) + \log P(c_2)} \right), c \in S(c_1, c_2)$$

$S(c_1, c_2)$ again is the set of common ancestors of terms c_1 and c_2 .

Gene1	Gene2	E-Value	Bit Score	Resnik	Lin
AAC1	AAC1	4.6e-145	1412	3.9049	1
AAC1	PET9	1.7e-115	1133	3.9049	1
AAC1	AAC3	3.7e-111	1092	3.9049	1
AAC1	YPR011C	3.1e-20	234	1.2790	0.3958
AAC1	LEU5	1.1e-14	171	1.2790	0.4096

AAC1	OAC1	9.9e-13	181	2.1438	0.4897
AAC1	YEA6	3.70E-11	169	1.2790	0.4934
AAC1	CTP1	5.30E-09	150	2.1438	0.7668
AAC1	ODC1	1.80E-07	100	1.2790	0.7073
AAC1	AGC1	2.40E-07	100	1.2790	0.5101

Table 1.1. Compare Sequential and Semantic Measures in High Sequentially Related Genes

Gene1	Gene2	GO Gene2	Occurrence(out of 184810)	LCA GO	LCA Occurrence
AAC1	AAC1	GO:0005471	23	GO:0005471	23
AAC1	PET9	GO:0005471	23	GO:0005471	23
AAC1	AAC3	GO:0005471	23	GO:0005471	23
AAC1	YPR011C	GO:0005215	9721	GO:0005215	9721
AAC1	LEU5	GO:0015228	2	GO:0005215	9721
AAC1	OAC1	GO:0008271, GO:0000227	21,2	GO:0015291	1327
AAC1	YEA6	GO:0051724, GO:0005215	3,9721	GO:0005215	9721
AAC1	CTP1	GO:0005371	9	GO:0015291	1327
AAC1	ODC1	GO:0005342, GO:0005478	850,224	GO:0005215	9721
AAC1	AGC1	GO:0015183, GO:0005313	9,34	GO:0005215	9721

Table 1.2. LCA for genes with multiple annotated GO terms

As you see in the Table 1.2 some genes are related to more than one GO terms. Lins and Resnik both suggest picking up the one with the maximum occurrence of Least Common Ancestors. These terms are marked as bold in the table. Here the scores calculated from Resnik and Lins which are semantic similarity measures are compared to the sequential scores called Bit Score and E-value. Bit Score is the score that two sequences of genes obtain for their structural similarities and the E-Value represents the error or the differences between the genes

In the following table the Resnik and Lins measures are calculated for those genes that have no sequential similarities with the selected gene (AAC1). These genes are selected from the genes that were not appearing among those that have sequential similarity with the selected gene.

Gene1	Gene2	Resnik	Lins
AAC1	15S_RRNA	0.1293	0.0816
AAC1	AAD10	0.1293	0.0476
AAC1	YPL206C	0.1293	0.0526
AAC1	YPL278C	0.1293	0.3642
AAC1	RIO1	0.1293	0.0860
AAC1	RIX1	0.1293	0.3642
AAC1	SCS7	0.1293	0.0442
AAC1	SSO1	1.2790	0.4934
AAC1	YPR158W	0.1293	0.3642
AAC1	tC(GCA)PI	0.1293	0.0668

Table 1.3. Semantic Measures in Low Sequentially Related Genes

1.2. How this thesis is organized

This chapter provides an introduction and overview to the task of similarity between genes and proteins using gene sequence data or gene annotation data from GO. Chapter 2 gives a review of the background about the gene ontology and the tools related to than in addition to the related work and the existing measures of gene similarity. In chapter 3, we propose novel measure called *PL* for measuring the functional similarity between genes using the GO annotations. One of the methods is based on calculating the simple path length (PL) between GO annotation terms of the genes. We evaluated our method with a series of experiments based on the correlation between our method and gene sequence similarity using Blast e-values. The experimental results proved that our

approach 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 three genome annotation datasets: SGD [53], FlyBase [67] and a Human-Yeast dataset of proteins[54]. 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.

In chapter 4 we have proposed another method of measuring the semantic similarity of GO terms based on path length and the number of minimum paths between GO terms in the GO graph. This method distinguishes between two types of paths and assigns different weights to determine the contributions of number of paths in the semantic similarity between the GO terms. To assess the similarity between two GO terms, our method considers all the possible paths between the two terms rather than considering only the distance to their least common ancestor LCA or the IC of their LCA [48], [23], [30], [61] . In the evaluation, we measured the semantic similarity of SGD (Saccharomyces Genome Database) genes from various SDG pathways (obtained from <http://www.yeastgenome.org>) and compared our results with two of the leading measures (Resnik [48] and Wang et al. [61]). In chapter 5 we extend our *PL* measure and came up to a new measure called *Sim_{PLD}* that uses the depth of least common ancestor of two gene series of related term and the path length between them [25]. We used the average of all *Sim_{PLD}* for the terms annotated for each gene. The method is evaluated by a series of

experiments based on the correlation between *Sim_{PLD}* and gene sequence similarity using Blast e-values.

2. BACKGROUND AND RELATED WORK

This chapter gives an introduction on the gene ontology which is one of the most comprehensive projects done in bioinformatics. It will also discuss about the tools and browsers available to search and navigate the terms in the gene ontology. Then the similarity measures that are proposed in different domains will be explained.

2.1. Gene Ontology

The Gene Ontology, created in 2000 by Gene Ontology (GO) Consortium [9], is an ontology which shows the functional and biological terms (*annotation terms*) related to genes and proteins in a hierarchical and structured way. Gene Ontology consists of a set of controlled vocabularies to describe the biology of genes in any organism [9]. GO annotations capture the available functional information of a gene or protein and can be used as a basis for defining a measure of functional similarity between genes. Besides the bioinformatics resources that hold data in the form of sequences, these data has represented as scientific natural language which is easier to be modeled and is more readable to human [32]. Gene Ontology has provided more accessible representation of the data related to the genes [47]. It is a dynamic evolving project of the Gene Ontology (GO) Consortium in which different sections of the ontology are expanded or reorganized

as more biological information becomes available. Therefore, GO project is a collaborative effort to address the need for consistent descriptions of genes in different databases. The project is collaboration between 35 model organism databases. Among them FlyBase (*Drosophila Melanogaster*), the *Saccharomyces* Genome Database (SGD) and the Mouse Genome Database (MGD), were the first groups of databases started the collaboration and after that other databases have joined them [9]. The ontology is represented as a network, directed acyclic graph (DAG), in which terms may have multiple parents and multiple relationships to their parents. In addition, each term inherits all the relationships of its parent(s). GO consists of three ontologies that describe the **molecular function** of a gene, the **biological process** in which the gene participates, and the **cellular component** where the gene can be found; see Figure 2.1. Figure 2.1 shows an excerpt of the gene ontology as it appears in the Amigo browser [7]. Each one of these three ontologies (molecular function, biological process, and cellular component) can be viewed as a root node and has children. For example, as shown in Figure 2.1, the node “molecular function” with the GO id number of GO:0003674 and has the following children: “GO:0016209 : antioxidant activity”, “GO:0015457 : auxiliary transport protein activity”, “GO:0005488 : binding”, “GO:0003824 : catalytic activity”, “GO:0060089 : molecular transducer activity”, “GO:0004871 : signal transducer activity”. The “signal transducer activity” is also the parent of “GO:0004872 : receptor activity” and other children. If we continue to see the next children we see “GO:0008188 : neuropeptide receptor activity” which is the child of “GO:0030594 : neurotransmitter receptor activity”. This term is the last node so-called a leaf and there is

no other term that can be categorized under this term. It has the smallest association value (the value inside the bracket) in compare with its parents and ancestors.

```

└─ all : all [221913]
  └─ GO:0008150 : biological_process [142636]
    └─ GO:0005575 : cellular_component [155182]
      └─ GO:0003674 : molecular_function [145259]
        └─ GO:0016209 : antioxidant activity [550]
          └─ GO:0015457 : auxiliary transport protein activity [165]
            └─ GO:0005488 : binding [43761]
              └─ GO:0003824 : catalytic activity [46436]
                └─ GO:0030188 : chaperone regulator activity [62]
                  └─ GO:0042056 : chemoattractant activity [13]
                    └─ GO:0045499 : chemorepellant activity [8]
                      └─ GO:0030234 : enzyme regulator activity [2471]
                        └─ GO:0016530 : metallochaperone activity [39]
                          └─ GO:0060089 : molecular transducer activity [7766]
                            └─ GO:0004871 : signal transducer activity [7766]
                              └─ GO:0004872 : receptor activity [5726]
                                └─ GO:0030594 : neurotransmitter receptor activity [287]
                                  └─ GO:0008188 : neuropeptide receptor activity [176]
                                    └─ GO:0001653 : peptide receptor activity [323]
                                      └─ GO:0008528 : peptide receptor activity, G-protein coupled [315]
                                        └─ GO:0008188 : neuropeptide receptor activity [176]
                                          └─ GO:0004888 : transmembrane receptor activity [4885]
                                            └─ GO:0004930 : G-protein coupled receptor activity [3889]
                                              └─ GO:0001584 : rhodopsin-like receptor activity [3434]
                                                └─ GO:0008528 : peptide receptor activity, G-protein coupled [315]

```

Figure 2.1. Overview of Gene Ontology

Each node is specified by a GO id number which is a unique identifier for the GO terms in the gene ontology, a name, and the number of genes associations (i.e. the number of genes that are annotated with this term in gene ontology) shown inside the brackets. The more specific term, the smaller number of gene is associated with it. Therefore a big number of associations mean that the term is a general term. Each node's association

number is the summation of the association number of its children plus the association number of itself. For example in Figure 2.2 we have “GO:0000146 *microfilament motor activity*” (association number of 63) with two children of “GO:0060001: *minus-end directed microfilament motor activity*”(association number of 2) and “GO:0060002 *plus-end directed microfilament motor activity*” (association number of 2) . The small value of the children shows the specificity of the two terms. Whereas the term “GO:0000146 *microfilament motor activity*” have larger number than its children which is compatible with “*true path rule*” that states that if a term describes a gene then all its parents must also apply to that gene [9]).

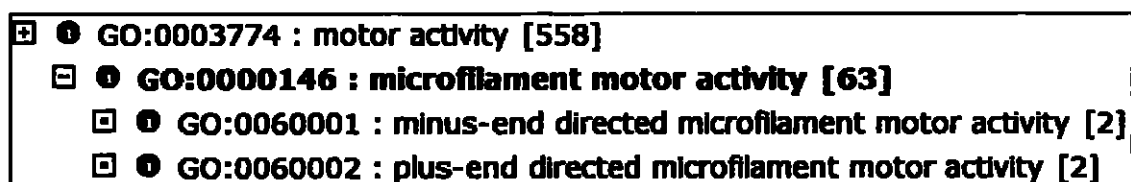


Figure 2.2. True path rule: The two children are more specified and have smaller association value than their parent

In GO, the terms are linked by two kinds of relationships that are *is_a* and *part_of*. The *is_a* relationship has the meaning of being a subclass. The *part_of* relationship means that if A is *part_of* B then whenever B exists A exists as a part of B. But A does not depend on B. Figure 2.3 shows some GO terms with *is-a* relationships between them in Gene Ontology.

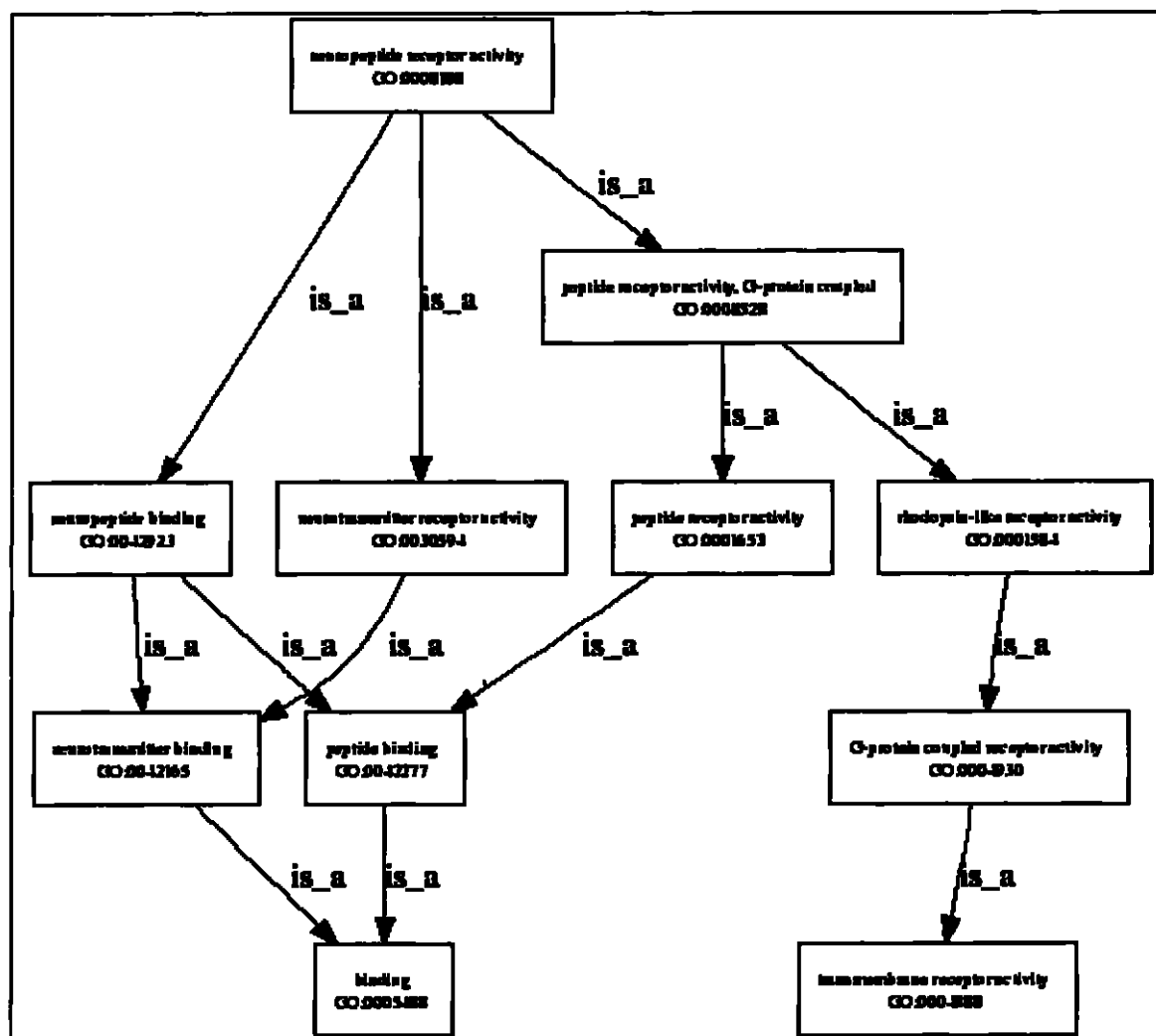


Figure 2.3. A tree view of some GO terms with is_a relationships between them (Picture is from Amigo browser [7])

2.2. GO Tools and Browsers

There are several software tools to navigate and browse through the Gene Ontology to shows the position of the terms within the GO hierarchy. In this section we take a look and review some of these tools.

- GOLEM [50] is an interactive graphic visualization tool for gene ontology that can be used for navigation and analysis of GO on the web. Users can also load annotation for various organisms to search particular genes. GOLEM is implemented in Java and both applet and web version of it is available. Figure 2.4 shows how this software looks like.

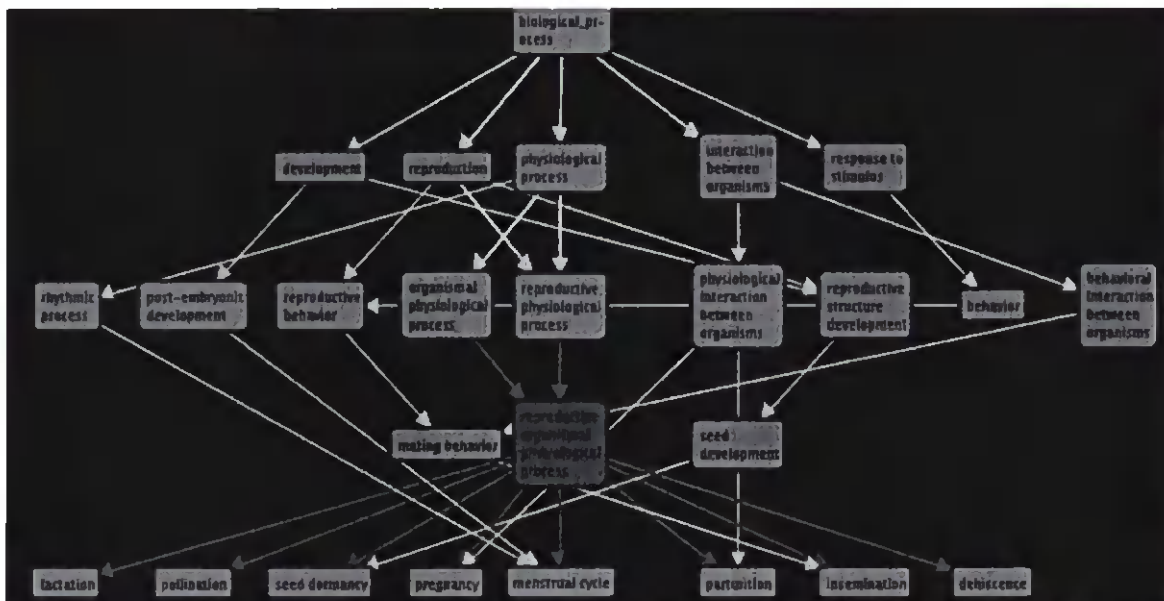


Figure 2.4. Each node in GO could have more than one parent. The picture is from GOLEM software [50]

- Amigo is a browser for gene ontology data that is used for browsing and searching the gene ontology [7]. Users can search for genes to see the terms associated with them. They can see a terms' position in the GO by using the Amigo interface. Amigo can be used to view all the genes associated with a GO term. The new added feature is BLAST search, which is useful to find the genes

that have the highest sequence similarity with the specified gene. Amigo uses the MySQL database. Figure 2.5 shows the genes associated with the term GO:0008188 in Amigo browser. By pushing the BLAST button we can have the FASTA format of the genes in addition to the genes that are sequentially similar to that gene based on their p-value.





Name/Symbol		Information	Evidence	Reference
<u>AKR2</u> allatostatin C receptor 2		gene from <i>Drosophila melanogaster</i>	<u>ISS</u>	<u>PMID:11381038</u>
<u>Alstr</u> Allatostatin Receptor		gene from <i>Drosophila melanogaster</i>	<u>ISS</u>	<u>PMID:11381038</u>
<u>AR-2</u> Allatostatin Receptor 2		gene from <i>Drosophila melanogaster</i>	<u>ISS</u>	<u>PMID:11381038</u>
<u>capaR</u> capa receptor		gene from <i>Drosophila melanogaster</i>	<u>ISS</u> <u>IDA</u>	<u>PMID:11381038</u> <u>PMID:12177421</u>

Figure 2.5. Genes associated with term GO:0008188 in Amigo Browser

Here is an example of *FASTA format* for gene *TVFV2E*. *FASTA format* starts with a single line description and the lines of sequence data comes after that. The ">" symbol at the beginning of the line distinguishes the description from the sequence data. See below:

```
>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLNGSYSENRT
QIWQKHRTSNDSALILLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHSQKYNLRLRQAWC
HFPSNWKGAWKEVKEEIVNLPKERYRGTDNPKRIFFQRQWGPETANLWFNCHGEFFYCK
MDWFLNYLNNLTVDADHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKK
TYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLS PQIESIWAAELDRYKLVEITPIGF
```

```
APTEVRRYTGGHERQKRVPFVXXXXXXXXXXXXXXXXXXXXXXXXXVQSQHLLAGILQQQKNL
LAAVEAQQQMLKLTIWGVK
```

There are lots of other navigation and analysis tools available on gene ontology website geneontology.org. The mentioned software tools are the ones used in this thesis.

```
⊕ all : all [219358]
  ⊖ GO:0003674 : molecular_function [143593]
    ⊕ GO:0016209 : antioxidant activity [540]
    ⊖ GO:0015457 : auxiliary transport protein activity [164]
      ⊖ GO:0016249 : channel localizer activity [1]
      ⊖ GO:0016247 : channel regulator activity [153]
        ⊖ GO:0005246 : calcium channel regulator activity [42]
          ⊖ GO:0019855 : calcium channel inhibitor activity [11]
```

Figure 2.6. Sample of Amigo Browser output

2.3. Distance between terms in GO

In Gene Ontology finding the number of the edges between two terms has not been automated by any software. In this thesis we have implemented a program that can quantify the distance between the terms, using the XML format of the Gene Ontology. The XML file is freely available and downloadable from www.geneontology.org.

```

<?xml version="1.0" encoding="UTF-8"?>
<rdf>
  <term about="http://www.geneontology.org/go#all">
    <accession>all</accession>
    <name>all</name>
    <definition>This term is the most general term possible</definition>
  </term>
  <term about="http://www.geneontology.org/go#0015488">
    <accession>0015488</accession>
    <name>glucuronide permease activity</name>
    <synonym>glucuronoside permease activity</synonym>
    <definition>Catalysis of the reaction: glucuronide(out)
+ monovalent cation(out) = glucuronide(in) + monovalent cation(in).
</definition>
    <is_a resource="http://www.geneontology.org/go#0015164" />
    <is_a resource="http://www.geneontology.org/go#0015486" />
    <dbxref parseType="Resource">
      <database_symbol>TC</database_symbol>
      <reference>2.A.2.3.1</reference>
    </dbxref>
  </term>
</rdf>

```

Figure 2.7. XML format of Gene Ontology

In this thesis we have calculated the distances between genes and proteins from different genomes [26]. The terms associated with each gene and protein is extracted from a database related to that genome. The process of assigning GO terms to genes is called *annotation*. The database provides us with terms that the genes are annotated with and the references that associated the terms to the genes. It also indicates the kind of *evidence code* available to support the annotation. For every evidence code, a curator judges about the quality of the evidence. Therefore the terms that have the evidence code of TAS (Traceable Author Statement) is completely different in terms of quality from those that have the evidence code of NR (Not Recorded). Some of other evidence codes are NAS: Non-traceable Author Statement, ISS: Inferred from Sequence or Structural Similarity,

IEA: Inferred from Electronic Annotation. More detail about the evidence code can be found in geneontology.org.

Each of these databases has downloadable files that contain all these associations. Some of the genomes that have their annotations available are:

- **SGD: This is a scientific database related to the genes of the yeast *Saccharomyces cerevisiae*, which is commonly known as baker's or budding yeast. It contains 6476 annotated genes in gene ontology [53].**
- **FlyBase: This database contains the molecular biology and genetics of the Fruit Fly (*Drosophila melanogaster*) that is used as a research tool and model organism. It contains 10581 annotated genes [67].**
- **WormBase: This is a database of the model organism *Caenorhabditis Elegans*. It contains 14156 annotated genes in gene ontology[63]**
- **Arabidopsis thaliana TAIR/TIGR: This database contains the genes from genome *Arabidopsis thaliana* which is a model organism for plants [8]. It contains 34683 annotated genes in gene ontology [8].**
- **Trypanosoma brucei Sanger GeneDB: Contains the genetics and molecular biology related to *Trypanosoma brucei* which causes the African trypanosomiasis (or sleeping sickness) disease. There are more than 60 million people at risk in Africa.[62] It contains 3921 annotated genes in gene ontology [59].**
- **MGI: Mouse Genome Informatics provides integrated access to data on the genetics, genomics, and biology of the laboratory mouse [39]. It contains 18052 annotated genes in gene ontology [39].**

2.4. Similarity Measures

Ontology-based semantic similarity measures have been investigated for long time in different domains. First it was proposed in English domain and later it was adapted in biomedical and bioinformatics domains. The first Ontology used for measuring the semantic similarities between its terms was WordNet [12, 37, 40]. Several measures were proposed, some were based on the structure of the ontology [32] and some were related to information content of the terms [12, 23, 30, 40, 48].

▪ *Resnik Measure*

Resnik [48] proposed an information-content (IC) based measure for semantic similarity between terms and these measures were designed mainly for WordNet [12, 37]. WordNet is a freely available lexical database that represents an ontology of approximately 100,000 general English concepts [12, 37]. These measures are proven to be useful in natural language processing (NLP) tasks [44]. 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 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 $p(t)$.

$$IC(c) = -\log p(c) \quad (1)$$

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. For example in Figure 2.8 the information content of node 1 is less

than all of its descendants and the leaves (nodes 10, ..15) have the most information content and are the most specialized terms. The probability of a term to occur is assumed to be equal to its frequency in the annotations in a database [32] [51]. In Gene Ontology the frequency of each term c is calculated by:

$$freq(c) = anno(c) + \sum_{h \in children(c)} freq(h) \quad (2)$$

where $anno(c)$ is the number of genes annotated with this term in the database, $children(c)$ is the set of children for term c in GO [54]. It means that the frequency of each term equals to the number of the time that genes are annotated by this term plus the number of the times that its children are used to annotate a gene.

The probability of term t is then defined as:

$$p(t) = freq(t) / freq(root) \quad (3)$$

where $freq(root)$ is the frequency of the root term [54].

The probability assigned to a term is defined as its relative frequency of occurrence.

$$sim_{Resnik}(t_1, t_2) = -\log p(t) \quad (4)$$

$t = LCA(t_1, t_2)$

The minimum similarity is zero and there is no maximum for this measure.

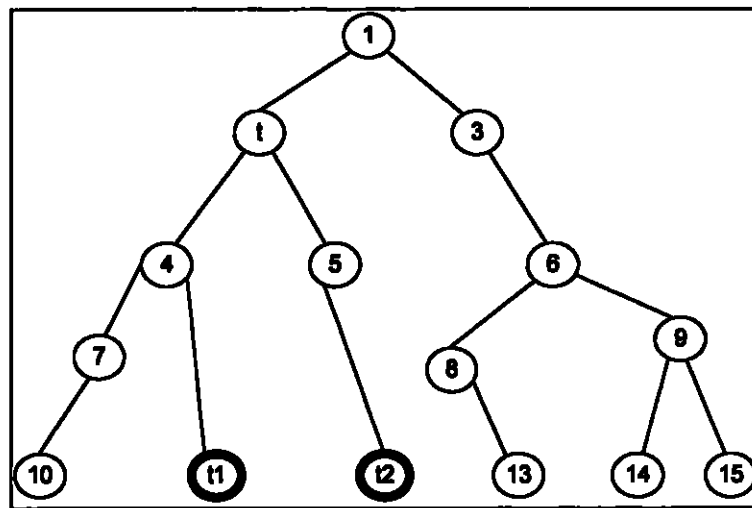


Figure 2.8. Example of a tree structure

The more frequency of occurrence means the more general term. The power of Resnik's measure is that both the relevance of the LCA itself and the distance to the LCA are taken into consideration [61]. Resnik's method only concentrates on the information content of a term derived from the corpus statistics and it ignores the structure of the ontology which is considered as a drawback of using his method in Gene Ontology in which the specificity of a GO term is usually determined by its location in GO-graph and the biological meaning of a term is inherited from all of the term's ancestors [61]. For this reason Wang et. al pointed out the information content is not an appropriate measure for the measuring the semantic similarity of the GO terms [61].

- *Jiang and Conrath*

Jiang and Conrath [23] proposed a different approach for the WordNet ontology by combining the edge based measure with information content calculation of node based techniques derived from Resnik's method. Their formula measures the distance between two terms. The distance is the reverse of their similarity measure.

$$\text{dist}_{JC}(t_1, t_2) = 2\log p(t) - (\log p(t_1) + \log p(t_2)) \quad (5)$$

$t = \text{LCA}(t_1, t_2)$

▪ *Lin's Measure*

Lin [30] 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.

$$\text{sim}_{Lin}(c_1, c_2) = \max_{c \in S(c_1, c_2)} \left(\frac{2 \cdot \log P(c)}{\log P(c_1) + \log P(c_2)} \right) \quad (6)$$

Here $S(c_1, c_2)$ is the set of common ancestors of terms c_1 and c_2 . In contrast to Resnik's similarity, the values range between 0 and 1.

▪ *Other Measures*

In 1994 Wu and Palmer [64] applied both the distance between each term with the LCA of two terms and the depth of LCA of them. Later in 1998 Leacock and Chodorow [29] proposed a formula for computing the semantic similarity or the relatedness between two terms in WordNet ontology as follows:

$$\text{sim}_{\text{LC}}(t_1, t_2) = -\log \frac{\text{Len}(t_1, t_2)}{2 \times \max_{c \in \text{WordNet}} \text{depth}(c)} \quad (7)$$

in which *Len* is the minimum path between t_1 and t_2 .

Biomedical Domain

In the Biomedical domain, measures of semantic similarity based on ontology were developed as early as 1989. Rada et al. [46] proposed the first semantic similarity measure in the biomedical domain by using *path length* between biomedical terms in the MeSH ontology [36] as a measure of semantic similarity. Al-Mubaid et al. (2007) [1] presented a technique for computing the semantic distance (similarity) between biomedical terms across multiple ontologies within a unified framework like UMLS. Also, Nguyen and Al-Mubaid (2006) [42] proposed a similarity measure for biomedical terms by combining both path length and depth features from biomedical ontologies.

In fact the path length is the distance between the terms in the ontology based on the edges needed to be traversed to reach to the other term. Path Length (PL) can be calculated easily for a tree structured Ontology such as WordNet. But for DAG-type ontology, 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) [60], have also been used for measuring semantic similarity in bioinformatics [1, 2, 4, 41], e.g. Snomed-ct [28, 40] and ICD9CM [58].

Lord et al. (2003) [32] 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 [48]. 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) [32].

Speer et al. (2004) [56] used a distance measure based on Lin's similarity for clustering genes on a microarray according to their function. Chang et al. (2001) [14] and MacCallum et al. (2000) [33] showed that Similarity between annotation and literature will augment sequence similarity searches [32]. They improved PSIBLAST (Altschul *et al.*, 1997 [6]) with similarity scores calculated over the annotations and Medline [35] references. Sevilla et al. (2005) [51] analyzed the correlation between gene expression and Resnik's, Jiang and Conraths' and Lin's measures of semantic similarity [51]. 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. On the other hand, Budanisky and Hirts [12] investigated the relatedness of Resnik [48], JC [23] and Lin's [30] measures in WordNet ontology and founded JC [23] as a superior measure to all other ones. These measures were all applied to the non-biomedical ontologies.

More recently, Schlicker et al. (2006) [54] 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

ancestor as well as how detailed the LCA is, i.e., distinguishes between generic and specific terms.

$$sim_{Rel}(c_1, c_2) = \max_{c \in S(c_1, c_2)} \left(\frac{2 \cdot \log P(c)}{\log P(c_1) + \log P(c_2)} \right) \cdot (1 - P(c)) \quad (8)$$

$S(c_1, c_2)$ is the set of common ancestors of terms c_1 and c_2 .

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 orthology¹ 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 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 [54].

xxviii_____

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 [47].

Wang et. al (2007) [61] proposed a measure to calculate the functional similarity of GO terms based on GO 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 compared with the results of Resnik's measure on the same genes [61].

Although Path length measure has been applied and explored with several biomedical ontologies [46] [44], 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 [54] [23] [32] [46].

3. A PATH LENGTH METHOD FOR GENE SIMILARITY USING GO ANNOTATIONS

This chapter presents the first gene similarity method which estimates the gene functional similarity based on the semantic similarity between the GO terms annotated for genes. As mentioned in chapter 2, Path length metric has been used in the biomedical domain as a good measure of term similarity [46] but has never been investigated in the context of gene functional similarity and gene ontology. We use the ontology structure, of the 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 similarity/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 MF. 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 . In this chapter the proposed measure is evaluated by comparing it with the sequence similarity measure.

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 terms that these genes are annotated with.

These annotation terms can be from each of biological process BP, molecular function MF, and cellular component CC ontologies. Recall that the GO is organized into these three ontologies BP, MF, and CC. For a given pair of genes (g_1 and g_2), in certain annotation database like SGD, the annotation terms for g_1 and g_2 in molecular functions will be extracted and stored in a link list. Then we calculate the first common ancestor of the terms related to the two genes. We used the February 2007 release of GO from the gene ontology website [22]. The yeast gene annotations were downloaded from the SGD site (Dec.2006) [53], FlyBase gene annotations were obtained from the GO website (Dec.2006) [22]. Here is simplified algorithm for the process:

1. For each pair of genes $\{g_1, g_2\}$ in the annotation file, the terms related to each gene are extracted from the database.
2. The path lengths between the GO terms are calculated from the GO DAG using edge counting.
3. The distance score between two genes is measured based on the average distance (shortest path length) between their GO annotation terms.

There were two ways for implementing our algorithm for computing the shortest path length between two GO nodes n_1 and n_2 :

1. Recording all the ancestors of each node (each node represents a GO term) till we reach the root. Then we compare the ancestors of n_1 and n_2 to find the common ancestors.
2. Recording just the first level ancestors of each node and comparing them to see if they have anything in common or not.

Since the second approach uses less memory and faster compared to the first approach we have applied it in our method. In next section the detail of the method is explained.

3.2. Algorithm for Distance Measure

To measure the distance between the genes we need to have distance (path length) between the terms related to each gene. In section 3.2.1 we explain how the distance between two terms is measured and in section 3.2.2 the distance between two genes are computed.

3.2.1. Distance between GO terms

To calculate the distances between each 2 terms in the gene ontology we have developed an application in .Net framework using C# language. The algorithm that is used in this program is as follows:

1. The LCA (least common ancestor) between two nodes is calculated first:

- a. The first level ancestors of each node are extracted from the gene ontology DAG.
 - b. The ancestors are then compared to each other to see if they have come up to a common ancestor or not.
 - c. When the ancestors of the two target nodes had any node in common it means that the common ancestor is found.
2. To measure the distance between two nodes we count the edges from each node to the common ancestor found in previous stage.

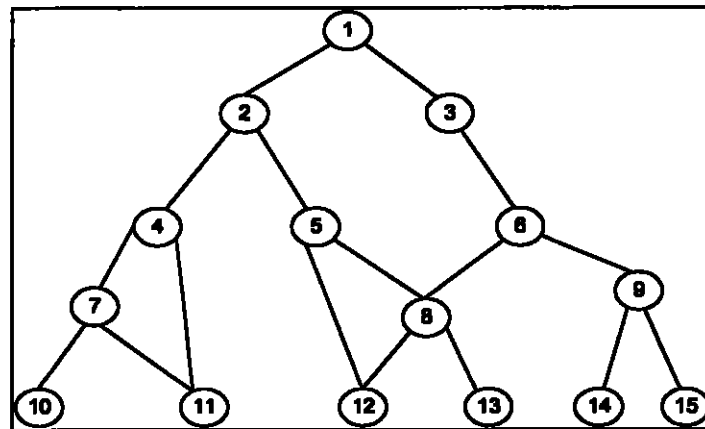


Figure 3.1. GO is a kind of DAG.

As an example we explain the algorithm of finding the first common ancestor of node 11 and node 12 in Figure 3.1. Some snapshot of the process is shown in figures 3.2 and 3.3. We have used linked list as the structure of storing the nodes in it. We have a pointer that moves from the beginning to the end of the link list to show which node's parent should be calculated. Here is the algorithm:

- 1- First the two nodes of 11 and 12 (the target nodes) are pushed as the first 2 elements of the link list. The pointer is now on the node 11 in the link list.



Figure 3.2. Stage 1 of the algorithm

- 2- The first level ancestors of the node 11 (which has the pointer on it) will be added to the list(7, 4). The pointer moves one cell further to the node 12.

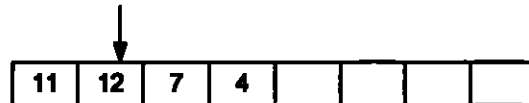


Figure 3.3. Stage 2 of the algorithm

- 3- The first level ancestors of the node 12 which are (8 and 5) are added to the list. Pointer will move further on to the node 7.

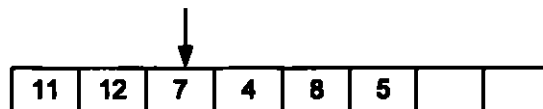


Figure 3.4. Stage 3 of the algorithm

- 4- The first level ancestor of node 7 is node 4 which had been added to the list before. Since there is no need to add the existing number to the list we just go to the next element.

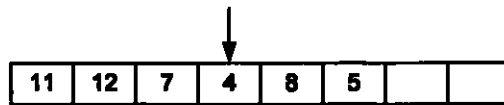


Figure 3.5. Stage 4 of the algorithm

- 5- Node 4 has the node 2 as its immediate ancestor. We add it to the list. The pointer moves on node 8.

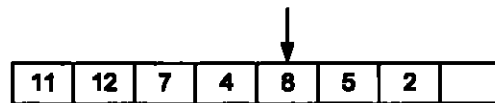


Figure 3.6. Stage 5 of the algorithm

- 6- The first level ancestors of node 8 are nodes 5 and 6. The node 5 is already in the list so we just add 6 to the list.

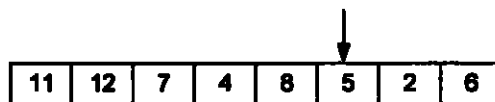


Figure 3.7. Stage 6 of the algorithm

- 7- The first level ancestor of node 5 is node 2. That has been added to the link list in the stage 5 as the parent of node 4 and node 4 was the parent of node 11. On the other hand node 5 was the ancestor of node 12. So we have reached to node 2 from two different target nodes (11 & 12) that make it the Least Common Ancestor of them.

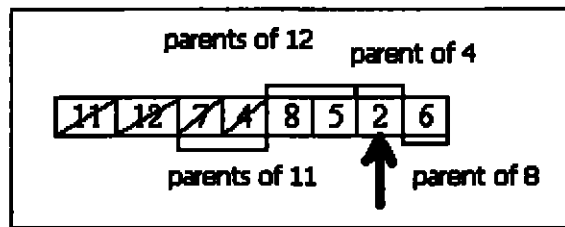


Figure 3.8. Reach the first common ancestor from two target nodes

Note: In this algorithm we keep the track of each path to see which source the ancestors are related to. If the program reaches a common ancestor from two different sources it means we have reached to the first common ancestor.

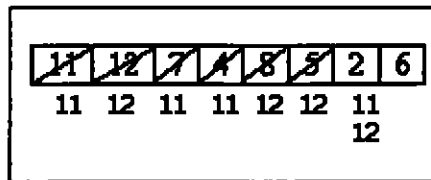


Figure 3.9. Source node(target node) of each node in the link list

Figure 3.10 shows a sample of the program run for genes AAD4 and NUP159 (from SGD). Moreover, more details about the implementation of the PL method are available in Appendix A.

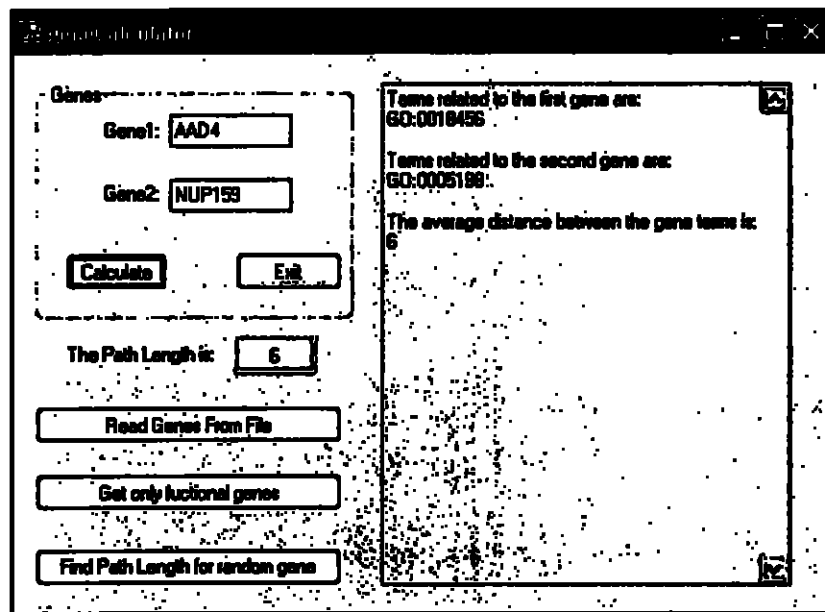


Figure 3.10. The Path Length Calculator application snapshot

3.2.2. Distance between genes

To find the distance between two genes we first calculate the distance between the GO-terms of each gene and then we derive a similarity score that represents all of them. This score could be calculated by one of the following ways:

- *Row Maxima and Column Maxima*

This is the method that has been used by Schlicker et. al [54]. They defined their measure of similarity between the genes based on the similarity value between their related terms using the maximum values of all rows and columns in the similarity matrix. As an example suppose that the Table 3.1 is the similarity matrix for GO-terms related to two genes:

	Gene ₁	term ₁	term ₂	term ₃	term ₄
Gene ₂	term ₅	d ₅₁	d ₅₂	d ₅₃	d ₅₄
	term ₆	d ₆₁	d ₆₂	d ₆₃	d ₆₄
	term ₇	d ₇₁	d ₇₂	d ₇₃	d ₇₄

Table 3.1. The similarity matrix between two genes

In this method, the maximum value in each row is extracted and the average of them forms the *rowScore*. Then the average of maximum value for each column is calculated that forms the *columnScore*. The final similarity measure is the maximum of the two values (rowScore and columnScore) [54]

$$\text{rowScore} = \frac{1}{N} \sum_{i=1}^N \max_{1 \leq j \leq M} d_{ij} \quad (1)$$

$$\text{columnScore} = \frac{1}{M} \sum_{j=1}^M \max_{1 \leq i \leq N} d_{ij} \quad (2)$$

$$\text{Similarity_Score} = \text{maximum}(\text{columnScore}, \text{rowScore}) \quad (3)$$

▪ *Average of all the GO-Distances*

For the pair of genes $\{g_1, g_2\}$ such that g_1 is annotated (for its MF) with the terms t_1, \dots, t_n while g_2 is annotated with terms t_1, \dots, 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:

$$\text{avg}\{ \text{dij} \mid i : 1..n, j : 1..m \} \quad (4)$$

For example, suppose that the 2 genes g_1 and g_2 are annotated with the following GO terms. $g_1 \rightarrow t_1, t_2, t_3, t_4$ and $g_2 \rightarrow t_1', t_2', t_5', t_6'$ where $t_1 = t_1'$ and $t_2 = t_2'$. Then their similarity matrix contains 16 values. To calculate the average we have:

$$\begin{aligned} \text{Average} = & [d(t_1, t_1') + d(t_1, t_2') + d(t_1, t_5') + d(t_1, t_6') + \\ & d(t_2, t_1') + d(t_2, t_2') + d(t_2, t_5') + d(t_2, t_6') + \\ & d(t_3, t_1') + d(t_3, t_2') + d(t_3, t_5') + d(t_3, t_6') + \\ & d(t_4, t_1') + d(t_4, t_2') + d(t_4, t_5') + d(t_4, t_6')] / 16 \end{aligned}$$

where $d(a, b)$ means the distance(or shortest path length between the 2 terms a and b).

If we simply measure the distance between each two term as mentioned above we would encounter a problem which is shown by example below.

Suppose that we have two genes that are annotated with exactly the same terms, that is $g_1 \rightarrow t_1, t_2$ and $g_2 \rightarrow t_1', t_2'$ where $t_1 = t_1'$ and $t_2 = t_2'$. The distance measure between the two genes would be $d(t_1, t_1') + d(t_1, t_2') + d(t_2, t_1') + d(t_2, t_2') = [0+1+1+0]/4 = 0.5$ which is not the desired result we expect from this measure. We expected to see the minimum distance which is zero between these two genes. Therefore we change our approach a little bit so that the distance of those terms that are common in two terms is not counted. Therefore in the above example that we had two genes of $g_1 \rightarrow t_1, t_2, t_3, t_4$ and $g_2 \rightarrow t_1', t_2', t_5', t_6'$ where $t_1 = t_1'$ and $t_2 = t_2'$ the average is calculated as follows:

$$\begin{aligned} \text{Average} = & [0 + 0 + 0 + 0 + \\ & 0 + 0 + 0 + 0 + \end{aligned}$$

$$\begin{aligned} & d(t_3, t_1') + d(t_3, t_2') + d(t_3, t_5') + d(t_3, t_6') + \\ & d(t_4, t_1') + d(t_4, t_2') + d(t_4, t_5') + d(t_4, t_6') \end{aligned} / 16$$

3.3. Comparing the results with Sequence Similarity

We used Blast tool [11] 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. [11]

In some experiments, we used another tool, WU-BLAST2 [52], 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 [11].

3.3.1. 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 [11]. 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 [11]. Bit-score is another metric of sequence similarity that 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 3.11. In which all the genes in group1 have high sequential similarity, all the genes in group1 have medium sequential similarity with group2 and all the genes in group1 have no sequential similarity with group3.

Gene1(group1)	Gene2(group1)	Path Distance	Score(bits)
AAD10	AAD4	0	1379
AAD10	AAD14	0	1362
AAD10	AAD3	0	1177
AAD10	AAD16	0	695
AAD10	AAD15	0	531
AAD10	AAD6	0	427
AAD10	YPL088W	0	227

Table 3.2. SGD genes with high sequence similarity with AAD10

Gene1(group1)	Gene2(group2)	Path Distance	Score(bits)
AAD10	POP3	9	39
AAD4	GRX4	5	0
AAD14	RRN5	8	0
AAD3	KAP95	8	0
AAD10	HUA1	5	47
AAD4	NUP159	6	-
AAD14	BFA1	8	0
AAD10	YMR041C	5	79
AAD10	RPL29	7	44
AAD10	ATP10	8	63

Table 3.3. Comparing Group1 with Group2 genes

Gene1(group1)	Gene2(group2)	Path Distance	Score(bits)
AAD10	ABZ1	8	0
AAD10	ACB1	9	0
AAD10	ACT1	7	0

AAD10	ADE17	9	0
AAD10	ADE8	10	0
AAD10	ADY2	10	0
AAD10	AGP1	9	0
AAD10	AHP1	6	0

Table 3.4. SGD genes with no similarities with AAD10

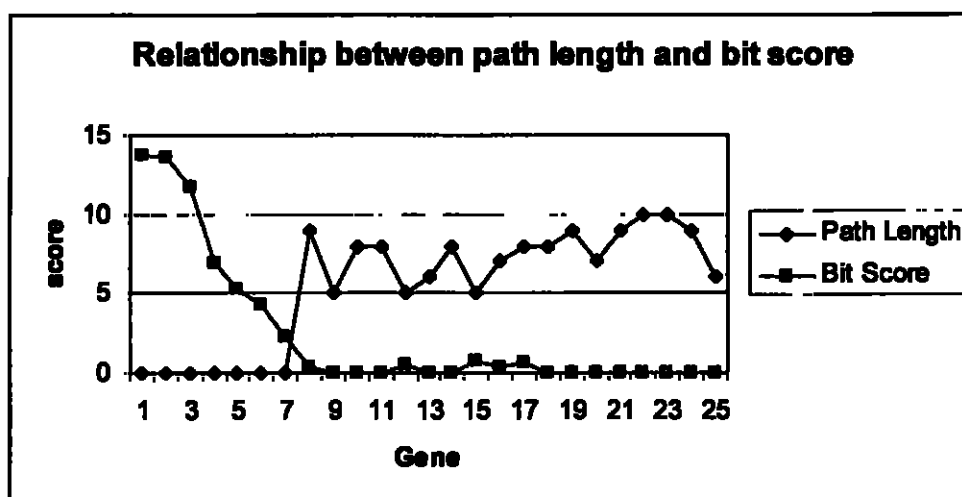


Figure 3.11. Relationship between path length and bit score

As it shown in the diagram the path length have the opposite trend compare with the bit score. The bit score values are divided into 100 to be shown easier in the diagram.

3.4. Experiments and Results

We developed a module, called *PathLengthCalculator*, to implement our proposed method for measuring the similarity between GO terms and between genes. We used the

PathLengthCalculator module to evaluate our methodology and measure the distance between the genes and proteins.

3.4.1. Distribution of Path Length

- *Distribution of PL in SGD Dataset*

We have explored 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 3.12. From this experiment (Figure 3.12) 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 [8-13] 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.

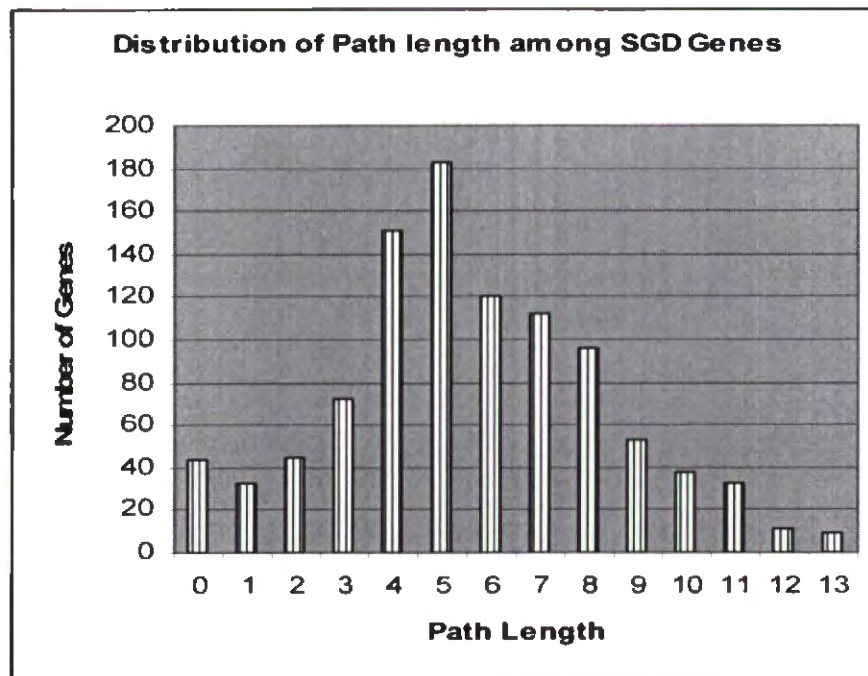


Figure 3.12. Distribution of path length among 1000 gene pairs randomly selected from SGD.

- *Distribution of PL in FlyBase Dataset*

To see the distribution of path length in FlyBase we have collected randomly 500 gene pairs from FlyBase annotation file. The path length distribution is illustrated in Figure 3.13. Again, no pattern or relation exists between FlyBase genes.

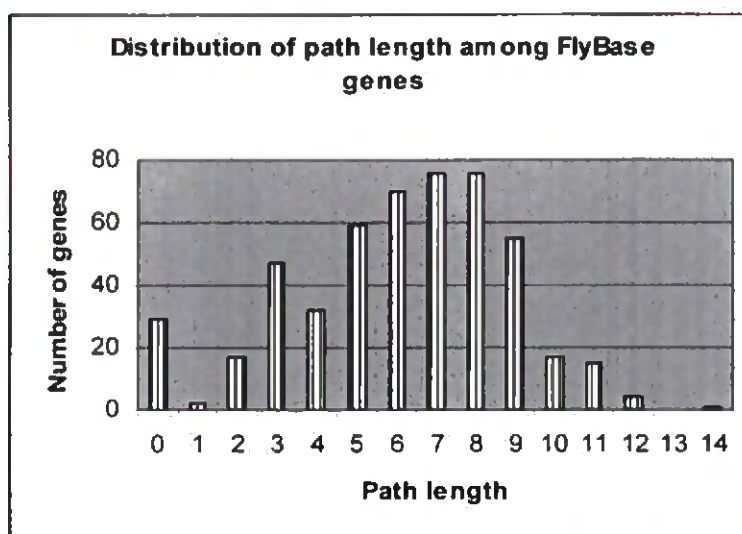


Figure 3.13. Distribution of path length among 500 gene pairs randomly selected from FlyBase

3.4.2. Evaluation based on Correlation with Sequence Similarity

In our experiments we have 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.

Gene1	Gene2	PL	BitScore	EValue
ABP140	OMS1	6	98	0.0024
ABF2	NHP10	3	102	1.80E-05
POP3	SIP3	8	79	0.13
ACE2	YPR013C	2	114	2.80E-05
AFT1	VAC7	3	85	0.82
ABF1	SRO9	3	89	0.02
ABM1	YIL102C	0	86	1.00E-05
ACA1	GCN4	4	68	0.83
ADP1	NEW1	5	83	0.0027
AAC1	MRS4	7	62	0.99
AAT1	TMA7	5	52	0.37
YMR289W	IMD3	5	70	0.83

Table 3.5. Example from SGD-LSS gene pairs

Table 3.5 shows a small part of the result for the LSS dataset as an example. We have plotted the percentages of each group (HSS, LSS, NSS) that have PL value less than 2, the PL value of greater than 2 but less than 7 and the PL value of greater than 7 in the following.

The PL measure is tested on the following datasets:

- Dataset 1 contains 200 gene pairs of HSS, 200 gene pairs of LSS, and 200 gene pairs of NSS extracted from SGD annotation database [53].

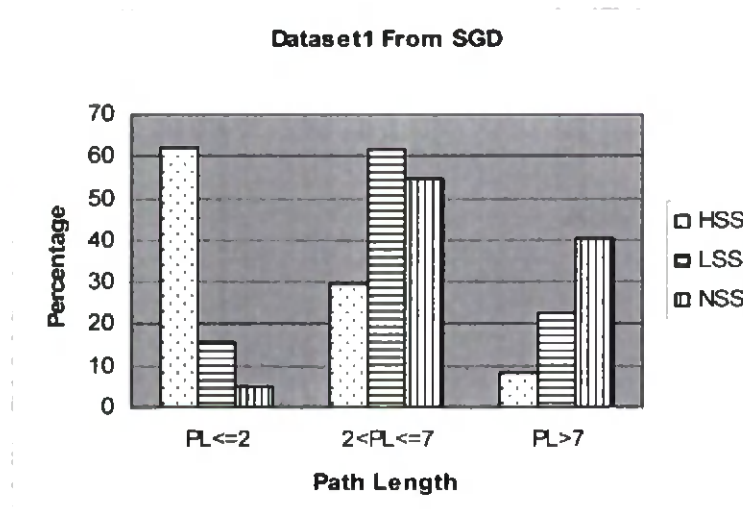


Figure 3.14. Distribution of path length between gene pairs in Dataset 1

Figure 3.14 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% of 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 3.15. 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 terms.

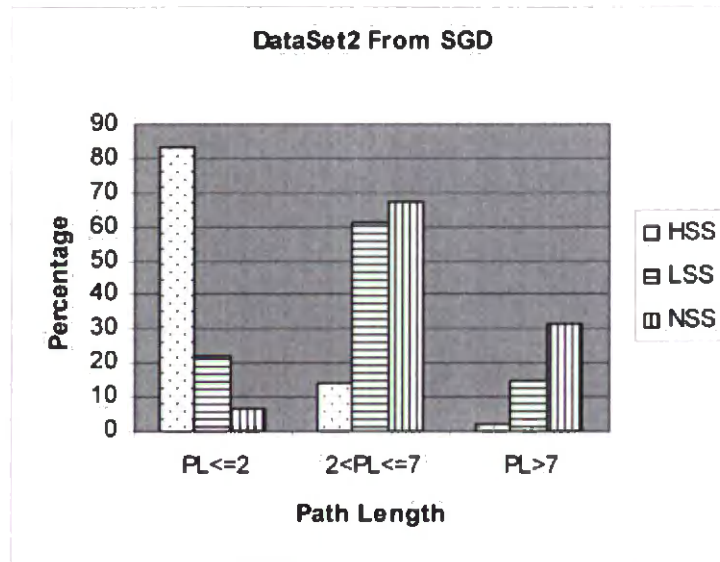


Figure 3.15. Distribution of path length between gene pairs in Dataset 2 from SGD

For example, more than 80% of HSS pairs have path length of 2 or less. Moreover, 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 3.16. 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.

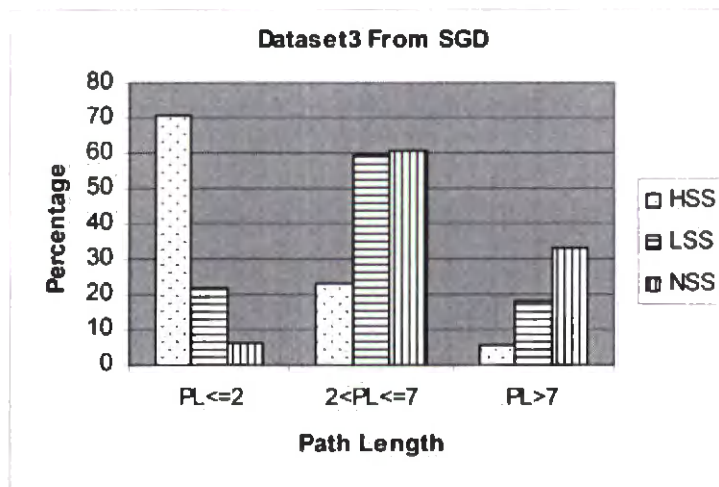


Figure 3.16. Distribution of path length between gene pairs in Dataset 3

- In another evaluation, we used genes from a different genome, the FlyBase annotation database [67]) in a new dataset (we call it 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 3.17. 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.

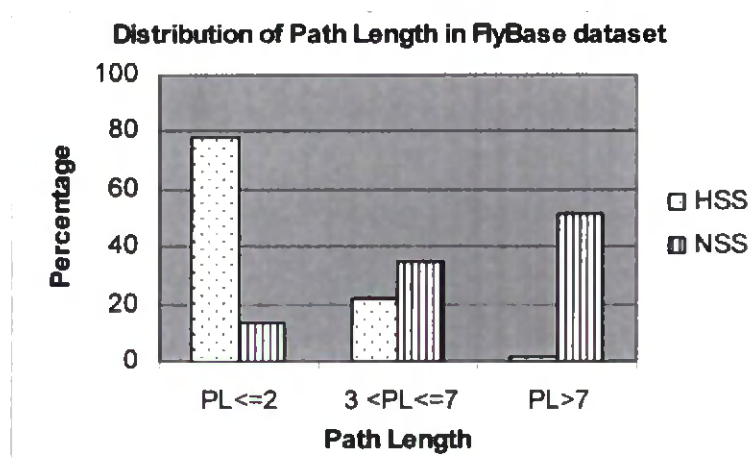


Figure 3.17. Distribution of path length between gene pairs in Dataset 4 from FlyBase

We include a listing of the gene pairs of each group (HSS, LSS, NSS) in each dataset in Appendix B.

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 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.

3.4.3. Compare Average and Maxima methods

We introduced two methods for calculating the distance between two genes in section 3.2.2: *Row Maxima and Column Maxima and Average of all the GO-Distances*. To compare between two methods, some experiments have been done. These experiments are applied on the dataset we explained in section 3.4. We call the first approach Maxima and the second approach PL in the figures:

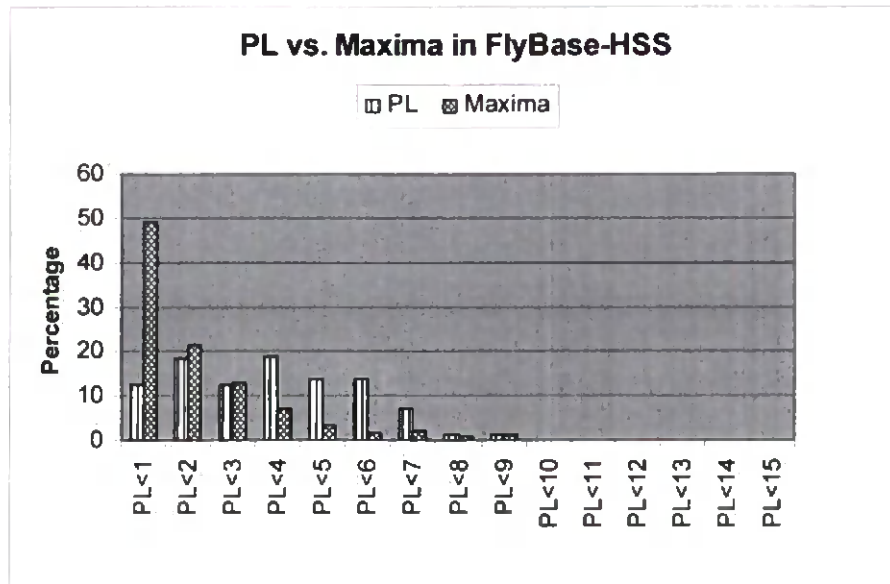


Figure 3.18. Comparison between PL and Maxima measure in HSS FlyBase dataset

As it is shown in the figure the maxima measure is doing very well in predicting the path length for the genes in FlyBase HSS. The results are even better in compare with PL measure. Near 50% of the gene pairs with high sequence similarity have the PL value of less than one. The PL is measured by considering the maximum of the rows and columns explained in section 3.2.2. Next we consider the diagram for FlyBase NSS. As it shown

below the two measures are similar to each others and both shows correlation with sequence similarity.

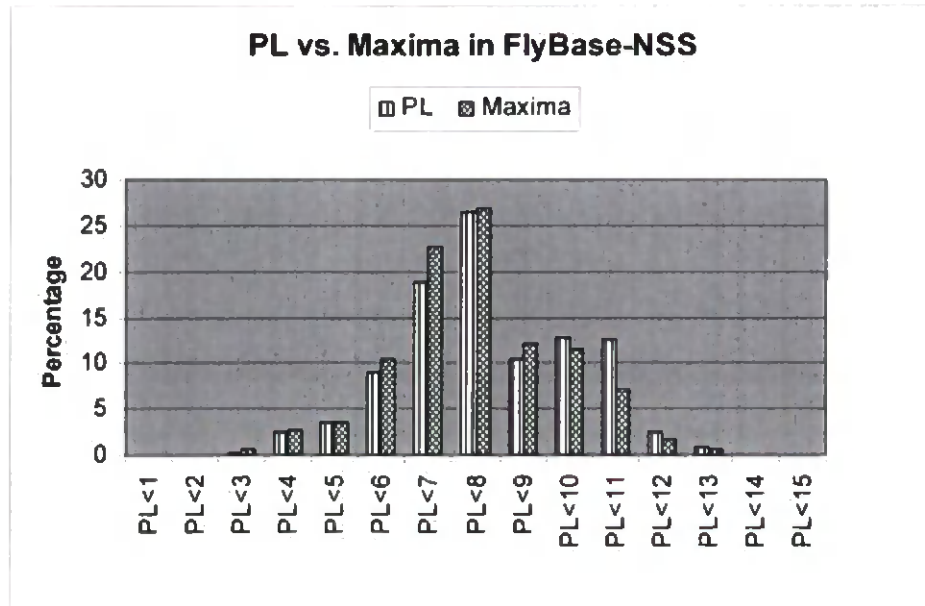


Figure 3.19. Comparison between PL and Maxima measure in NSS FlyBase dataset

The 3 datasets of SGD is also used to compare the two approaches. As you see in figure below both of the measures have correlation with sequence similarity. With PL measure 37 percent and with Maxima measure 42% of the gene pairs with high sequence similarity have the PL value less than 1.

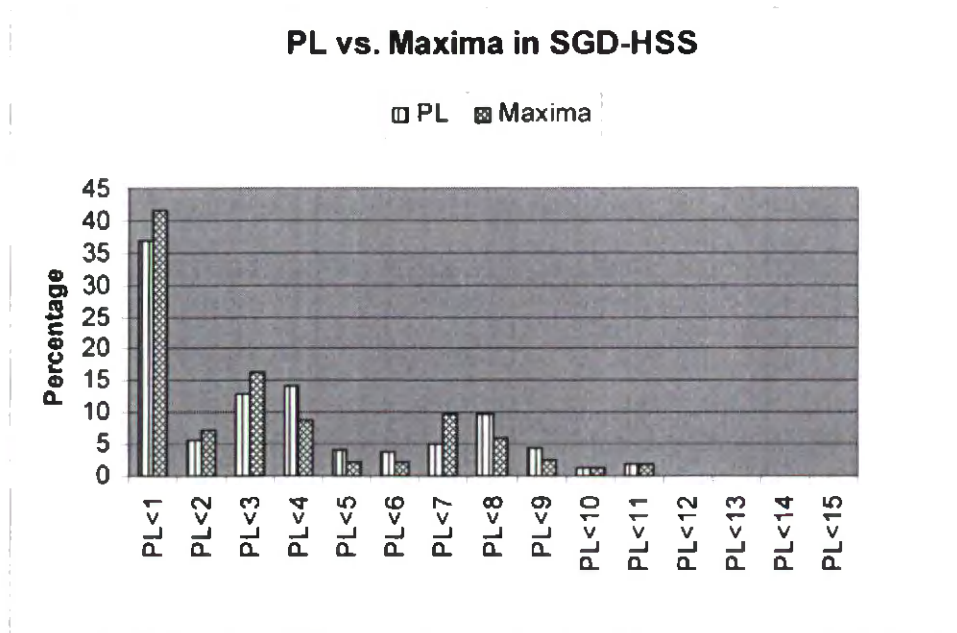


Figure 3.20. Comparison between PL and Maxima measure in HSS SGD dataset

For LSS and NSS we also can see the difference between these two measures. As it is shown below most of the pairs have the PL value of 6 in both measures which is approximately a medium distance for the gene pairs. Since we consider the PL measure less than 2 as close distance and between 2 and 7 is considered as medium distance and the PL value of greater than 7 shows a far distance between the gene pairs.

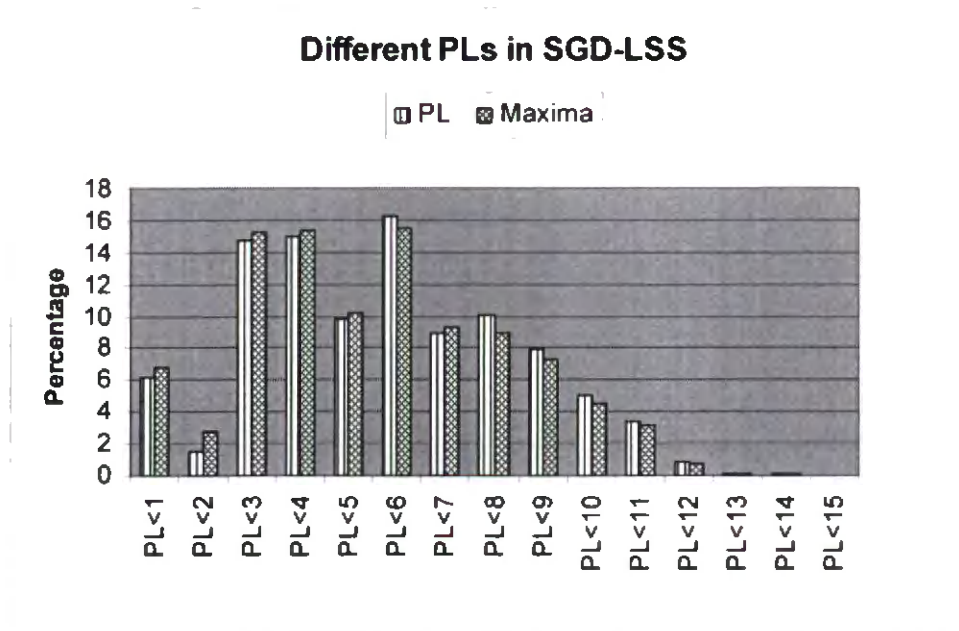


Figure 3.21. Comparison between PL and Maxima measure in LSS SGD dataset

As it is shown below more than 50% of the gene pairs have the PL measure greater than 7. Less than 5% have the PL value of less than 2 and the rest have the PL value between 2 and 7. Still the correlation can be seen clearly for both measures.

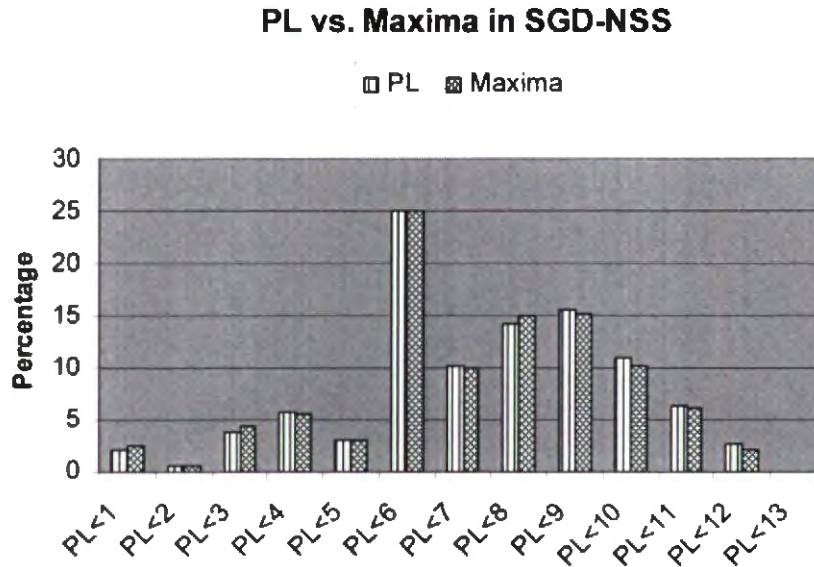


Figure 3.22. Comparison between PL and Maxima measure in NSS SGD dataset

We have also applied these two approaches to the datasets from [54]. This dataset is being further used in the rest of this thesis. It contains 4 groups of the protein pairs. Those with very high sequence similarity that is called IO dataset, those with high sequence similarity called HSS, those with low sequence similarity and no sequence similarity called LSS and NSS respectively.

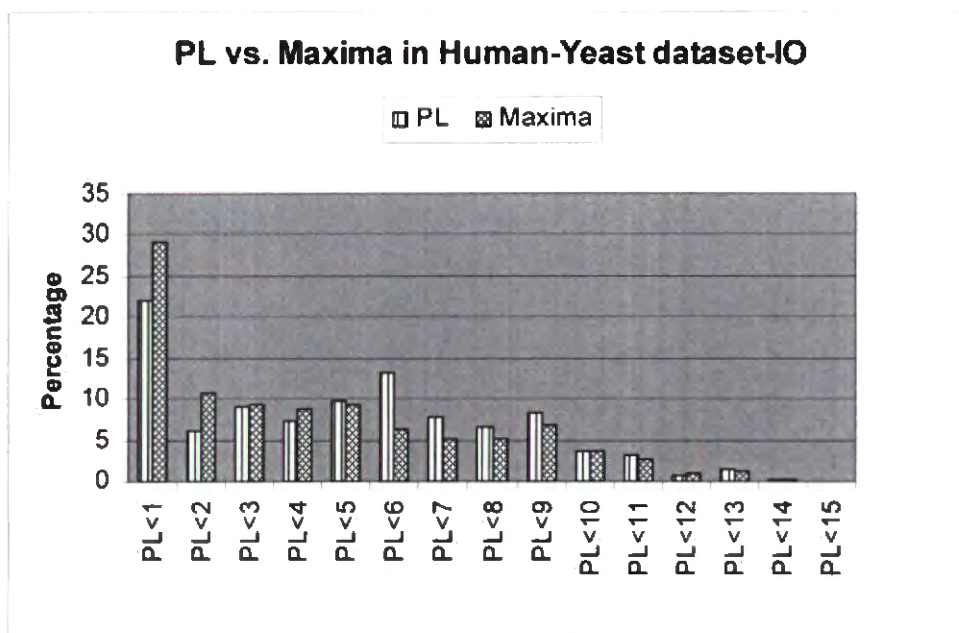


Figure 3.23. Comparison between PL and Maxima measure in IO Human-Yeast dataset

The PL and Maxima measures both show the highest percentage of protein pairs in the PL value range of less than 1. In HSS, LSS and NSS dataset we also can see that the result is the same as what we expected.

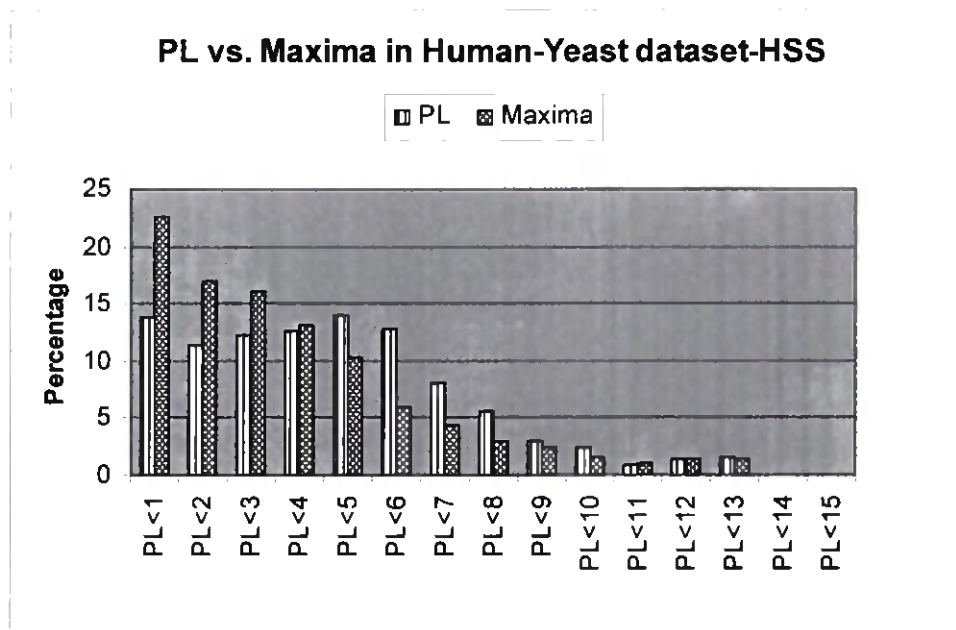


Figure 3.24. Comparison between PL and Maxima measure in HSS Human-Yeast dataset

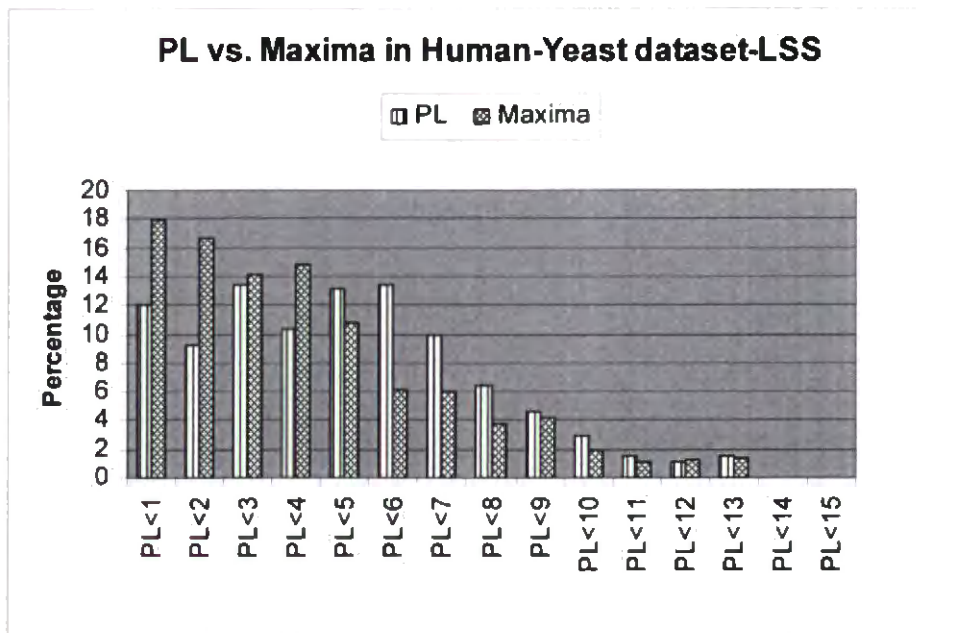


Figure 3.25. Comparison between PL and Maxima measure in LSS Human-Yeast dataset

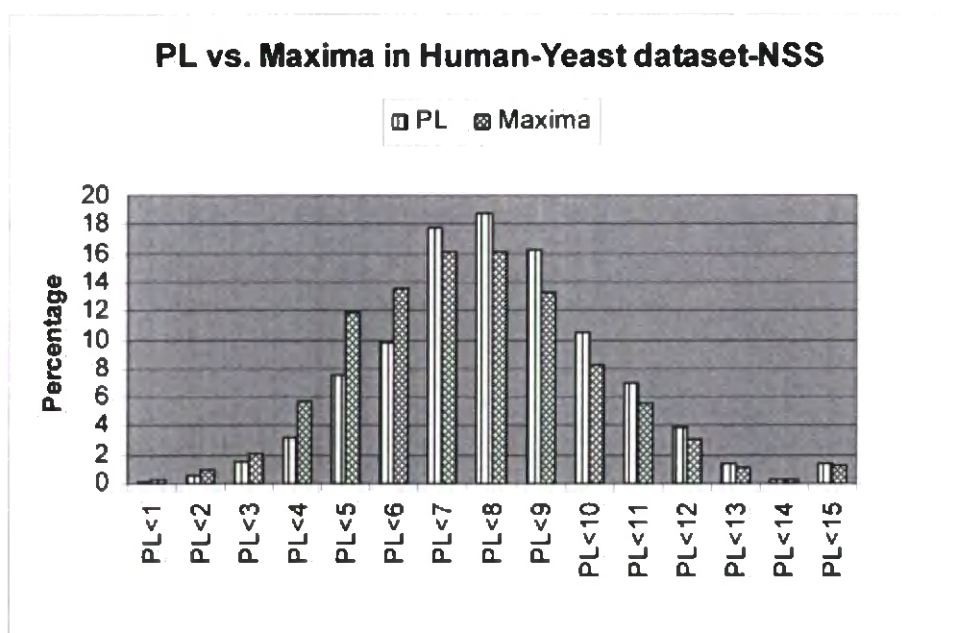


Figure 3.26. Comparison between PL and Maxima measure in NSS Human-Yeast dataset

3.4.4. Compare terms in Biological Process and Molecular Function ontologies

We have done some experiments to compare the Biological Process (BP) distance versus the Molecular Function (MF) distance in the gene ontology. We have used 2 data sets for our comparison. First we applied it to 2000 genes from FlyBase dataset.

In FlyBase HSS dataset which are those genes with high sequence similarity, it is expected that the PL measure would be small. Therefore it is more desirable for us to have the genes with $PL = 0, 1$ rather than 6, 7 and more. As shown in Figure 3.27 the MF datasets acts as what we expected. For example, most of the gene pairs (near 70%) with

high sequence similarity have the path length value less or equal to two. the percentage decreases as the distance (PL value) increases.

In BP dataset as it is shown in the Figure 3.27 less than 5% have the PL value less than or equal to two. When the path length increases the percentage of the genes with greater distance (bigger PL value) also increases.

This shows that the PL would not be a suitable measure to be used in biological process (BP) ontology.

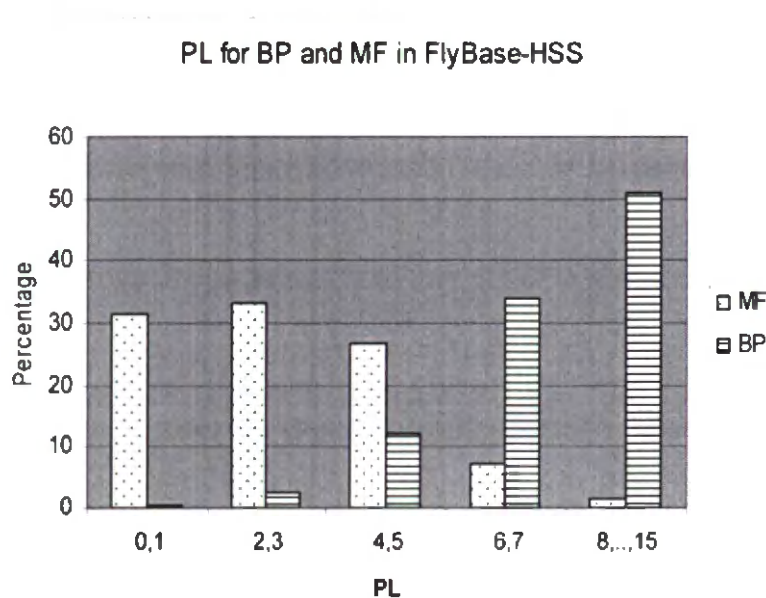


Figure 3.27. Comparison PL between BP and MF in HSS FlyBase dataset

For the genes with no sequence similarity both ontologies of BP and MF show correlation with sequence similarity. See Figure 3.28.

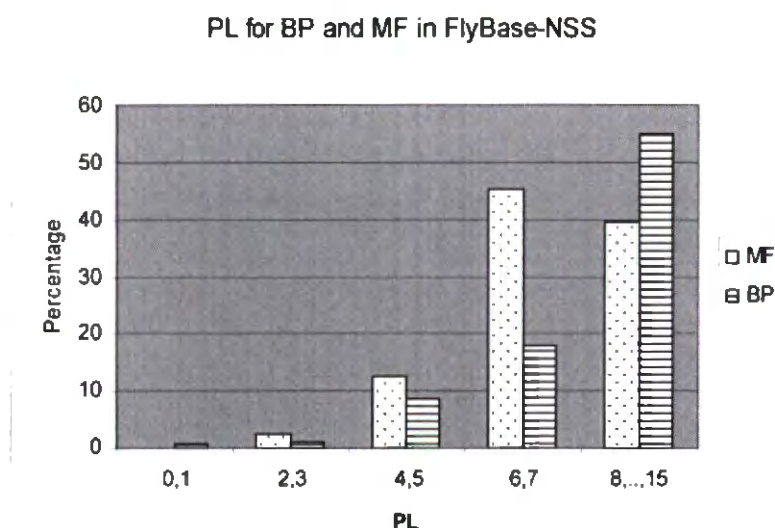


Figure 3.28. Comparison PL between BP and MF in NSS FlyBase dataset

However the desired trend has been observed in another experiment with a dataset of 4000 protein pairs from Human-Yeast [54]. Each Biological Process and Molecular Function datasets are shown separately in Figure 3.29 and Figure 3.30.

As it is shown in Figure 3.29 the highest percentage of the gene pairs with path length of less than two is related to the genes with high sequence similarity (HSS) and the highest percentage of the gene pairs with the PL value of greater than 7 is for the gene pairs with no sequence similarity (NSS).

For those gene pairs that we measured their PL value based on their annotated terms in MF ontology (Figure 3.30) we see that the highest percentage of the gene pairs with path length of less than two is related to the genes with very high sequence similarity (IO set) and the highest percentage of the gene pairs with the PL value of greater than 7 is for the gene pairs with no sequence similarity (NSS).

This also shows that MF in dataset shows more correlation with sequence similarity.

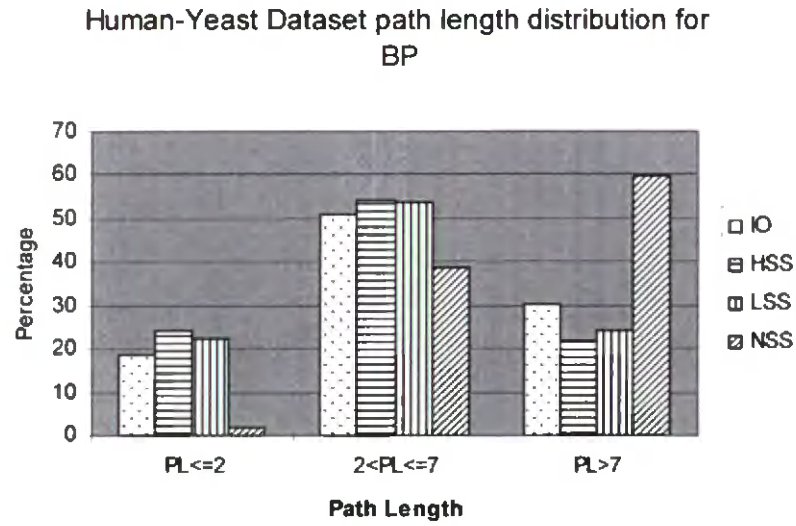


Figure 3.29. Distribution of PL in Human-Yeast dataset using BP terms

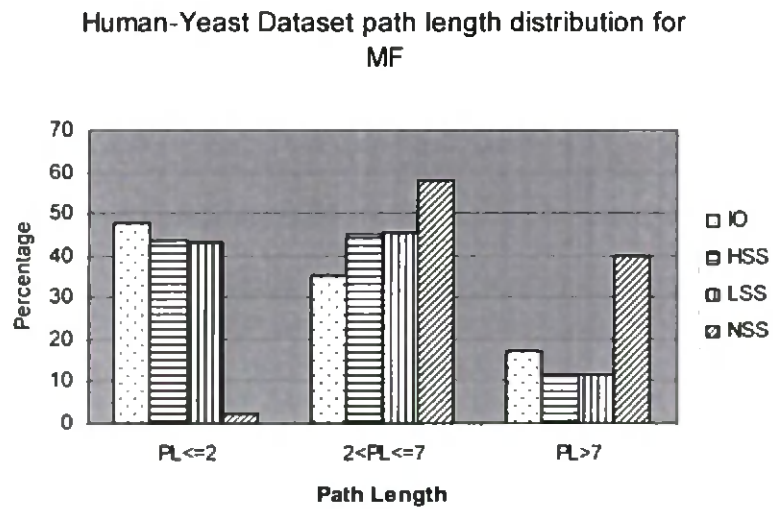


Figure 3.30. Distribution of PL in Human-Yeast dataset using MF terms

Now we consider each dataset of gene pair based on their sequence similarity separately. In IO dataset high percentage (48%) of the protein pairs have the path length less than 2 for the time that we consider their molecular function (MF) terms to calculate the PL value. The percentages of the protein pairs with the PL value between 2 and 7 and PL value greater than 7 decreases to 35%, 18% respectively that is what we expect from the pairs that have the very high sequence similarity (IO).

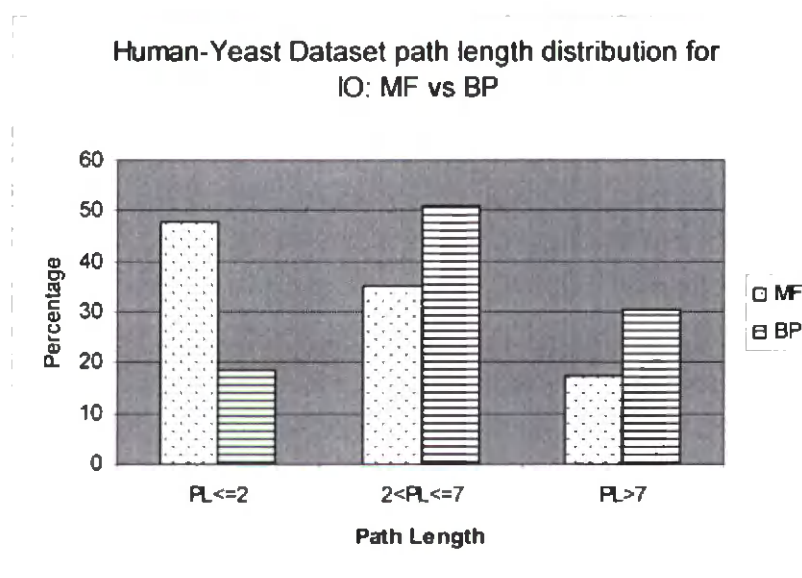


Figure 3.31. MF vs. BP in Human-Yeast IO dataset

On the other hand, both the Molecular Function and Biological Processes datasets in HSS and LSS show high percentage of protein pairs with the path length value greater than 2 and less than 7. See Figure 3.32 and Figure 3.33.

Human-Yeast Dataset path length distribution for
HSS: MF vs BP

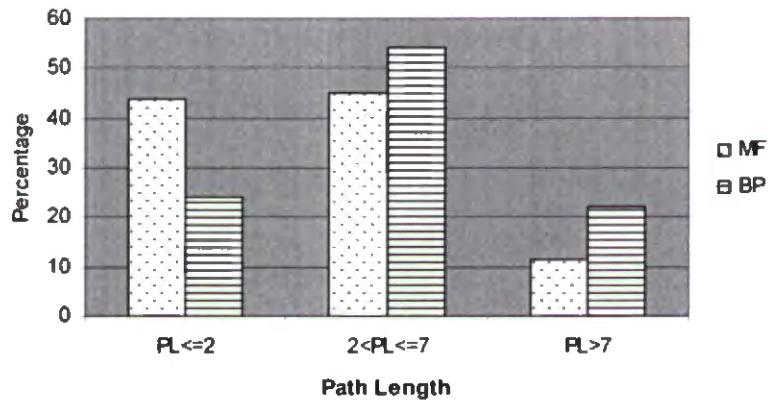


Figure 3.32. MF vs. BP in Human-Yeast HSS dataset

Human-Yeast Dataset path length distribution for
LSS: MF vs BP

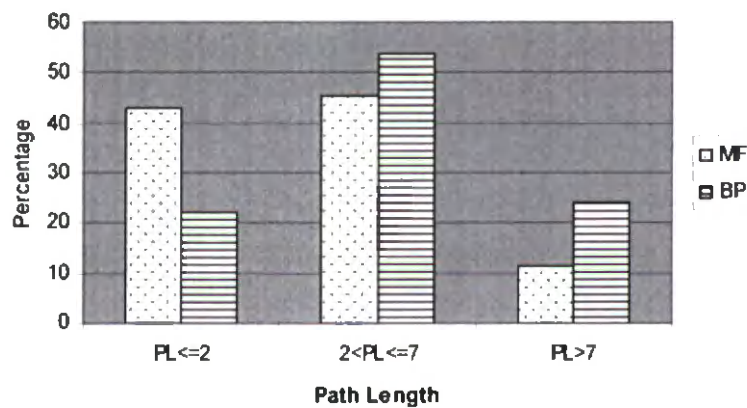


Figure 3.33. MF vs. BP in Human-Yeast LSS dataset

For NSS dataset BP shows high percentage of pairs with PL value greater than 7. Although the MF shows lesser percentage in compare with the BP, still the result is acceptable (40% of the pairs have the PL greater than 7). See Figure 3.34.

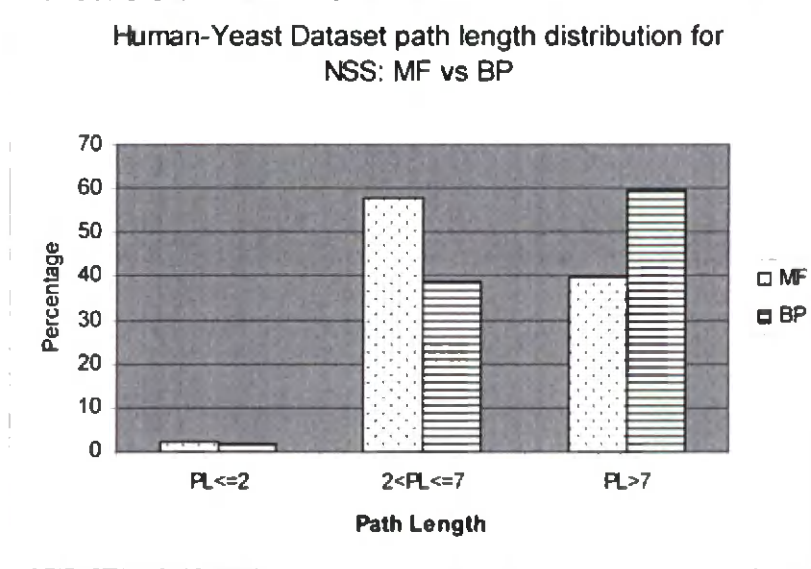


Figure 3.34. MF vs. BP in Human-Yeast NSS dataset

In general using BP terms in our measure to calculate the biological similarity between the genes shows less correlation with sequence similarity in compare with the time that we want to use MF terms to find the functional similarity between the genes.

3.5. 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. Little effort has been done for investigating the Path length feature as a metric or indicator for gene functional similarities. The work presented in this chapter 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.

4. A NEW GO STRUCTURE BASED MEASURE WITH EVALUATION USING SGD PATHWAYS

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 [1, 46, 12, 13, 44]. In this chapter, we compute path length between GO terms and modify it by considering the number of distinct minimum-length paths between the terms. Then we measure the similarity between two genes by using the semantic similarity values between their GO annotation terms and also considering the number of common GO terms between the two genes.

4.1. Distance between GO terms

To measure the similarity between genes we need to compute the distance (shortest path length) between GO terms annotated for those genes. The following are some notes that we should consider:

- 1- Each gene or protein is annotated with one or more GO terms.
- 2- Each two GO terms could have more than one minimum path among them. So that there may be more than one Least Common Ancestor (LCA) between two terms. As an example, consider the Figure 4.1 in which, each node represents a

GO-term. The LCAs between node_6 and node_1 are node_10 and node_11, because, the two nodes could be reach from 2 paths of “6-10-7-5-1” and “6-10-11-5-1”. Either of these paths has the Path Length of 4 which are the reason for the existence of two different LCAs.

- 3- In this algorithm the number of LCAs affects the measure of functional similarity. If two genes are related to each other from several different paths, it means that they have more functional similarity that those who have only one path between them

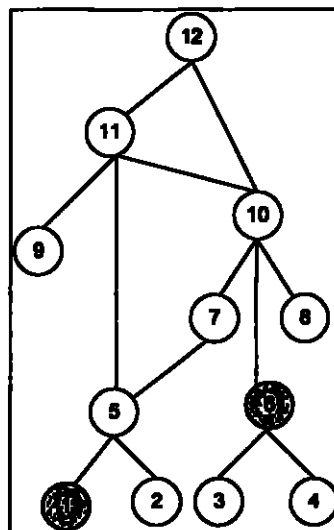


Figure 4.1. A graph to represent multiple paths in GO

As an example consider the following gene pair from FlyBase [67]:

The first gene *InR* is annotated with 4 Go-terms and the second gene *Ror* is annotated with 3 GO-terms. See Table 4.1.

Gene: InR	GO:0004716		GO:0005009		GO:0005520		GO:0005520	
Gene: Ror	PL	Nmp	PL	nmp	PL	nmp	PL	nmp
GO:0005520	0	0	2	1	10	1	11	1
GO:0005520	1	1	1	1	9	1	10	1
GO:0005520	9	1	5	1	3	1	8	2

Table 4.1. Path Length (PL) and number of minimum path (nmp) between the GO-terms for InR and Ror genes from FlyBase organism

Let us define the path length function between two GO terms go_x and go_y as follows:

$PL(go_x, go_y)$ = *the minimum path length in the GO graph between the two GO terms go_x and go_y*

(1)

But there might be more than one minimum-length path between go_x and go_y . We count number of distinct paths between go_x and go_y in the GO hierarchy. Two GO nodes might have several paths between them and among which there are two or more paths with the minimum length. This means that we can have more than one Least Common Ancestor (LCA) for two GO terms in the GO tree. The larger the number of minimum paths between two GO terms, the more similar they are. To test this hypothesis we modified the PL, Eq(1), by dividing it by number of minimum paths nmp between go_x and go_y , we call modified path length PL_m . Then $PL_m(go_x, go_y)$ is defined as:

$$\left\{ \begin{array}{ll} PL(go_x, go_y) & \text{if } nmp = 1 \\ PL(go_x, go_y)/w_1.nmp, & \text{otherwise} \end{array} \right. \quad (2)$$

where nmp is the number of minimum paths between go_x and go_y and w_1 is a weight factor to determine the contribution of nmp in PL_m . In our evaluations, we found that $w_1 = 0.6$ gives the best results.

Example: As an example, in Figure 4.2, the minimum path length between the two GO terms GO:0042626 and GO:0004129 is 7 using edge counting:

$$PL(GO:0042626, GO:0004129) = 7.$$

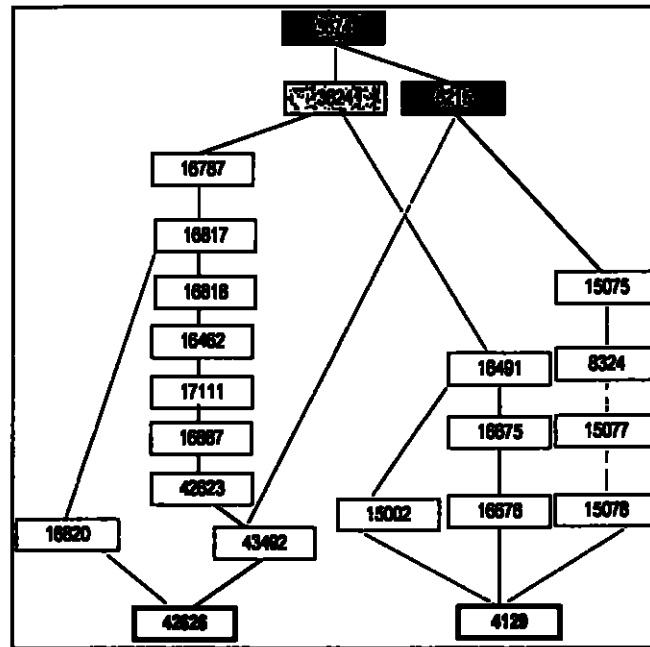


Figure 4.2. Part of the GO to illustrate the paths between two GO terms 0042626 and 0004129

We notice that there are 3 paths between GO:0042626 and GO:0004129. The first path of length 7 is via the *LCA* node GO: 0003824, while the second and third paths are via the *LCA* nodes GO: 0003674 and GO: 0002215 respectively.

$LCA(GO:0042626, GO:0004129) = \{GO:0003824, GO: 0003674, GO: 0002215\}$

Minimum-Paths (GO:0042626, GO: 0004129) =

{ 42626-16820-16817-16787-3824-16491-15002-4129; 42626-43492-5215-3674-3824-16491-15002-4129; 42626-43492-5215-15075-8324-15077-15078-4129 }

The 3824 and 5215 that have the bold format are the least common ancestor of the two target nodes. All the relations (edges) in Figure 4.2. are an “is-a” relationship, *i.e.*, each node has an “is-a” relationship with its parent node. Using Eq (2) the modified path length (PL_m) between these two GO terms is calculated as follows:

$$PL_m(GO:0042626, GO:0004129) = 7 \times \frac{1}{0.6 \times 3} = 3.89$$

4.2. Distance 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 : $GO_p = \{go_p^1, go_p^2, \dots, go_p^n\}$, and similarly, the annotation set for gene $G_q = GO_q = \{go_q^1, go_q^2, \dots, go_q^m\}$; that is, gene G_q is annotated with m different GO terms. From these two sets, GO_p and GO_q , we compute an $n \times m$ matrix of PL_m values between GO term pairs $PL_m(go_p^i, go_q^j)$ for all $i = 1, \dots, n$ and $j = 1, \dots, m$. Then we calculate the average of all PL_m values in the matrix which will be the PL_m for the two genes, that is:

$$PL_m(G_p, G_q) = \frac{\sum_{i=1}^n \sum_{j=1}^m PL_m(go_p^i, go_q^j)}{n \times m} \quad (3)$$

Now, number of minimum paths (nmp) between the two GO terms has been considered as a positive feature for similarity and thus contributed to similarity as we have seen in Eq(2). As we mentioned earlier, our method distinguishes between two different paths:

paths of length > 0 and paths of length 0 (common terms). Paths of length > 0 has been considered in calculating PL_m of two GO terms (in Eq.2) while the contribution of paths of length 0 will affect the PL_m of two genes. That is, if there is one or more paths of length 0 (*i.e.*, one or more common GO terms) in the annotation terms of the two genes then this affects their PL_m value. If the two genes G_p and G_q have one or more common terms between them, then we divide their PL_m (eq.3) by 2 times the number of common terms between G_p and G_q :

$$PL_m(G_p, G_q) = \frac{1}{2 \times nct} \frac{\sum_{i=1}^n \sum_{j=1}^m PL_m(go_i^p, go_j^q)}{n \times m} \quad (4)$$

where nct is the number of common GO terms between G_p and G_q . If G_p and G_q have no common terms between them ($nct = 0$) then we use equation (3). Notice that the number of common terms (nct) is not considered in the summation of PL_m in equation (2) because path length is 0 and dividing it by $wl * nmp$ will not reduce the result (eq.2). To have common terms between two genes means that the genes are closer and have common functionality. So the distance (path length) between them should be less.

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

$GO_{ABF1} = \{3682, 8301, 3677, 3700, 16563, 16564\}$

$GO_{IFH1} = \{3700, 3704\}$

The 6×2 matrix containing the pair-wise path length (PL) and nmp between their GO terms is shown in Table 4.2. The PL_m between IFH1 and ABF1 is computed as follows:

$$PL_m(IFH1, ABF1) = \frac{4 \times \frac{1}{2 \times 0.6} + 2 \times \frac{1}{1} + 1 \times \frac{1}{1} + 2 \times \frac{1}{1} + 2 \times \frac{1}{1} + 5 \times \frac{1}{1} + 7 \times \frac{1}{1} + 6 \times \frac{1}{1} + 3 \times \frac{1}{1} + 3 \times \frac{1}{1} + 3 \times \frac{1}{1}}{\frac{3 \times 4}{2 \times 1}} = 1.6$$

		IFH1			
		GO:0003700		GO:0003704	
ABF1		PL	nmp	PL	Nmp
	GO:0003682	4	2	5	1
	GO:0008301	2	1	7	1
	GO:0003677	1	1	6	1
	GO:0003700	0	0	3	1
	GO:0016563	2	1	3	1
	GO:0015564	2	1	3	1

Table 4.2. PL and nmp values between GO terms of two SGD genes (ABF1 and IFH1).

4.3. Similarity between Genes

Finally, the functional similarity between two genes G_p and G_q is as follows:

$$\text{Sim}(G_p, G_q) = \max_{g \in \mathcal{P}} -PL_m(G_p, G_q) \quad (5)$$

Therefore, for the last example we have:

$$\text{Sim}(G_p, G_q) = 15 - 1.6 = 13.4$$

The max_{go_pl} in the formula above is the maximum PL value in GO, in our experiments, we used $\text{max}_{go_pl} = 15$ because, according to the research done by Delfs et. al [15] the Gene Ontology had a depth of 13 levels based on the study they had in the year 2003. The depth of gene ontology never remains the same and it would be gradually increasing by the advent of new GO terms. We have used depth 15 in our experiments but the depth and the number of the words in gene ontology tend to be changed in future.

4.4. Experimental Results and Evaluation

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 [12, 13]. In this thesis since we are in the context of gene functional similarity using GO annotations, the evaluation methodologies include: - comparing the computed similarity values with gene sequence similarity [23, 13, 54, 1] with gene expression profiles [51], or using gene pathways and clusters information to validate the results [61]. In this chapter we followed the third approach, as in [61], and we compared our measure with two measures [48, 61] .

The semantic similarity measure of Resnik [48] calculates the similarity between two terms $[t_1, t_2]$ in Ontology (e.g., WordNet) as the information content (IC) of the least common ancestor (LCA) of t_1, t_2 . As what Sevilla et al. (2005) [51] found from the analysis of the correlation between gene expression and other IC based measures (Resnik, 1995[48]; Jiang and Conrath, 1997 [23]; Lin, 1998 [30]), Resnik's measure turned out to be more accurate than the others. For this reason, we chose to compare our method experimentally with Resnik's measure. For that, we measured the similarity of gene pairs in SGD pathways obtained from <http://pathway.yeastgenome.org/>. We have obtained pathways #5 (*allantoin degradation*) and #6 (*arginine biosynthesis*) containing 4 and 7 genes respectively (pathways 1 to 4 contains less than 3 genes each). The similarity values among the gene pairs of pathways 5 & 6 are shown in Table 4.3 for both our method and Resnik's measure. First, we notice that in pathway #5 with 4 genes (*DAL1, DAL2, DAL3, DUR1,2*) and 6 gene pairs, both techniques produced consistent results.

	Gene1	Gene2	Resnik	Proposed
Pathway 5	DAL1	DAL2	2.47	11
	DAL1	DAL3	2.47	11
	DAL1	DUR1,2	1.74	9.5
	DAL2	DAL3	5.22	13
	DAL2	DUR1,2	1.74	9.5
	DAL3	DUR1,2	1.74	9.5
Pathway 6	ARG1	ARG2	0.28	8.5
	ARG1	ARG3	0.28	8
	ARG1	ARG4	0.28	8
	ARG1	ARG5,6	0.28	6.67
	ARG1	ARG8	0.28	8
	ARG1	ECM40	0.28	6.67
	ARG2	ARG3	1.38	7.5
	ARG2	ARG4	0.28	5.83
	ARG2	ARG5,6	1.01	6.67
	ARG2	ARG8	1.38	7.5
	ARG2	ECM40	5.76	14.5
	ARG3	ARG4	0.28	7

ARG3	ARG5,6	1.01	8.5
ARG3	ARG8	1.38	9
ARG3	ECM40	1.38	7.5
ARG4	ARG5,6	0.28	7.67
ARG4	ARG8	0.28	7
ARG4	ECM40	0.28	5.83
ARG5,6	ARG8	1.01	8
ARG5,6	ECM40	1.10	6.67
ARG8	ECM40	1.38	7.5

Table 4.3. Comparison of our result with Resnik's result in two pathways from SGD.

For example, both measures gave the gene pair (DAL2, DAL3) the highest similarity whereas the 3 pairs (DAL1, DUR1,2; DAL2, DUR1,2; DAL3, DUR1,2) received the lowest similarity.

Pathway #6 demonstrated some differences in the similarity values produced by our measure and Resnik's measure. For example, if we compare the two pairs (ARG2, ARG3) and (ARG3, ARG5,6) we see that Resnik's measure gives higher similarity value (1.38) for (ARG2, ARG3) than for (ARG3, ARG5,6) (1.01), however, in GO tree, the distance between the terms annotating (ARG2, ARG3) and (ARG3, ARG5,6) are 9 and 6 respectively. Our measure gave higher similarity (8.5) for (ARG3, ARG5,6) than for the other pair (7.5) which is more consistent with the annotations in the GO tree. Let us consider the pair (ARG4, ARG8) with the pair (ARG1, ARG8). Both pairs have the same similarity of 0.28 based on Resnik measure, but in GO graph we notice that the distance between the GO terms annotating ARG4 and ARG8 is larger than the distance of the GO terms of ARG1 and ARG8. Our measure reflects this fact and gives higher similarity for the pair (ARG1, ARG8) than for the pair (ARG4, ARG8), see Table 4.3. Thus our measure is closer to human sense than Resnik's measure. Comparing (ARG1, ARG5,6) and (ARG1, ARG2) shows that there are three paths of minimum length 7 between the

GO terms of first gene pair, and for the second gene pair there are 2 paths with the minimum length of 10 between them. Therefore, it is a logical perspective that the first pair (*i.e.* (ARG1, ARG5,6)) is more similar than the second one. Again, Resnik's measure gives the same similarity value (of 0.28) for these two pairs while our measure gives similarity values of 8.5 and 6.6 to them, respectively, which shows that the first pair is more similar and this is closer to the human (curators) similarity estimates when they annotated these genes. Let's examine, further, the two pairs of (ARG4, ARG5,6) and (ARG3, ARG4). In GO hierarchy there are 3 distinct paths of length 8 between the terms of first pair (ARG4, ARG5,6) while there is only one path, also of length 8, between the GO terms of the second pair. Therefore the genes in the first pair are more bounded to each other compared with the second pair. As we see in Table 4.3, both pairs have the equal similarity value of 0.28 by Resnik's measure whereas the proposed measure gives the value of 7.6 to the first and 7 to the second pair which is again evidence that the proposed measure produces better results.

In another evaluation phase, we examined the proposed measure along with a newly published measure (Wang et al. 2007) [61]. In experimenting with the same pathways as [61], our measures produced results that are very competitive and sometimes closer to human perspective which is the criteria that Wang et al. have emphasized the most [61].

	- ARO8	ARO9	ARO10	PDC6	PDC5	PDC1	SFA1	ADH5	ADH4	ADH3	ADH2	ADH1
ARO8		15	7.3	7	7	7	7	7	6	7	7	7
ARO9			7.3	7	7	7	7	7	6	7	7	7
ARO10				14.9	14.9	14.9	7.3	7.3	6.3	7.3	7.3	7.3
PDC6					15	15	7	7	6	7	7	7
PDC5						15	7	7	6	7	7	7
PDC1							7	7	6	7	7	7
SFA1								14.7	11	14.7	14.7	14.7
ADH5									14	15	15	15
ADH4										14	14	14
ADH3											15	15
ADH2												15

ADH1												
------	--	--	--	--	--	--	--	--	--	--	--	--

Table 4.4. Similarity values among genes in tryptophan degradation pathway based on our algorithm

In [61], the proposed measure is used to cluster the genes in each pathway and reported in their paper the results for pathway #141 (*tryptophan degradation pathway*). We tested our method on SGD pathway 141 and the similarity values for our measure and their measure are shown in Tables 4.4 and 4.5, respectively. Moreover, Figures 4.3 and 4.4 show the clusters that resulted from both methods.

	ARO8	ARO9	ARO10	PDC6	PDC5	PDC1	SFA1	ADH5	ADH4	ADH3	ADH2	ADH1
ARO8		1	0.22	0.20	0.20	0.199	0.199	0.199	0.199	0.199	0.173	0.199
ARO9			0.217	0.199	0.199	0.199	0.199	0.199	0.199	0.199	0.173	0.199
ARO10				0.898	0.898	0.898	0.221	0.217	0.217	0.217	0.190	0.217
PDC6					1	1	0.199	0.199	0.199	0.199	0.173	0.199
PDC5						1	0.199	0.199	0.199	0.199	0.173	0.199
PDC1							0.199	0.199	0.199	0.199	0.173	0.199
SFA1								0.779	0.779	0.779	0.877	0.779
ADH5									1	1	0.869	1
ADH4										1	0.869	1
ADH3											0.869	1
ADH2												0.869
ADH1												

Table 4.5. Similarity values among genes in tryptophan degradation pathway based Wang et al.'s measure [61].

Threshold	Initial	15	14.9	14.7	14	7.3
	ADH1	ADH1	ADH1	ADH1	ADH1	
	ADH2	ADH2	ADH2	ADH2	ADH2	
	ADH3	ADH3	ADH3	ADH3	ADH3	
	ADH4	ADH4	ADH4	ADH4	ADH4	
	ADH5	ADH5	ADH5	ADH5	ADH5	
	ARO8	ARO8	ARO8	ARO8	ARO8	
	ARO9	ARO9	ARO9	ARO9	ARO9	
	ARO10	ARO10	ARO10	ARO10	ARO10	
	PDC6	PDC6	PDC6	PDC6	PDC6	
	PDC5	PDC5	PDC5	PDC5	PDC5	
	PDC1	PDC1	PDC1	PDC1	PDC1	
	SFA1	SFA1	SFA1	SFA1	SFA1	
	ADH1	ADH1	ADH1	ADH1	ADH1	
	ADH2	ADH2	ADH2	ADH2	ADH2	
	ADH3	ADH3	ADH3	ADH3	ADH3	
	ADH4	ADH4	ADH4	ADH4	ADH4	
	ADH5	ADH5	ADH5	ADH5	ADH5	
	ARO8	ARO8	ARO8	ARO8	ARO8	
	ARO9	ARO9	ARO9	ARO9	ARO9	
	ARO10	ARO10	ARO10	ARO10	ARO10	
	PDC6	PDC6	PDC6	PDC6	PDC6	
	PDC5	PDC5	PDC5	PDC5	PDC5	
	PDC1	PDC1	PDC1	PDC1	PDC1	
	SFA1	SFA1	SFA1	SFA1	SFA1	
	ADH1	ADH1	ADH1	ADH1	ADH1	
	ADH2	ADH2	ADH2	ADH2	ADH2	
	ADH3	ADH3	ADH3	ADH3	ADH3	
	ADH4	ADH4	ADH4	ADH4	ADH4	
	ADH5	ADH5	ADH5	ADH5	ADH5	

	SFA1	SFA1	SFA1			ADH4
	PDC1	PDC1	PDC1	PDC1	PDC1	SFA1
	PDC2	PDC2	PDC2	PDC2	PDC2	PDC1
	PDC3	PDC3	PDC3	PDC3	PDC3	PDC2
	PDC4		ARO10	ARO10	ARO10	PDC3
	ARO10	ARO10				ARO10
	ARO4	ARO4	ARO4	ARO4	ARO4	ARO4
	ARO5	ARO5	ARO5	ARO5	ARO5	ARO5

Figure 4.3. Clustering genes in tryptophan degradation pathway based on our algorithm

Comparing these two measures on this particular gene group, we found that both measures give very similar and consistent results (Tables 4.4 & 4.5) with few differences in the resulted similarity values as follows. The similarity value by our measure is 14.7 for the pair (SFA1, ADH5) and 14.0 for the pair (ADH4, ADH5); therefore SFA1 will be clustered with ADH5 group sooner than ADH4 according to our measure. But in Wang's method ADH4 is clustered with ADH5 before SFA1 is clustered with the ADH5 group, since the similarity values are 0.87 and 0.78 for (ADH4, ADH5) and (SFA1, ADH5) respectively.

Threshold	Initial	1.000	0.890	0.860	0.770	0.220	0.210
Clustering Results	ADH1	ADH1	ADH1	ADH1	ADH1		
		ADH2	ADH2	ADH2	ADH2	ADH1	ADH1
	ADH2	ADH3	ADH3	ADH3	ADH3	ADH2	ADH2
		ADH4	ADH4	ADH4	ADH4	ADH3	ADH3
	ADH3			ADH4	ADH4	ADH4	ADH4
		ADH4	ADH4		SFA1	ADH4	ADH4
	ADH4			SFA1		SFA1	SFA1
		SFA1	SFA1			EDC1	EDC1
	ADH5				EDC1	EDC1	EDC1
				EDC1	EDC1	EDC1	EDC1
	SFA1	EDC1	EDC1	EDC1	EDC1	ARO10	ARO10
		EDC1	EDC1	EDC1	ARO10		ARO10
	EDC1	EDC1	EDC1	ARO10			ARO10
			ARO10				
	EDC1	ARO10			ARO10		
				ARO10	ARO10	ARO10	
	EDC1		ARO10	ARO10		ARO10	
		ARO10	ARO10				
	ARO10	ARO10					
	ARO10						
	ARO10						

Figure 4.4. Clustering genes in tryptophan degradation pathway based on [61].

By examining the GO annotation terms of these genes, we find that SFA1 and ADH5 are both annotated with the same GO term “*alcohol dehydrogenase activity*”, while ADH4 & ADH5 have no common terms between them; See table 4.6. This confirms that our measure is closer to human perspective than the measure of Wang et al. [61].

ADH5	
GO:0004022	alcohol dehydrogenase activity
SFA1	
GO:0004022	alcohol dehydrogenase activity
GO:0004327	formaldehyde dehydrogenase (glutathione) activity
ADH4	
GO:0004024	alcohol dehydrogenase activity, zinc-dependent

Table 4.6. Three SGD genes with their annotation by GO terms.

4.5. Discussion and 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 features of the GO. Specifically, our measure estimates the semantic similarity between two GO terms using the various paths between them. We assign a higher weights in the similarity metric for gene pairs having common GO terms (having paths of length = 0) between their annotation sets. We also assign weights for number of minimum length paths between two terms. The strength of our measure comes from the idea that we consider all paths between the GO terms, and the paths of length zero (common terms) between two genes are treated differently. If two GO terms have multiple minimum paths between them then they have more than one LCA (least common ancestor) and hence they share more commonalities than those GO terms with one minimum path between them. We examined our measure with a large number of gene groups from SGD

pathways (*we cannot report all the results for space limitations*). The experimental results showed that our method 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.

5. CORRELATION BETWEEN DEPTH AND PATH LENGTH OF GO NODES WITH GENE SEQUENCE SIMILARITY

In this chapter we present another new similarity measure (Sim_{PLD}) for calculating the semantic similarity of terms in Gene Ontology based on the depth and path length features in GO hierarchy. That is, this method is based strictly on the ontology structure features (*i.e.*, depth and path length) without using any other information sources (like biomedical text literature, or gene expression data). The method computes the similarity between two genes as numeric figure based on the average of Sim_{PLD} between the GO terms annotated for both genes in a given gene pair.

5.1. Semantic Similarity between GO terms

In Chapter 3 we proved that the length of the shortest path (PL) between two terms in a given ontology is a suitable measure of the semantic similarity between the two GO annotation terms. In this chapter, we also consider the depth of the least common ancestor of the two terms in the previous measure which was the path length between the two terms. Then the similarity value between two genes will be the semantic similarity values between their GO term annotations.

The similarity between two GO terms is defined as

$$Sim_{PLD}(go_x, go_y) = \log\left(\frac{depth(lca(go_x, go_y))}{Max_dpth}\right) - \log\left(\frac{PL(go_x, go_y)}{2 \times Maxdpth}\right) \quad (1)$$

$PL(go_x, go_y)$ is the minimum path length in the GO graph between the two GO terms go_x and go_y . In formula 1, the first phrase is divided by the maximum of depth in the GO and second phrase is divided by 2 times the maximum depth in GO which implies the maximum PL in the gene ontology. The division operation is for the purpose of normalization and has scaled down the value of Sim_{PLD} in our computations. There is no bottom or upper limit for Sim_{PLD} value but in our experiment we got the values ranged between -2 and 2.

5.2. The Semantic Similarity of 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 : $GO_p = \{go_p^1, go_p^2, \dots, go_p^n\}$, and similarly, the annotation set for gene $G_q = GO_q = \{go_q^1, go_q^2, \dots, go_q^m\}$; that is, gene G_q 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.

$$\text{sim}_{\text{PLD}}(g_p, \text{gene}_q) = \text{avg} \{ \text{sim}_{\text{PLD}}(go_x, go_y) \mid x : 1..n, y : 1..m \} \quad (2)$$

5.3. Experiments and Results

5.3.1. Dataset

The sample size which is used in this chapter consists of 1000 gene pairs from SGD(*Saccharomyces cerevisiae*) [53] and 2000 pairs from FlyBase (*Drosophila melanogaster*) [67] genomes in one experience and 4000 protein pairs from a dataset that is used on [54]. The sample size is consistent with those researches done on the same similar subject. Indeed the size is not exactly the same or larger, still it is considered as a reasonable size. We mention some examples as the proof of this claim: Schlicker et al. 2006 [54] has applied their measure on 682 protein pairs from human and *saccharomyces cerevisiae* proteins with very high sequence similarity (IO set), 989 protein pairs with high sequence similarity (HSS set) and 989 protein pairs with low sequence similarity (LSS set). They have applied their measure to 1356 protein pairs with no sequence similarity (NSS set). Another research done by Lord et al. [31] has applied their measure of semantic similarity to those proteins with the evidence code of TAS extracted from approximately 7000 human proteins in Swiss-Prot. Dolan et al. [18] investigated on the consistency of the annotations for genes related to mouse and human. They could find out, of the complete set of human and mouse and 11860 MGI curated genes, 3948 genes

have only MGI GO annotation and 4994 genes have only GOA annotation and only 1572 genes are annotated by both groups. Khatri et al. [27] worked on genes from Homo Sapiens genome. From the 11203 genes and 5201 ontology category and 58 millions gene-function association they could extract 212 additional gene-function assignments, out of which 161 were confirmed in later releases of gene ontology database. Therefore the size of the dataset used in this chapter is consistent with the size of dataset used in other researches.

In this chapter as what we did in chapter 3 for the evaluation, we divided the datasets into different groups based on the Blast E-value of the gene pairs. 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. Remember that we grouped the gene pairs with the Blast E-value $\leq 10^{-5}$ 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.

5.3.2. Distribution of Sim_{PLD}

As it is shown in Figure 5.1, 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.

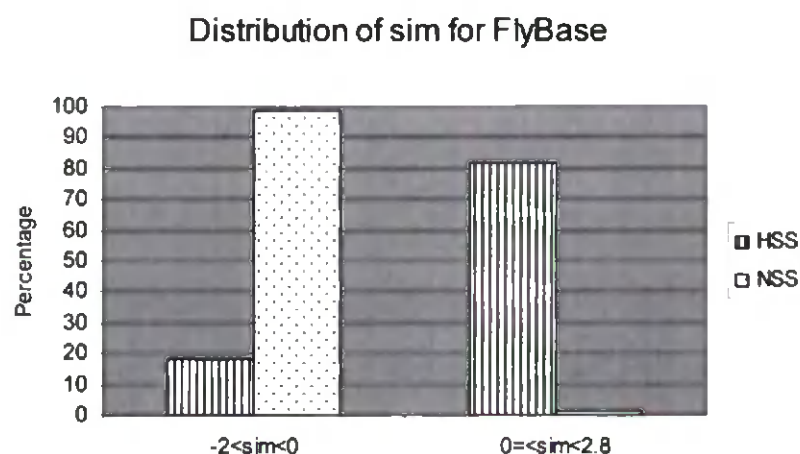


Figure 5.1. Distribution of Sim_{PLD} value between gene pairs in FlyBase dataset

In Figure 5.2 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.

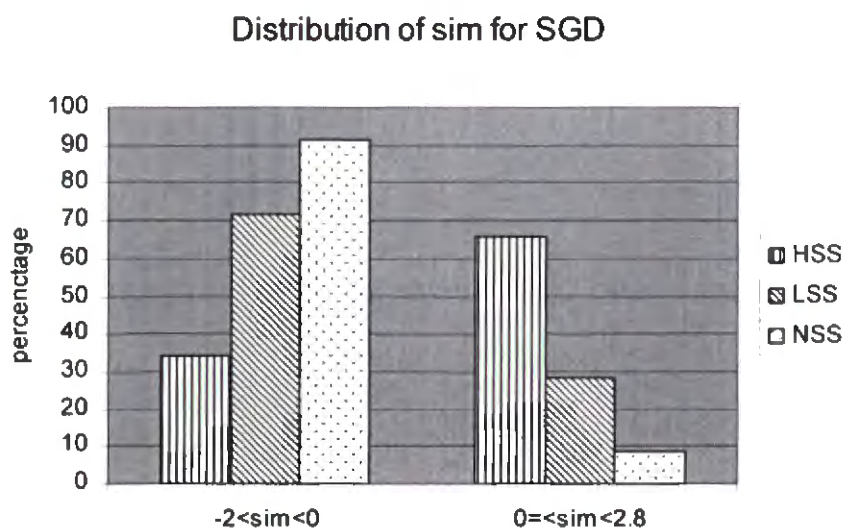


Figure 5.2. Distribution of Sim_{PLD} value between gene pairs in SGD dataset

In Figure 5.3 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.

Distribution of sim for Human-Yeast dataset

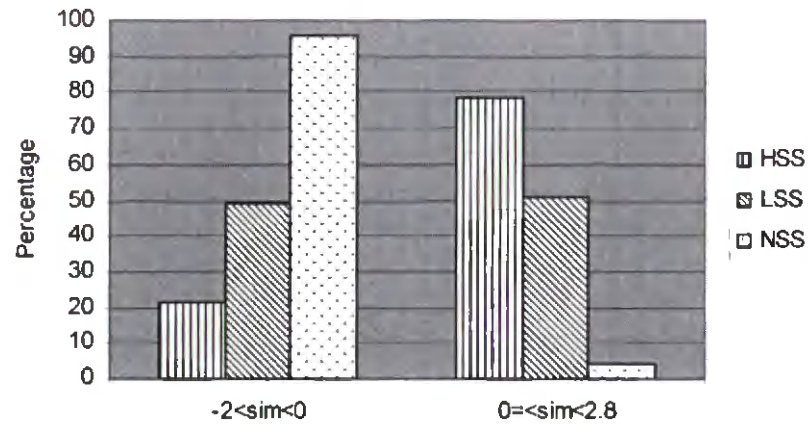


Figure 5.3. Distribution of Sim_{PLD} value between gene pairs in Human-Yeast dataset

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.

We have also plotted the distribution of Sim_{PLD} separately for each dataset that we had. Here we analyze it shortly. In figures below the Y axis is the value of Sim_{PLD} and the gene pairs are along the X axis that are sorted by their Sim_{PLD} value. For example, in Figure 5.4 FlyBase gene pairs have the minimum Sim_{PLD} value of -1.5 and the maximum Sim_{PLD} value of 2.5. The first dataset is for FlyBase gene pairs with high sequence

similarity. Although we have some gene pairs with Sim_{PLD} of negative values but most of them have the positive value. It shows compatibility with sequence similarity.

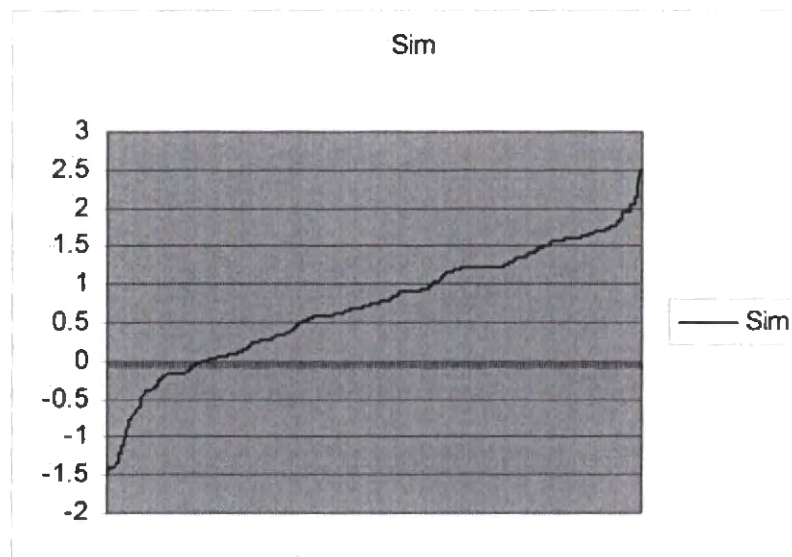


Figure 5.4. Sim_{PLD} in FlyBase HSS dataset.

The second dataset is for FlyBase gene pairs with low sequence similarity. Although we have some gene pairs with Sim_{PLD} of positive values but most of them have the negative values. It also shows correlation with BLAST value.

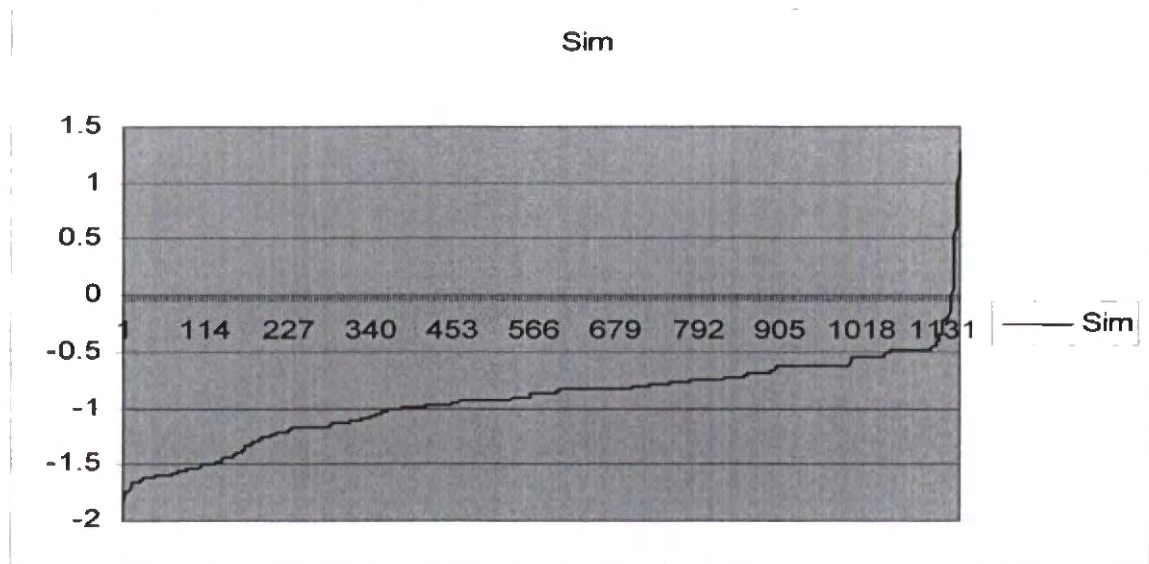


Figure 5.5. Sim_{PLD} in FlyBase NSS dataset

We applied the same measure to SGD and observed that for the pairs with high sequence similarity some of the gene pairs have the Sim_{PLD} of negative, we had lots of value zero and some of the positive values. For the dataset with low and no sequence similarity the number of zero and positive values decreases and the number of negative increases as we move to the lower sequence similarity. It is also showing a good correlation with sequence similarity.

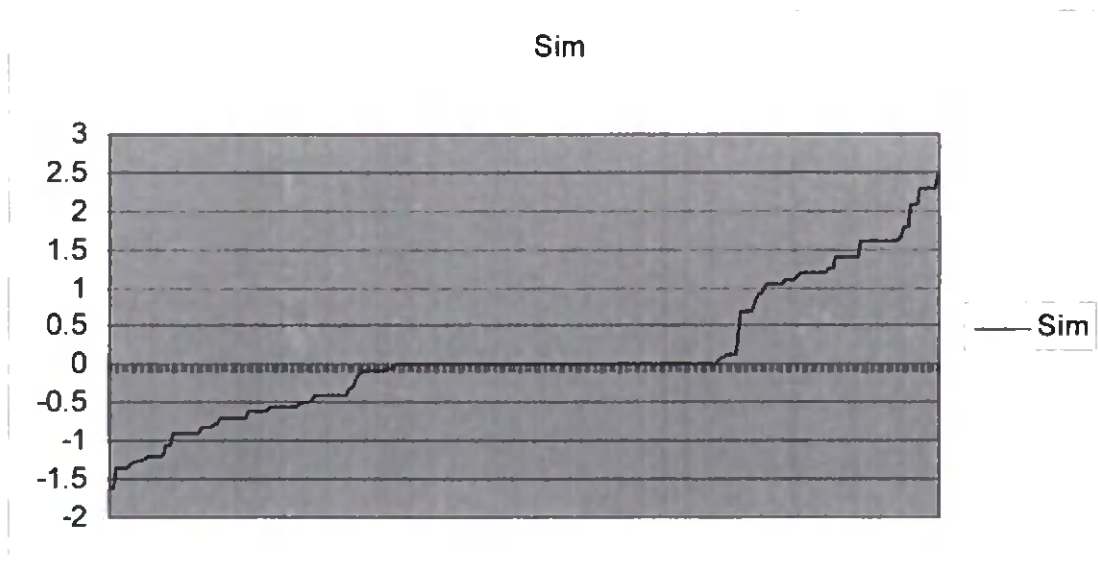


Figure 5.6. Sim_{PLD} in SGD HSS dataset

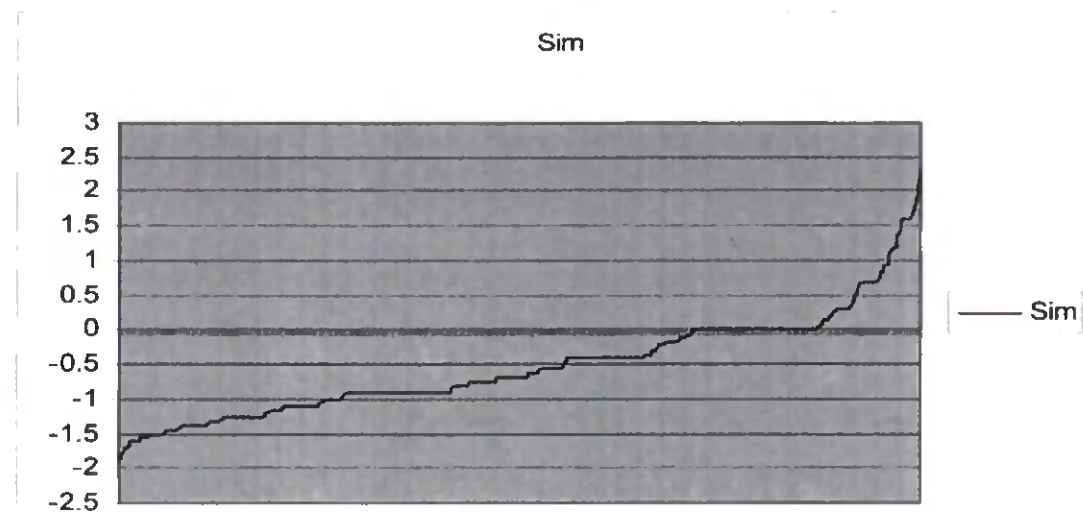


Figure 5.7. Sim_{PLD} in SGD LSS dataset

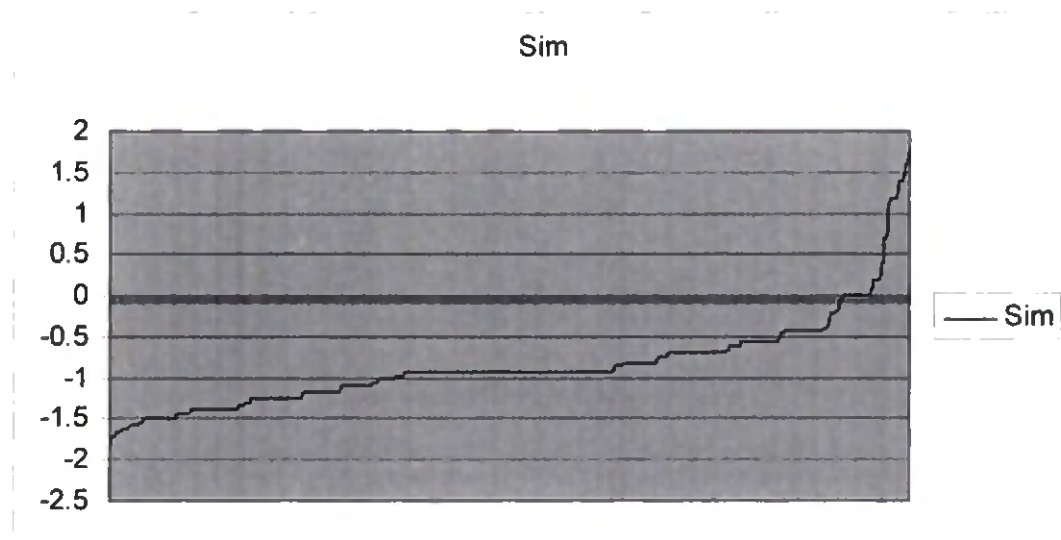


Figure 5.8. Sim_{PLD} in SGD NSS dataset

For Human-Yeast dataset HSS, LSS and NSS show the correlation with sequence similarity but the IO dataset with the highest sequence similarity is expected to have higher Sim_{PLD} value in compare with HSS. But as it shown in Figure 5.9 the number of gene pairs with positive Sim_{PLD} value is less than those in HSS. This might have the meaning that the sequence similarity in IO dataset does not necessarily mean that the gene pairs are more functionally similar. It means that they are sequentially similar, but they are not functionally similar.

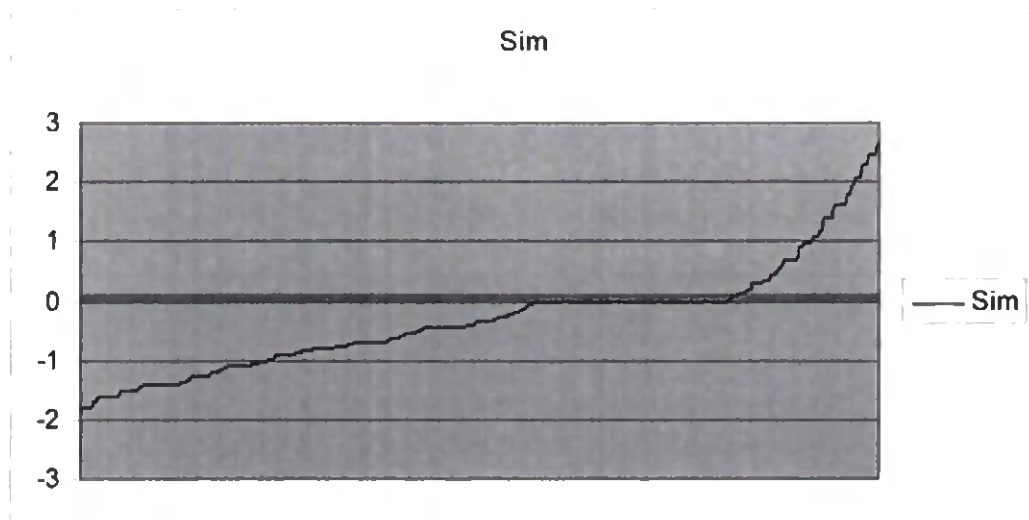


Figure 5.9. Sim_{PLD} in Human-Yeast IO dataset

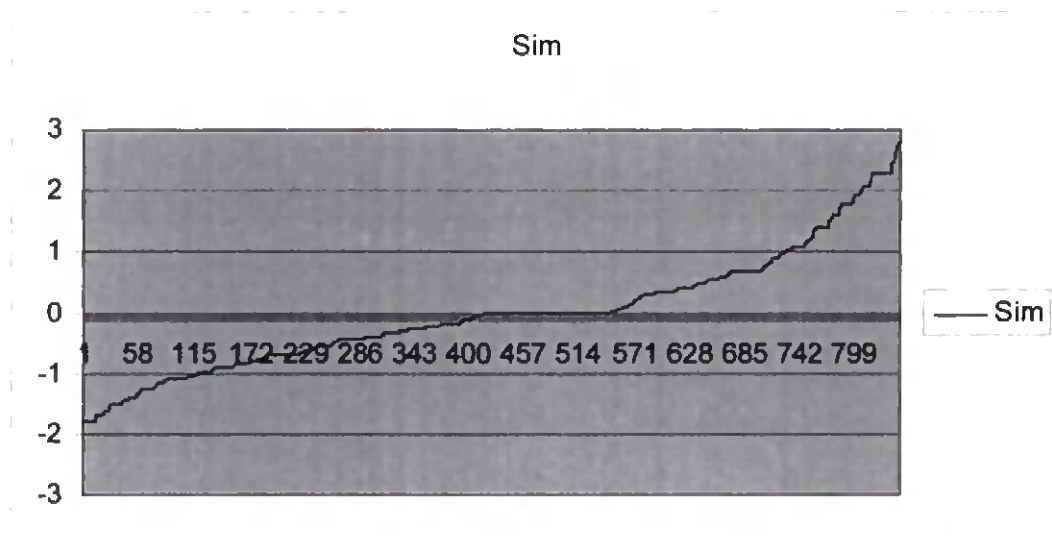


Figure 5.10. Sim_{PLD} in Human-Yeast HSS dataset

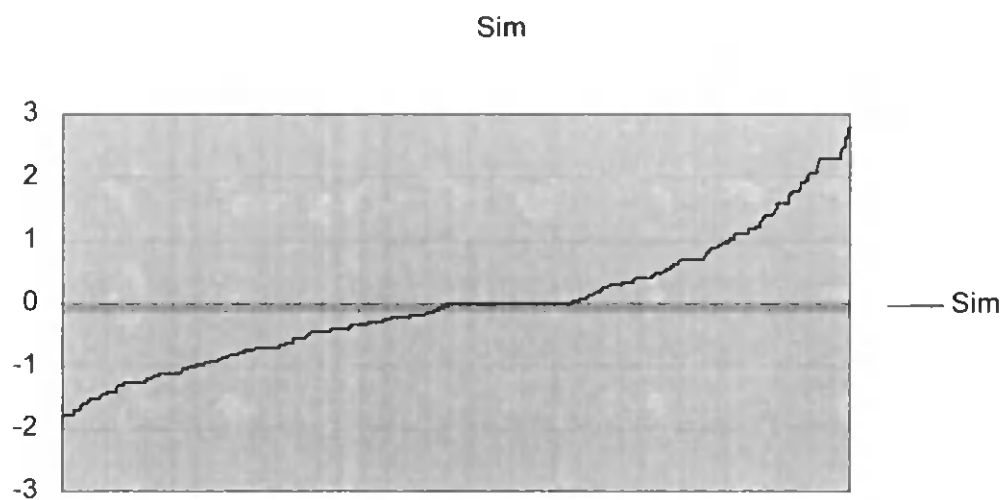


Figure 5.11. Sim_{PLD} in Human-Yeast LSS dataset

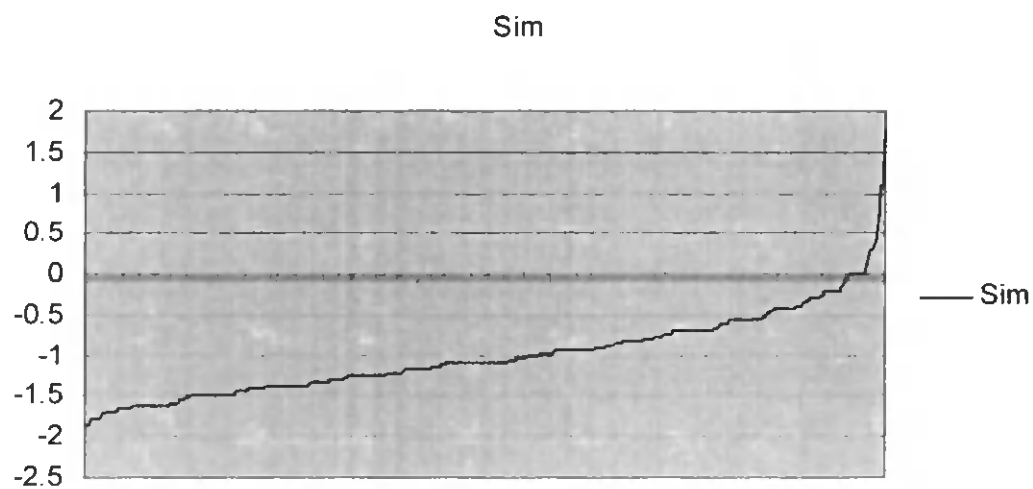


Figure 5.12. Sim_{PLD} in Human-Yeast NSS dataset

Figure 5.13 shows some snapshots of the running program. The program gets the annotation files for the three datasets (FlyBase, Human-Yeast, SGD) as its input in

addition to the excel file that contains the name of genes in each gene pair with its associate E-Value for later comparison and based on what is selected by the user in the first menu of the application, the sequence similarity menu will be populated accordingly. For example, for FlyBase we have two items of HSS and NSS in the sequence similarity menu and for Human-Yeast dataset we have IO, HSS, LSS, NSS items and for SGD dataset we have HSS, LSS and NSS items.

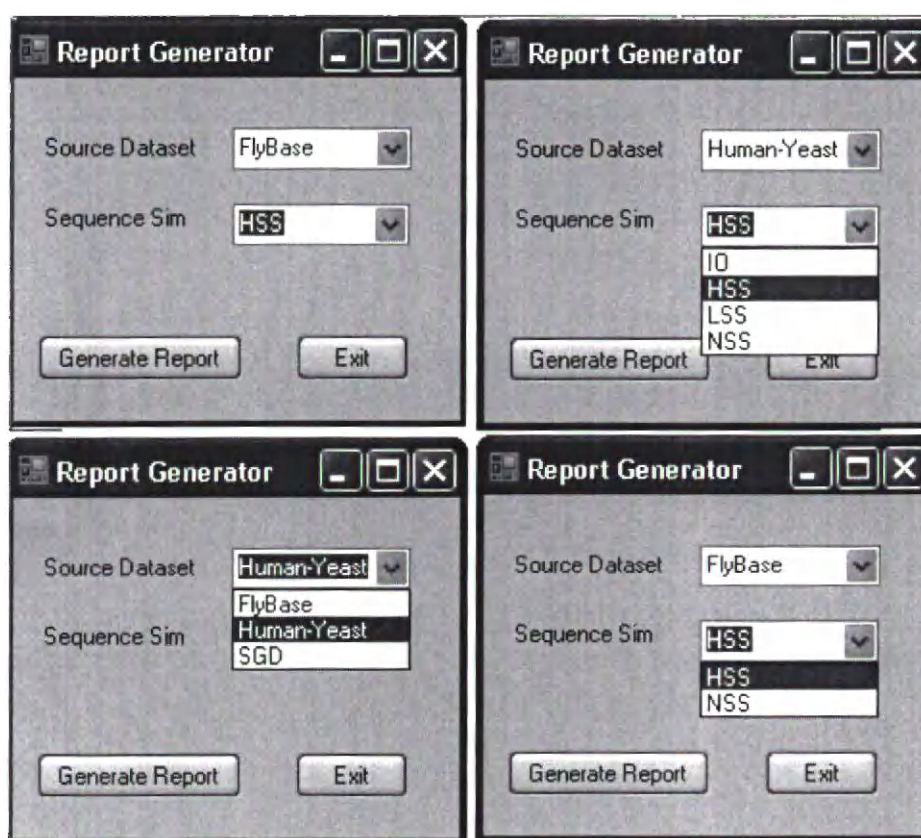


Figure 5.13. Sample of running of the program

The output of the program is the files in excel format that contains the path length between the GO terms and the depth of Least common ancestor of the terms related to each gene in a gene pair. A part of it is shown in Table 5.1. Sample of the output of application

Gene1	Gene2	Evalue	PL_List	depth_list
InR	Ror	1.10E-50	0/1/9/2/1/5/7/10/9/3/7/11/10/8/7/	6/6/1/6/6/3/1/1/4/1/1/1/
Alk	Nrk	1.50E-52	0/1/7/1/0/7/1/2/7/	6/6/6/6/6/5/
htl	dnt	1.00E-25	0/1/7/2/1/7/	6/6/6/6/
Pak	slik	4.80E-43	1/2/7/0/1/7/	6/6/4/4/
Cad96Ca	Nrk	5.90E-42	1/2/7/0/1/7/1/0/7/	5/5/6/6/6/6/
Eph	Cad96Ca	1.10E-41	1/0/1/7/2/1/0/7/3/2/1/7/	5/6/6/5/6/6/5/6/6/
Eph	shark	1.30E-39	0/1/7/1/2/7/2/3/7/	6/6/6/6/6/5/
Ror	Ret	4.10E-38	0/1/7/1/0/7/9/4/7/	6/6/6/6/1/3/
tak1	Tak1l	3.40E-54	3/7/1/7/0/7/	5/5/5/
tak1	CG5169	2.50E-21	2/7/2/7/1/7/	5/4/4/
tak1	CG7097	4.00E-19	1/2/7/3/2/7/2/1/7/	5/5/6/4/6/4/
Pak3	CG11870	7.70E-25	1/7/	6/
CG5169	hpo	1.40E-70	1/0/5/7/	6/4/1/
CG5169	Dsor1	3.00E-40	0/1/4/7/	4/4/5/

Table 5.1. Sample of the output of application

As you see the PL_List contains the list of path length between the GO terms in gene pairs. Consider the first row of the output in table above.

Ror is a gene that is annotated with three GO-terms that are GO:0004713, GO:0004714, GO:0005030. *InR* is a gene that is annotated with four GO-terms that are GO:0004713, GO:0005009, GO:0005520, GO:0005520. The PL_List for these two genes is 0/1/9/2/1/5/7/10/9/3/7/11/10/8/7/. Each three number is separated with a separator for being used later to build a matrix. The first Go term of *InR* which is GO:0004713 is compared with all the three GO terms of *Ror*. Then a matrix can be built from this PL-List. See Table 5.2.

	GO:0004713	GO:0005009	GO:0005520	GO:0005520
GO:0004713	0	2	10	11
GO:0004714	1	1	9	10
GO:0005030	9	5	3	8

Table 5.2. Path Length between Ror and InR GO-terms

The depth_list (6/6/1/6/6/3/1/1/4/1/1/1/) also contain the depth between them. Table 5.3 shows how they are placed inside our matrix.

Gene1: InR	GO:0004713		GO:0005009		GO:0005520		GO:0005520	
Gene2: Ror	PL	depth	PL	depth	PL	depth	PL	depth
GO:0004713	0	6	2	6	10	1	11	1
GO:0004714	1	6	1	6	9	1	10	1
GO:0005030	9	1	5	3	3	4	8	1

Table 5.3. Depth and PL between Ror and InR GO-terms

Then the formula introduces in sections 5.1 and 5.2 is applied to these values to find the semantic similarity between two genes.

5.4. Discussion and Conclusion

We have used the path length along with the depth of LCA of two terms to measure the semantic similarity between GO terms that leads to functional similarity measure between genes. We called this measure Sim_{PLD} (short for **S**imilarity measure based on **PL** and **D**epth) 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).

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.

6. CONCLUSION AND FUTURE WORK

Gene Ontology is the main and most comprehensive resources for research on gene and protein functions and structure. It consists of a set of controlled vocabularies to describe the biology and functions of genes and proteins in any organism [9]. GO annotations capture the available functional information of a gene or protein and can be used as a basis for a measure of functional similarity between genes. Besides the bioinformatics resources that hold data in the form of sequences, these data has represented as scientific natural language which is easier to be modeled and is more readable to human [32]. Gene Ontology is a dynamic evolving project of the GO Consortium in which different sections of the ontology are expanded or reorganized as more biological information becomes available. In this thesis we proposed new similarity techniques for finding gene functional similarity based mainly on the shortest path length between the GO terms annotated for both genes in a given gene pair. For example in chapter 3 we presented a measure based on plain path length that simply considered the distance between the GO terms in gene ontology and then used the average of these distances to find the similarity between the genes. In chapter 4, we considered the number of minimum paths, *nmp*, and the number of common terms, *nct*, in a given gene pair as contributing features in computing the similarity between genes. In chapter 5, we added the depth feature of the least common ancestor of two terms in gene ontology to the measure introduced in

Chapter 3. Then the similarity between the genes was calculated based on the average of this measure between the GO terms. The existing techniques for finding gene functional similarity based on GO rely mainly on the information content of (IC) of the GO terms. PL has never been investigated in the context of GO to estimate the functional similarity between genes based on GO annotation terms. PL has been used extensively as a measure of similarity in the general English domain using, for example, the WordNet ontology [12]. It also has been used in the bioinformatics domain [Rada-1989] [13]; for MeSH [36] ontology and from these applications proved that PL in general can be used as a good indicator of semantic similarity between terms in a given ontology. This research used the PL as one of the most important features in gene ontology.

The proposed measures have been fully implemented and extensively evaluated. In the evaluation, we compared our proposed measure with the BLAST [11] sequence similarity between the sequences of the genes in a given gene pair. We also compared our measure with other IC measures like Resnik based on the human perception [54, 61]. Our evaluation was similar to other research projects in this field like Schlicker et. al [54] that evaluated their work based on the sequence similarity and Wang et. al [61] that compared their measure with Resnik measure [49] based on the justifiability of their result with the human perception. In chapters 3 and 5 we used the first approach of the evaluation while in chapter 4 the second approach has been used.

The experiments were applied on large sets genes from two genomes SGD (*Saccharomyces cerevisiae*) [53] and FlyBase (*Drosophila melanogaster*) [67]. We also tested our measure on a dataset of proteins that Schlicker et. al [54] have used in their experiments.

The experimental results proved the effectiveness of the proposed techniques in measuring the similarity in the GO and gene function domain. See for examples, Figures 3.14, 3.15, 3.16, 3.17 that shows the correlation between the plain path length and sequence similarity. The comparison of our PL_m measure with Resnik and Wang measures shows better or equal estimation of similarity between the genes in several pathways. For example see the Table 4.3. Based on PL_m measure (Chapter 4) we could cluster the genes more accurately than using Resnik measure based on the human perception; see Table 4.4. We also showed , in Chapter 5, that the result of using depth and path length along with each other also correlates very well with the sequence similarity. For example see figures 5.1, 5.2 and 5.3. We applied our plain path length measure to compute the distance between genes based on using terms in molecular function (MF) ontology and terms in biological process (BP) ontology. We found that the MF dataset correlates much better with sequence similarity rather that BP dataset.

6.1. Future Work

In future work of this research we would like to apply path length-based measures to more datasets from different model organisms. For more accurate evaluation we also would like to measure the similarity between the genes using other information sources like the biomedical literature (*e.g.* Medline). We can also use the microarray data analysis to determine expression levels of genes and find the correlation between gene expression data with our semantic similarity measure. Furthermore, we would like to consider the number of distinct paths between two GO terms as a potential feature contributing into

the semantic distance between the genes. In this research we just considered the number of minimum path (nmp) and not the total number of all distinct paths.

Another interesting feature that we would like to study in the future of this research is the effect of the various evidence codes on the performance of the gene similarity measures.

7. 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] Al-Mubaid H and Nguyen HA. (2006) "A Cross-Cluster Approach for Measuring Semantic Similarity Between Concepts." *The 2006 IEEE International Conference on Information Reuse and Integration IRI'06*. Hawaii, USA, 2006.
- [3] Al-Mubaid H. (2006) "Context-Based Technique for Biomedical Term Classification." *Proceedings of the 2006 IEEE Congress on Evolutionary Computation CEC-2006*, Vancouver, BC, Canada, pp.5726-5733, 2006.
- [4] Al-Mubaid H and Nguyen HA. (2006) "Using MEDLINE as Standard Corpus for Measuring Semantic Similarity in the Biomedical Domain." *Proceedings of the IEEE 6th Symposium on Bioinformatics and Bioengineering BIBE06*. pp.315-318, Washington DC USA, 2006.
- [5] Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) "Basic local alignment search tool." *J. Mol. Biol.*, 215:, 403–410. [PubMed].
- [6] Altschul, S. F., Madden, T. L., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein data base search programs". *Nucl. Acids Res.* 25, 3389-3402.
- [7] Amigo Browser. Available:
<http://amigo.geneontology.org/cgi-bin/amigo/go.cgi>
- [8] Arabidopsis Thaliana. Available:
<http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?3769>
- [9] 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.
- [10] Aubry M., Monnier A., Chicault C., Tayrac M., Galibert M.D., Burgun A., and Mosser J. (2006) "Combining evidence, biomedical literature and statistical

dependence: new insights for functional annotation of gene sets”, *BMC Bioinformatics*.

- [11] Blast Tool. Available:
<http://www.ncbi.nlm.nih.gov/blast/>
- [12] Budanitsky A. and Hirst G. (2006) “Evaluating WordNet-based measures of semantic distance,” *Computational Linguistics*, vol.32,1, March 2006.
- [13] Caviedes JE, Cimino JJ. (2004) “Towards the development of a conceptual distance metric for the UMLS.” *Journal of Biomedical Informatics*, vol. 37, no. 2, pp. 77-85, 2004.
- [14] Chang,J., Raychaudhuri,S. and Altman,R. (2001) “Including biological literature improves homology search.” *Pac. Symp. Biocomput.*, 6, 374–383.
- [15] Delfs R., DomsA., Kozlenkov A., and SchroederA. (2004) “GoPubMed: ontology-based literature search applied to Gene Ontology and PubMed” In *Proc. of German Bioinformatics Conference*, Bielefeld, Germany, 2004. LNBI Springer.
- [16] Devos D, Valencia A. (2001) “Intrinsic errors in genome annotation.” *Trends Genet.*
- [17] Devos D & Valencia A. (2000) “Practical limits of function prediction.” *PROTEINS, Structure, Function, and Genetics* 41, 98-107.
- [18] Dolan M. E., Ni L., Camon E. and Blake J. A. (2005) “A procedure for assessing GO annotation consistency”, *Bioinformatics*.
- [19] Expert Protein Analysis System. Available:
<http://expasy.org/spot/>
- [20] Finn RD, Mistry J, Schuster-Boeckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer ELL, Bateman A (2006) “Pfam: clans, web tools and services.” *Nucleic Acids Res.*
- [21] Fox, Michael Allen (1986) “The Case for Animal Experimentation: An Evolutionary and Ethical Perspective.” Berkeley and Los Angeles, California: *University of California Press*
- [22] Gene Ontology. Available:
www.geneontology.org
- [23] 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.

- [24] Jiang T. and Kcating A.M. (2005) "AVID: An integrative framework for discovering functional relationships among proteins", *BMC Bioinformatics*.
- [25] Khabiri E. (2007) "A Preliminary study of Correlation between depth and Path Length of GO nodes with Gene Sequence Similarity." *IEEE 7 International Conference on BioInformatics and BioEngineering BIBE07*, Boston, Massachusetts USA, 2007
- [26] Khabiri E., Al-Mubaid H. (2007) "A path length method for gene functional similarity using GO annotations." *16th International Conference on Software Engineering and Data Engineering SEDE 2007*. Las Vegas, Nevada USA, 2007.
- [27] Khatri P., Done B., Rao A., Done A. and Draghici S. (2005) "A semantic analysis of the annotations of the human genome", *Bioinformatics*.
- [28] 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
- [29] 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.
- [30] Lin, D. (1998) "An information-theoretic definition of similarity." *In Proc. of the Int'l Conference on Machine Learning*.
- [31] Lord P. W., Stevens R. D., Brass A. and Goble C. A. (2003) "Semantic Similarity Measures as Tools for Exploring the Gene Ontology." *Pac Symp Biocomput.*
- [32] Lord P. W., Stevens R. D., Brass A., Goble C. A. (2002) "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation." *Bioinformatics*, 19, pp. 1275-1283.
- [33] 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.
- [34] McGinnis S., Madden T. L. (2004) "BLAST: at the core of a powerful and diverse set of sequence analysis tools." *Nucleic Acids Res.*
- [35] MEDLINE. Available:
<http://www.cas.org/ONLINE/DBSS/medliness.html>
- [36] MeSH. Available:
<http://www.nlm.nih.gov/mesh/meshhome.html>

- [37] Miller G. A. (1995) "WordNet: A Lexical Database for English," *Comm. ACM*, vol. 38, no. 11, pp. 39-41.
- [38] Chagoyen M., Carmona-Saez P., Gil C., Carazo J. M., Pascual-Montano A. (2006) "A literature-based similarity metric for biological processes." *BMC Bioinformatics*.
- [39] Mouse Genome Informatics (MGI). Available:
<http://www.informatics.jax.org/>
- [40] Nguyen H., Al-Mubaid H. (2006) "New Semantic Similarity Techniques of Concepts applied in the biomedical domain and WordNet." MS Thesis, University of Houston Clear Lake, Houston, TX USA, 2006.
- [41] Nguyen H. A., Al-Mubaid H. (2006) "New Ontology-based Semantic Similarity Measure for the Biomedical Domain." *Proceedings of the IEEE conference on Granular Computing GrC-2006*. pp. 623-628, 2006.
- [42] Nguyen H. A., Al-Mubaid H. (2006) "A Combination-Based Semantic Similarity Measure Using Multiple Information Sources." *Proc. of the 2006 IEEE Int'l Conference on Information Reuse and Integration IRI'06*. Hawaii, USA, 2006.
- [43] Pan H., Zuo L., Choudhary V., Zhang Z., Leow S. H., Chong F.T., Huang Y., Wui V. Siong Ong, Mohanty B., Tan S.L., Krishnan S. P. T., Bajic V. B. (2004), "Dragon TF Association Miner: a system for exploring transcription factor associations through text-mining", *Nucleic Acids Res.*
- [44] Pedersen T., Pakhomov S. V., Patwardhan S., Chute C. G. (2006) "Measures of Semantic Similarity and relatedness in the biomedical domain." *Journal of Biomedical Informatics*.
- [45] PubMed. Available:
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed>
- [46] 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.
- [47] Remm M., Storm C. E., Sonnhammer E. L. L. (2000) "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." *J Mol Biol.* 2001;314:1041-52. doi: 10.1006/jmbi.2000.5197. [PubMed]
- [48] Resnik, P. (1995) "Using Information Content to Evaluate Semantic Similarity in a Taxonomy." *Proc 14th Int'l Joint Conf Artificial Intelligence*. pp. 448-453.

- [49] Resnik P. (1999) "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language." *J Artif Intell Res.* 1999;11:95–130.
- [50] Sealfon R. S., Hibbs M. A., Huttenhower C. E., Myers C. L., Troyanskaya O. G. (2006) "GOLEM: an interactive graph-based gene-ontology navigation and analysis tool" *BMC Bioinformatics*
- [51] Sevilla Jose' L., Segura V., Podhorski A., Guruceaga E., Mato Jose' M., Marti'nez-Cruz L. A., Corrales F. J., Rubio A. (2005). "Correlation between Gene Expression and GO Semantic Similarity" *IEEE/ACM Transaction on computational biology and bioinformatics*, vol.2, No. 4.
- [52] S.Cerevisiae WU-BLAST2 Search. Available:
<http://seq.yeastgenome.org>
- [53] Saccharomyces Genome Database. Available:
<http://www.yeastgenome.org/>
- [54] Schlicker A., Domingues FS., Rahnenführer J., Lengauer T. (2006). "A new measure for functional similarity of gene products based on Gene Ontology." *BMC Bioinformatics*.
- [55] Sohler F., Hanisch D., Zimmer R. (2004), "New methods for joint analysis of biological networks and expression data." *Bioinformatics*.
- [56] Speer. N., Spieth, C., Zell A. (2004) "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)*.
- [57] Tatusova T. A., Madden T. L. (1999) "BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences." *FEMS Microbiol Lett*.
- [58] The International Classification of Diseases, 9th Revision, Clinical Modification" (ICD-9-CM) Available:
<http://icd9cm.chrisendres.com/>
- [59] Trypanosoma brucei Genome Project. Available:
http://www.sanger.ac.uk/Projects/T_brucei/
- [60] UMLS. Available:
<http://www.nlm.nih.gov/research/umls/>

- [61] Wang J. Z., Du Z., Payattakool R., Yu P. S., Chen C. F. (2007) "A new method to measure the semantic similarity of GO terms." *Bioinformatics*.
- [62] WHO Media centre (2006) "Fact sheet N°259: African trypanosomiasis or sleeping sickness"
- [63] WormBase. Available:
<http://www.wormbase.org/>
- [64] 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.
- [65] Zhang P., Zhang J., Sheng H., Russo J., Osborne B., Buetow K. (2006) "Gene functional similarity search tool (GFSST)" *BMC Bioinformatics*.
- [66] Zhaotao C., Xizeng M., Songgang L., Liping W. (2006) "Genome comparison using Gene Ontology (GO) with statistical testing", *BMC Bioinformatics*
- [67] FlyBase. Available:
<http://flybase.bio.indiana.edu/>
- [68] Deonier R. C., Tavaré S., Waterman M. S. (2005) "Computational Genome Analysis, An Introduction" Springer, 2005
- [69] Verspoor K, Cohen J, Mniszewski S, Joslyn C. (2006) "A Categorization Approach to automated ontological function Annotation". *Protein Science* vol. 15, pp. 1544-1549.
- [70] Azuaje F., Wang H., Zheng H., Bodenreider O., Chesneau A. (2006) "Predictive Integration of gene ontology driven similarity and functional interaction" *Proc. of IEEE International Conference on Data Mining (ICDM) 2006*,

APPENDIX A: IMPLEMENTATION DETAILS

Here we want to show parts of the program that is developed to calculate the path length between the genes. The detailed of the program is as the following: We used linked list as the structure of storing the GO nodes in computing the shortest path length (Please refer to Sec. 2.3 and Figure 3.2 in Chapter 3.). Each cell in the linked-list has the following properties:

```
class CellArray
{
    String _goID;
    String _goParent;
    int _goPathLen;
    String _goParent2;
    int _goPathLen2;
    int _distance;
}
```

All the properties are private and we used setter and getter to access them. Like:

```
public String GoID
{
    get { return _goID; }
    set { _goID = value; }
}
public String GoParent
{
    get { return _goParent; }
    set { _goParent = value; }
}
public int GoPathLen
{
    get { return _goPathLen; }
    set { _goPathLen = value; }
}
```

We use a method of *getParent* to get all the parents of a node. Details are as the following:

```
private ArrayList getParents(String termID)
{
    ArrayList is_a_ArrayList = new ArrayList();
    XmlDocument goDoc = new XmlDocument();
    String GOPath = Application.StartupPath + "\\summerizedGO.xml";
    goDoc.Load(GOPath);
    XmlElement root = goDoc.DocumentElement;
    XmlNodeList goList = root.GetElementsByTagName("term");

    IEnumerator inum = goList.GetEnumerator();
    while (inum.MoveNext())
    {
        XmlNode node = (XmlNode)inum.Current;
        String temp = node.Attributes.GetNamedItem("about").Value;
        int startTrim = temp.IndexOf('#') + 1;
        String term = temp.Substring(startTrim);
        //if the term was the same as the input term
        if (termID == term)
        {
            XmlNodeList list = node.ChildNodes;
            IEnumerator ienum = list.GetEnumerator();

            while (ienum.MoveNext())
            {
                XmlNode currentChild = (XmlNode)ienum.Current;
                if (currentChild.Name == "is_a")
                {
                    String templ =
currentChild.Attributes.GetNamedItem("resource").Value;
                    int startTrim1 = templ.IndexOf('#') + 1;
                    String parent = templ.Substring(startTrim1);
                    is_a_ArrayList.Add(parent);
                }
            }
        }
    }
    return is_a_ArrayList;
}
```

This part of the code called *getDistance* get two terms and returns the number of minimum paths and the distance between the two terms.

```

public void getDistance(String term1, String term2, ref double
distance, ref int nmp)
//number of minimum path
{
    distance = -1;
    nmp = 0;
    if (term1 == term2)
    {
        distance = 0;
        return;
    }
    int counter = 0;
    int minDistance = 100;
    ArrayList list = new ArrayList();
    //contains terms + the parents of each terms + the
    //parents of the each node that is being added

    //calculate goID, goParent, goPathLen
    CellArray cell1 = new CellArray(term1);
    cell1.GoParent = term1;
    list.Add(cell1);

    CellArray cell2 = new CellArray(term2);
    cell2.GoParent = term2;
    list.Add(cell2);

    CellArray currentCell = (CellArray)list[counter];
    //counter and currentcell points to a cell that its parents should
be found

    //minDistance keeps the minimum distance between the two GO nodes.
    while (list.Count > counter && currentCell.GoPathLen < minDistance)
    {
        currentCell = (CellArray)list[counter];
        ArrayList parents = getParents(currentCell.GoID);
        //gets the first upper level parents
        bool found = false;
        for (int i = 0; i < parents.Count; i++)//for i{
            found = false;
            String parent = parents[i].ToString();
            IEnumerator ienum1 = list.GetEnumerator();
            while (ienum1.MoveNext())//to compare from the beginning of
the list
            {
                //see if there exist the same GOID from before.
                {
                    CellArray currentEnum =
(CellArray)ienum1.Current;//checker from beginning to end
                    if (currentEnum.GoID != parent)//not found any goID that added
before
                    {
                        // ienum1.MoveNext();

```



```

        }
        else//if current.GOID == parents[i]
        {if (currentEnum.GoParent == currentCell.GoParent)
//if l1=l1//come from the same path
{
found = true;
ienum1.MoveNext();
} else//if l1!=l2 //come to the same LCS: not the same path
{
found = true;
if (currentEnum.GoParent2 == "" && currentEnum.GoPathLen2 == 0)
{
currentEnum.GoParent2 = currentCell.GoParent;
currentEnum.GoPathLen2 = currentCell.GoPathLen + 1;
currentEnum.Distance = currentEnum.GoPathLen +
currentEnum.GoPathLen2;
}
if (currentEnum.Distance < minDistance)
minDistance = currentEnum.Distance;
} //end else
} //end else
} //end while
if (found == false)//if not found the GO add it to the list.
{
CellArray cell = new CellArray(parent);
cell.GoPathLen = currentCell.GoPathLen + 1;
cell.GoParent = currentCell.GoParent;
list.Add(cell);
}
}
counter++;
} //while and
distance = minDistance;
for (int i = 0; i < list.Count; i++)
{
CellArray current = (CellArray)list[i];
if (current.Distance == minDistance)
{
nmp++;
}
}
} //calculate number of minimum distance
}

```

Here is the code for getting the name of organism and the sequence simialrity of the dataset and finding the similarity between the genes inside the dataset.

```

public static void readAnnotation(String source, String seqSim){

String path1 = "\\Variation1\\" + source + "_" + seqSim +
"_variation1.csv";
String file1 = Application.StartupPath + path1;
StreamReader reader = File.OpenText(file1);

String path2 = "\\All_Output\\" + source + "_" + seqSim + "_All.csv";
StreamWriter writer = File.CreateText(Application.StartupPath + path2);

writer.WriteLine("Gene1,Gene2,Evalue,PL_List,NP_List,Depth_List,simGO,S
im");

String fileLine = reader.ReadLine();
while (!reader.EndOfStream)
{
    String gene1 = "";String gene2 = "";String evalue = "";
    String pl_list = "";String np_list = "";String depth_list = "";
    String plv1 = "";

    gene1 = fileLine.Substring(0, fileLine.IndexOf(","));
    fileLine = fileLine.Remove(0, fileLine.IndexOf(",")+1);

    gene2 = fileLine.Substring(0, fileLine.IndexOf(","));
    fileLine = fileLine.Remove(0,fileLine.IndexOf(",") + 1);

    evalue = fileLine.Substring(0,fileLine.IndexOf(","));
    fileLine = fileLine.Remove(0, fileLine.IndexOf(",") + 1);

    plv1 = fileLine.Substring(0, fileLine.IndexOf(","));
    fileLine = fileLine.Remove(0, fileLine.IndexOf(",") + 1);
    pl_list = fileLine.Substring(0, fileLine.IndexOf(","));
    fileLine = fileLine.Remove(0, fileLine.IndexOf(",") + 1);
    np_list = fileLine.Substring(0, fileLine.IndexOf(","));

    fileLine = fileLine.Remove(0, fileLine.IndexOf(",") + 1);
    depth_list = fileLine.Substring(0);
    //create pl_list ArrayList
    ArrayList PL_ArrayList = new ArrayList();
    String[] PL_List = pl_list.Split('/');
    for (int i = 0; i < PL_List.Length - 1; i++)
    {
        PL_ArrayList.Add(PL_List[i]);
    }
    //create np_list ArrayList
    ArrayList NP_ArrayList = new ArrayList();//1/2/3/4/5/
    String[] NP_List = np_list.Split('/');
    for (int i = 0; i < NP_List.Length - 1; i++)
    {
        NP_ArrayList.Add(NP_List[i]);//1,2,3,4,5
    }
    //create depth_list ArrayList
    ArrayList Depth_ArrayList = new ArrayList();//1/2/3/4/5/
    String[] Depth_List = depth_list.Split('/');
    for (int i = 0; i < Depth_List.Length - 1; i++)
    {

```

```

        Depth_ArrayList.Add(Depth_List[i]); //1,2,3,4,5
    }

    //calculate similarity for GOs
    //log(Depth(LCA(gox, goy)/maxDepth)-log(PL(gox,
goy)/2*maxDepth)
    String simGOString = "";
    double simGO = 0; //similarity measure for GO terms
    double sim = 0; //similarity measure for Genes

    for (int i = 0; i < PL_ArrayList.Count; i++)
    {

        double depth = Double.Parse(Depth_ArrayList[i].ToString());
        double PL = Double.Parse(PL_ArrayList[i].ToString());
        if (PL != 0)
        {
            double aa = PL / (2 * maxDepth);
            double bb = (maxDepth-depth)/maxDepth;
            double cc = aa*bb+1;
            double dist = Math.Log(cc, 2);

            simGO = 1-dist;
        }
        else
        {
            simGO = 1;
        }

        simGO = Math.Round(simGO, 2);
        sim += simGO;
        simGOString += simGO.ToString() + " ";
    }
    sim = sim / PL_ArrayList.Count;
    sim = Math.Round(sim, 2);
    writer.WriteLine(gene1 + "," + gene2 + "," + evalue + "," + pl_list +
    "," + np_list + "," + depth_list + "," + simGOString + "," + sim);
    writer.AutoFlush = true;
    fileLine = reader.ReadLine();

    }
    writer.Close();
    reader.Close();
}

```

APPENDIX B: SAMPLE OUTPUT AND RESULT TABLES

Here we show some parts of the output generated from by *PathLengthCalculator* application. All the results can not be shown in here. This is only a small part of it. The output of the program contains the name of the genes that are compared with each other. *PL_List* contains the plain path length between the GO terms associated with a given gene pair. *NP_List* contains the number of minimum paths between the GO terms. *Depth_List* contains the depth of the least common ancestor (LCA) of the two terms. If there are more than 1 term related to one gene in a gene pair then we have several PLs in our *PL_List*, several NPs in our *NP_List* and several depths in our *Depth_List* that are separated by a “slash”.

The following output is for Human-Yeast-IO dataset:

Gene1	Gene2	PL LHA	NP LHA	Depth LHA	SimGO	Sim
Q05636	Q06265	10/0/	1/0/	1/7/	0.61; 1;	0.8
P07347	P41227	10/1/	2/1/	1/7/	0.61; 0.97;	0.79
Q05506	Q96FU5	11/0/4/	1/0/1/	1/7/3/	0.58; 1; 0.85	0.81
P22438	Q96GW9	11/0/4/	1/0/1/	1/7/3/	0.58; 1; 0.85	0.81
P53043	P53041	8/0/	1/0/	1/8/	0.68; 1;	0.84
P32906	Q9UKM7	11/0/	1/0/	1/8/	0.58; 1;	0.79
P36017	P20339	9/0/	1/0/	1/8/	0.64; 1;	0.82
P06245	P17612	13/0/11/2/	2/0/2/1/	1/8/1/6/	0.51; 1; 0.58	0.76
P22137	Q00610	0/	0/	2/	1;	1
Q03940	Q9Y265	8/	1/	2/	0.7;	0.7
P19882	P10609	8/	1/	2/	0.7;	0.7
P41921	P00390	0/	0/	2/	1;	1
P32604	Q60825	11/	1/	2/	0.6;	0.6
P23615	Q7KZB5	3/	1/	2/	0.88;	0.88
P38888	Q92611	2/	1/	2/	0.92;	0.92
P14743	P30419	6/	2/	2/	0.77;	0.77
P53941	Q96G21	5/	1/	2/	0.81;	0.81
Q02939	Q92759	3/	1/	2/	0.88;	0.88
P16120	Q86YJ6	4/	1/	2/	0.84;	0.84
P36007	Q9GZT4	7/	1/	2/	0.73;	0.73
P47039	Q16773	7/	1/	2/	0.73;	0.73
P53686	Q9NTG7	8/	1/	2/	0.7;	0.7
P06105	Q00341	3/	1/	2/	0.88;	0.88
P38152	P53007	8/	1/	2/	0.7;	0.7
P38702	P16260	8/	1/	2/	0.7;	0.7
P23968	Q99437	4/	1/	2/	0.84;	0.84
Q03529	Q96DK1	1/	1/	2/	0.96;	0.96
P40556	Q9H2D1	7/	1/	2/	0.73;	0.73
P53731	O15144	1/	1/	2/	0.96;	0.96
P39706	P33316	13/	2/	2/	0.54;	0.54
Q05787	Q8VWH5	9/	1/	2/	0.67;	0.67
P36070	P23193	3/	1/	2/	0.88;	0.88

Table 0.1. Human-Yeast-IO dataset

The following output is for SGD HSS dataset:

Gene1	Gene2	Expr	BLUP	Number	Depth	SRGO	Sim
ANP1	MNN9	3.90E-164	1/0/	1/0/	6/6/	0.97; 1;	0.98
ABF2	KR1	5.40E-09	0/1/	0/1/	4/4/	1; 0.97;	0.98
PMC1	NEO1	7.10E-127	4/0/	2/0/	2/5/	0.84; 1;	0.92
ADH3	YAL061W	2.20E-07	1/	1/	5/	0.97;	0.97
ADH1	SOR2	2.20E-07	1/	1/	5/	0.97;	0.97
ADP1	MDL1	8.50E-114	1/3/0/	1/1/0/	3/6/3/	0.96; 0.93; 1;	0.96
AFT1	AFT2	2.30E-168	0/3/	0/1/	3/2/	1; 0.88;	0.94
PMC1	DRS2	1.50E-70	4/1/	2/1/	2/4/	0.84; 0.97;	0.9
PMC1	DNF3	1.50E-70	4/1/	2/1/	2/4/	0.84; 0.97;	0.9
PMC1	DNF2	1.50E-70	4/1/	2/1/	2/4/	0.84; 0.97;	0.9
PMC1	DNF1	3.00E-70	4/1/	2/1/	2/4/	0.84; 0.97;	0.9
PMC1	PMR1	1.10E-193	0/9/2/	0/3/1/	5/1/5/	1; 0.64; 0.94;	0.86
PMC1	ENA2	6.10E-104	2/2/1/	1/1/1/	4/4/4/	0.93; 0.93; 0.97;	0.94
PMC1	YOR291W	3.50E-53	5/	3/	1/	0.79;	0.79
ALD3	ALD4	6.00E-97	2/	1/	5/	0.94;	0.94
ALD3	UGA2	2.70E-94	2/	1/	5/	0.94;	0.94
PMC1	ENA1	2.40E-93	2/2/	1/1/	4/4/	0.93; 0.93;	0.93
ACO1	LYS4	3.90E-93	2/	1/	5/	0.94;	0.94
ADP1	PDR12	1.70E-83	3/1/	1/1/	2/3/	0.88; 0.96;	0.92
AOS1	UBA1	1.80E-72	2/	1/	3/	0.93;	0.93
ADH2	XYL2	1.70E-06	2/	1/	4/	0.93;	0.93
ACE2	PZF1	1.60E-06	2/	1/	2/	0.92;	0.92
ACT1	ARP8	1.60E-55	2/7/	1/3/	1/1/	0.91; 0.72;	0.82
ACT1	ARP5	9.20E-40	7/	3/	1/	0.72;	0.72
AHA1	PMR1	6.50E-41	8/7/8/7/	3/2/4/3/2/4/3/2/4/	1/1/1/1/1/1/2/1/	0.68; 0.72; 0.68; 0.72; 0.75; 0.72; 0.68; 0.81; 0.68;	0.72
ACE2	AZF1	3.00E-08	5/0/	1/0/	1/3/	0.79; 1;	0.9
ACC1	HFA1	2.50E-07	0/5/	0/1/	6/3/	1; 0.82;	0.91
PMC1	PCA1	1.00E-62	10/2/1/2/	2/1/1/1/	1/5/4/4/	0.61; 0.94; 0.97; 0.93;	0.66
AHA1	PMA2	2.00E-43	8/7/8/	3/3/3/	1/1/1/	0.68; 0.72; 0.68;	0.69
AHA1	PMA1	6.70E-42	8/7/8/	3/3/3/	1/1/1/	0.68; 0.72; 0.68;	0.69
AHA1	ENA1	3.80E-40	8/8/7/7/	3/3/3/3/3/	1/1/1/1/1/1/	0.68; 0.68; 0.72; 0.72; 0.68; 0.68;	0.69

Table 0.2. SGD HSS dataset

The following output is for FlyBase NSS dataset:

Gene1	Gene2	Exp	Recall	NPQ	Deletion	AmGO	Sim
ama	beta4GalN	1	7/	2/	1/	0.72;	0.72
CG11870	betaTub97	1	8/	2/	1/	0.68;	0.68
Fak56D	3L-B	1	6/7/	2/2/	1/1/	0.75; 0.72;	0.74
Cad96Ca	3L-B	1	5/5/	2/2/1/	1/1/1/	0.79; 0.75;	0.78
elik	3N5	1	3/4/	1/1/	1/1/	0.87; 0.83;	0.85
Fak56D	3N5	1	6/7/	2/2/	1/1/	0.75; 0.72;	0.74
Cad96Ca	3N5	1	5/5/	2/2/1/	1/1/1/	0.79; 0.75;	0.78
Iti1	alpha4GT1	1	8/10/	2/3/	3/1/	0.72; 0.61;	0.66
Ror	alpha4GT1	1	8/9/9/	2/3/2/	3/1/1/	0.72; 0.64;	0.67
shark	alpha4GT1	1	8/9/	2/2/	3/3/	0.72; 0.69;	0.7
Fak56D	alpha4GT1	1	8/9/	2/2/	3/3/	0.72; 0.69;	0.7
Cad96Ca	alpha4GT1	1	7/8/9/	2/2/3/	3/3/1/	0.75; 0.72;	0.7
elpr	alpha4GT1	1	7/11/8/10/	2/3/2/2/	3/3/3/1/	0.75; 0.63;	0.68
CG3277	alpha-Cat	1	10/9/10/	2/2/2/	1/1/1/	0.61; 0.64;	0.62
Iti1	alpha-Cat	1	10/9/10/10/9/10/	2/2/2/1/1/1/	1/1/1/1/1/1/	0.61; 0.64;	0.62
Cad96Ca	3L-211-27	1	5/5/	2/2/1/	1/1/1/	0.79; 0.75;	0.78
Iti1	alpha4GT1	1	8/10/	2/3/	3/1/	0.72; 0.61;	0.66
Ror	alpha4GT1	1	8/9/9/	2/3/2/	3/1/1/	0.72; 0.64;	0.67
shark	alpha4GT1	1	8/9/	2/2/	3/3/	0.72; 0.69;	0.7
Fak56D	alpha4GT1	1	8/9/	2/2/	3/3/	0.72; 0.69;	0.7
Cad96Ca	alpha4GT1	1	7/8/9/	2/2/3/	3/3/1/	0.75; 0.72;	0.7
elpr	alpha4GT1	1	7/11/8/10/	2/3/2/2/	3/3/3/1/	0.75; 0.63;	0.68
CG3277	alpha-Cat	1	10/9/10/	2/2/2/	1/1/1/	0.61; 0.64;	0.62
Iti1	alpha-Cat	1	10/9/10/10/9/10/	2/2/2/1/1/1/	1/1/1/1/1/1/	0.61; 0.64;	0.62
shark	alpha-Cat	1	10/9/10/11/10/11/	2/2/2/2/2/2/	1/1/1/1/1/1/	0.61; 0.64;	0.6
Fak56D	alpha-Cat	1	10/9/10/11/10/11/	2/2/2/2/2/2/	1/1/1/1/1/1/	0.61; 0.64;	0.6
Eph	alpha-Cat	1	10/9/10/9/8/9/10/9/10/	2/2/2/1/1/1/1/1/1/	1/1/1/1/1/1/1/1/1/	0.61; 0.64;	0.63
elpr	alpha-Cat	1	9/8/9/9/8/9/10/9/10/4/3/4/	2/2/2/1/1/1/2/2/2/1/1/1/	1/1/1/1/1/1/1/1/3/3/3/	0.64; 0.68;	0.7
CG3277	alpha-Est1	1	9/	2/	2/	0.67;	0.67
shark	alpha-Est1	1	9/10/	2/2/	2/2/	0.67; 0.63;	0.65
shark	thetaTry	1	10/11/	2/2/	2/2/	0.63; 0.6;	0.62
Fak56D	thetaTry	1	10/11/	2/2/	2/2/	0.63; 0.6;	0.62
Cad96Ca	thetaTry	1	9/10/11/	2/2/1/	2/2/1/	0.67; 0.63;	0.63

Table 0.3. FlyBase NSS dataset