Computing Gene Functional Similarity Using Combined Graphs

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

Hisham Al-Mubaid
University of Houston- Clear Lake, Houston, TX USA
Hisham@uhcl.edu

Said Bettayeb
University of Houston- Clear Lake, Houston, TX USA
Bettayeb@uhcl.edu

ABSTRACT

The Gene Ontology has been used extensively for measuring the functional similarity among genes of various organisms. All the existing gene similarity methods use either molecular function or biological process taxonomies in computing gene similarity. In this paper, we apply an algorithm for combining graphs to connect the molecular function (F) and biological process (P) taxonomies into one FP taxonomy graph. We then measure the functional similarity of two genes using the resulting FP graph with path length function. The two aspects of GO, molecular function and biological process, are combined by connecting F nodes with P nodes using gene ontology annotation, GOA, databases. By combining two GO graphs, we can have more comprehensive way to explore the functional relationships between genes. We conducted the evaluation on a dataset of OMIM disease phenotypes to estimate the similarity of disease proteins from various diseases.

Categories and Subject Descriptors

E.1 [Graph and Networks]: trees. J.3 [Computer and Applications]: Life and Medical Sciences – Medical information systems.

General Terms

Algorithms, Experimentation.

Keywords

Gene functional similarity, gene ontology, ontology integration.

1. INTRODUCTION

The gene ontology is used in all research related to gene and protein functional similarity [1, 6, 7]. It is effectively the central source of information on functions, processes, and localizations of gene products [7]. The gene ontology (GO) have been studied and investigated extensively for decades, for computing gene similarity and relationships among gene products in various organisms [10, 11, 12]. Moreover, most of the approaches for discovering new gene functions and identifying gene disease associations are also based on GO. Gene ontology is a structured vocabulary of gene functions and related information at the molecular level, biological process and cellular localization. Therefore, GO is composed of three orthogonal sub-ontologies: *molecular function* (F), *biological process* (P), and *cellular component* (C). The existing techniques for measuring the functional similarity of genes and proteins rely on the gene ontology annotation (GOA) terms of the target genes from either molecular function (F) or biological process (P) independently as there are no links inter-ontology relationships between the molecular function and biological process ontologies [6, 10, 12]. Table 1 includes an example of GOA annotation terms for four genes. In this paper, we want to explore the functional relationship between two genes given their GOA terms from the *molecular function* (F) and *biological process* (P) graphs combined [1, 6, 12]. However, there has not been any work that explores the functional relationships between gene products in terms of their F and P annotation terms combined. For that, we introduce an algorithm for combining two graphs based on given shared knowledge sources. The algorithm assumes that the graphs represent knowledge sources from certain domain. It connects the nodes from the two disconnected graphs, with disjoint vertex sets, based on a given tuple set that summarizes domain knowledge from the same domain of the graphs.

Usually, a graph represents an aspect or a branch (e.g., molecular function) of knowledge base (e.g., gene ontology) and the edges represent the relationships (e.g. *is_a*) between the knowledge terms or entities which are represented by the nodes. The set of tuples, which is the shared domain knowledge, is used to connect the nodes from the two graphs, see example in Figure 1. Such edges are called *bridge-edges*, see Figure 2. Figure 2 shows a bridge-edge e_{pq} that connects node t_p from one graph to node t_q in another graph. The path between two nodes in two graphs can pass through one or more *bridge-edges*. Figure 3 shows two graphs, G1 and G2, connected by four *bridge-edges*. This method of connecting two graphs, or two knowledge sources, will enable us to explore and understand the degree of relatedness of the nodes in two graphs based on a given domain knowledge summarized in the set of tuples.

2. COMBINING ONTOLOGY

In this section, we explain how two ontology graphs (e.g., molecular function and biological process of GO) will be combined based on a shared set of knowledge tuples; see example set of tuples in Figure 1. In general, an ontology graph consists of nodes where each node is a *term*, and the edges depict the relationships, *e.g. is_a*, between the nodes. The GO graphs are directed acyclic graph (DAG). A DAG is a graph that has no cycles and each edge has a direction. A GO graph like the molecular function (F) graph has *root* node, *internal* nodes, and *leaf* nodes [7]. As we go down the graph from the root towards the leaves, the nodes, or *terms*, become more specific and the root is the most general knowledge term. *Connecting nodes from two graphs:* The nodes from two graphs will be connected using shared domain knowledge presented as a set of tuples TP ={P1, .., Pn}.

SAC’12, March 26-30, 2012, Riva del Garda, Italy.

P1 = {tc, tg}
P2 = {tc, td, tg}
P3 = {te, th, ti}
P4 = {ta, tc, te}
P5 = {te, th}

Figure 1: A small tuple set represented in the example graph in Figure 3.

Each tuple Pi $_{(i=1, .., n)}$ is a set of terms Pi={t1,…,tk}. Moreover, each tuple Pi includes terms (nodes) from both graphs (e.g., F and P). That is, Pi={…,t_p,….,t_q,…} where t_p is a node in graph F whereas t_q is a node in graph P. Since the pair (t_p, t_q) is in a single tuple, we draw an edge e_{pq} to connect nodes t_p and t_q , i.e., connecting graph F with graph P. The edge e_{pq} is called *bridge-edge* because it has one endpoint in F and the other in P, as shown in Figure 2 and Figure 3. Figure 3 shows four bridge-edges connecting G1 with G2. These bridge-edges were created based on the tuple set in Figure 1. Notice here the term *bridge-edge* is used differently than its use in graph theory (in graph theory, a *bridge edge* is an edge that its removal will disconnect the graph).

The GOA databases contain a huge amount of gene annotation information for large number of model organisms [9]. If we used GOA database as a shared knowledge source, we can connect F nodes with P nodes. Table 1 contains biological process GOA terms (P terms) and molecular function terms (F terms) for a sample of four genes. For example, if a gene product g_i is annotated with F term t_f and P term t_p then we can assume that this is one (inter-ontology) relation between t_f and t_p nodes. Each link between F and P has a *b_count* (bridge count) value as follows:

b_count(tp, tq) = number of genes in the GOA annotated with both terms tp and tq …(1)

Also each bridge-edge has a weight *w()* defined as follows:

$$w(epq) = \frac{\text{b\_count(tp,tq)}}{\text{max(b\_count())}} \ldots\ldots\ldots\ldots\ldots(2)$$

Thus the weight of each bridge-edge e_{pq} is:
$0 < w(e_{pq}) \le 1$. Based on a predetermined threshold (*thrs*) value, we remove all bridge-edges with weight less than the *thrs*. In this work we use *thrs=0.50.* This thrs value (0.50) was selected experimentally after extensive tests with multiple GOA databases for various organisms.

3. PATH LENGTH SIMILARITY

A number of similarity measures based on the Gene Ontology (GO) annotation terms have been proposed and applied in the past several years for measuring the functional similarity of genes and proteins [1, 6, 12]. Path length measure (PL) is a direct technique that relies on the ontology structure for computing the similarity of genes [12]. In this measure, PL, we compute path length (PL) between GO terms and between genes/proteins.The path length between two GO terms in the same graph is computed straight forward by edge counting. If there is more than one path, then the shortest path is taken as follows:

PL(t1, t2) = the shortest path length between nodes *t1* and *t2*, ……………..…(3)

where t1 and t2 are two GO terms in a single ontology graph; in this case, either F (*molecular function*) or P (*biological process*). If the two nodes belong to two ontology graphs, then each path between them passes through a *bridge-edge*. Define a path length between two nodes belonging to two ontology graphs as follows: If nodes t_i and t_j are not in the same ontology graph then:

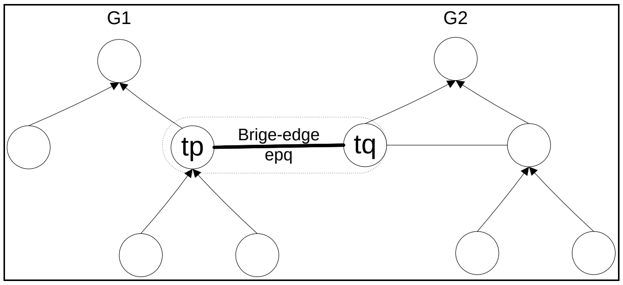


Figure 2: bridge edge and bridge nodes

$$PL(t_i, t_j) = PL(t_i, t_p) + PL(t_j, t_q) + 1/w(e_{pq}) \ldots\ldots\ldots\ldots(4)$$

where t_p is a bridge node in the path from t_i to the root, and similarly t_q is a bridge node in the path from t_j to the root; and *w()* is weight function shown in equation (2). The path length between two proteins is computed as average of PL of all GO terms of the two proteins as follows:

$$PL\ (P_p, P_q) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} PL\ \ (go_p^i, go_q^j)}{n \times m} \quad \ldots..(5)$$

where go^i_p and go^j_q are annotation terms of proteins P_p and P_q respectively. The similarity between two proteins is based on a PL similarity method proposed in previous work [4]. The transfer function for mapping the PL distance into similarity value as follows:

$$\text{Sim(p1, p2)} = e^{-\ f\ *PL(p1,p2)} \ldots\ldots(6)$$

where PL(p_1,p_2) is the path length between the two proteins p_1, p_2 based on their GO annotation terms and *f* is a tuning parameter (*f=0.20* in this research).

4. EVALUATION AND DISCUSSION

We conducted the evaluation on a dataset of 100 disease phenotypes from the OMIM database [13] and UniprotKB [www.uniprot.org/help/uniprotkb]. Each disease phenotype is associated with several proteins. The GO annotation terms of human proteins associated with these diseases are taken from Human UniProtKB GOA database [14]. In the evaluation, we created two sets each containing 50 pairs of proteins selected randomly. Each pair in the first set includes two proteins taken from the same disease (we call it set *S1*) while each pair in the second set contains two proteins taken from two different diseases (we call it set *S2*). We applied the method to measure the similarity between the two proteins in every protein pair using the annotation terms from molecular function (F), biological process (P), or FP combined. The results are shown in Table 2 for set *S1* and in Table 3 for set *S2*.

We measured the similarity between two proteins in every pair using the PL similarity measure explained in Section 3. For each pair of proteins, the similarity is measured using their GOA terms using (1) molecular function (F) terms only (2) biological process (P) terms only, or (3) F and P terms combined, and we call them *SIM_F*, *SIM_P*, and *SIM_FP_combd* respectively; see Tables 2 and 3. The average similarity SIM_F (using only F terms) of proteins in S1 (0.40) is higher than that of set S2 (0.29) as expected. Similarly, the average SIM_P and SIM_FP_combd for S1 (0.31 and 0.22) are higher than for S2 (0.23 and 0.18) as shown in Tables 2 and 3. These results are also illustrated in

Figure 4. This proves that combining F and P ontologies with the proposed approach produces similarity that is streamlining with similarity pattern using only F terms or only P terms. Furthermore, Table_2 shows that the mean value of *SIM_FP_combd* is 0.22 which is lower than the mean SIM_F and mean SIM_P and this is expected. When we use the path length between all F and P terms of both proteins in the combined FP graph, we will get larger path length values and hence lower similarity values. There is a clear difference between the mean value of SIM_FP_combd between S1 (same disease proteins) which is 0.22 and set S2 (different disease proteins) which is 0.18. The average path length between same disease proteins (set S1) is 7.90 while for different-disease proteins (set S2) is 8.63.

Clearly, the idea of combining two GO graphs into one is novel. In this work, we used a limited shared knowledge to combine the GO graphs. We used the GOA data on 100 OMIM diseases. In total, we had 100 diseases, and on average, each disease is associated with about less than 5 proteins for a total of 445 proteins. From the GOA annotation data of these proteins, we removed all the cellular components (C) term. The total F and P gene ontology annotation terms for all the proteins are 8255 with an average of 18.5 terms per protein. From these GOA data our method was able to create 30594 bridge-edges between F and P graphs. 86.7% of the bridge-edges have *b_count() = 1*.The system found only 31 bridges-edges with weight ≥ 0.5. The *max(b_count())* of all bridge-edges is 26 which is the bridge-edge connecting the F term *protein binding* (GO:0005515) with the P term *blood coagulation* (GO:0007596). That is, among the 445 proteins associated with these 100 diseases, there are 26 proteins (almost 6%) associated with both *protein binding* and *blood coagulation*.

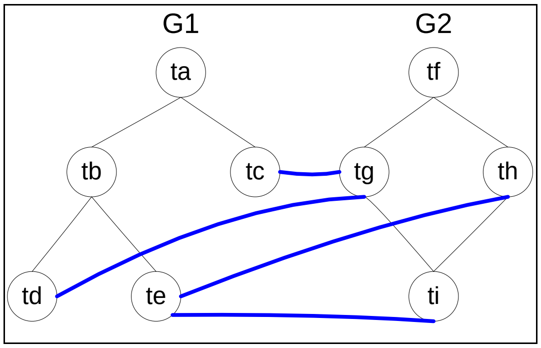


Figure 3: Four bridge-edges between G1 and G2

Table 1: Example of GOA data for four genes.

Gene id	GO Annotation	
	P terms	F terms
AAC1	GO:0006783, GO:0006810, GO:0006839, GO:0009060 GO:0015886, GO:0055085	GO:0005488, GO:0005215 GO:0005471
AAC3	GO:0006783, GO:0006810 GO:0009061, GO:0015886 GO:0055085	GO:0005488, GO:0005215 GO:0005471
ROD1	GO:0042493, GO:0070086	GO:0031625
SNM1	GO:0006379, GO:0006364	GO:0003723, GO:0000171 GO:0016787, GO:0004518

5. CONCLUSION

We presented an approach for combining two GO graphs, the molecular function and biological process ontology graphs. The functional similarity of proteins is measured from the GO terms using the combined graph. All the existing approaches for measuring the similarity between genes and proteins rely on GO annotation terms from either molecular function or biological process ontology. The approach is based on an algorithm for combining graphs using shared knowledge. We used the GOA database as the shared knowledge to combine the molecular function and biological process graphs. Clearly, the idea of merging these two GO graphs is novel and will enable for more comprehensive way of estimating the degree of relationship between genes using their function terms and process terms combined. Also, connecting the two graphs can enable more comprehensive viewing, exploring, and understanding of the knowledge with its both aspects. The external knowledge is represented as a set of tuples to allow for determining bridge nodes and creating bridge-edges between the graphs. The evaluation was conducted on two datasets of proteins pairs related to 100 disease phenotypes extracted from OMIM. We used path length based similarity measure applied to the GO function and process combined graph.

6. REFERENCES

[1] J. J. Goeman and U. Mansmann, Multiple Testing on the Directed Acyclic Graph of Gene Ontology. Bioinformatics 2008, 24(4):537-544.

[2] G. Chartrand; and L. Lesniak, Graphs and Digraps, 4th Edition, 2004, CRC press.

[3] T. H. Cormen, C.E. Leiserson, R. L. Rivest and C.L. Stein, Introduction to Algorithms, 3rd Edition, 2009, MIT Press.

[4] Al-Mubaid H. and Nagar A., A New Path Length Measure Based on GO for Gene Similarity with Evaluation Using SGD Pathways. Proc of 2st IEEE Int'l Symposium on Computer-Based Medical Systems CBMS, 2008.

[5] G. Ganapathy and R. Lourdusamy. Matching and Merging of Ontologies Using Conceptual Graphs. Proc. of the World Congress on Engineering WCE 2011, July, 2011, London, UK.

[6] H. Al-Mubaid and Hoa A. Nguyen. Measuring Semantic Similarity between Concepts within Multiple Ontologies in the Biomedical Domain. IEEE Trans. SMC-C, Vol.39, No.4, pp. 389-398, July 2009.

[7] The Gene Ontology: www.geneontology.org

[8] R. Ambauen, S. Fischer and Horst Bunke. Graph Edit Distance with Node Splitting and Merging, and Its Application to Diatom Identification. In Graph Based Representations in Pattern Recognition, Lecture Notes in Computer Science, 2003, Volume 2726, 2003.

[9] GOA: http://www.ebi.ac.uk/GOA/

[10] Vaida Jakoniene and Patrick Lambrix. Ontology-based integration for bioinformatics. Proc of 31st VLDB Conf, Trondheim, Norway, 2005.

[11] Robinson PN, Mundlos S. The Human Phenotype Ontology. Clinical Genetics, 2010: 77: 525–534.

[12] H. Al-Mubaid and A. Nagar, A New Path Length Measure Based on GO for Gene Similarity with Evaluation Using SGD Pathways, Proc. IEEE CBMS'2008, 2008.

[13] OMIM database: http://www.ncbi.nlm.nih.gov/omim

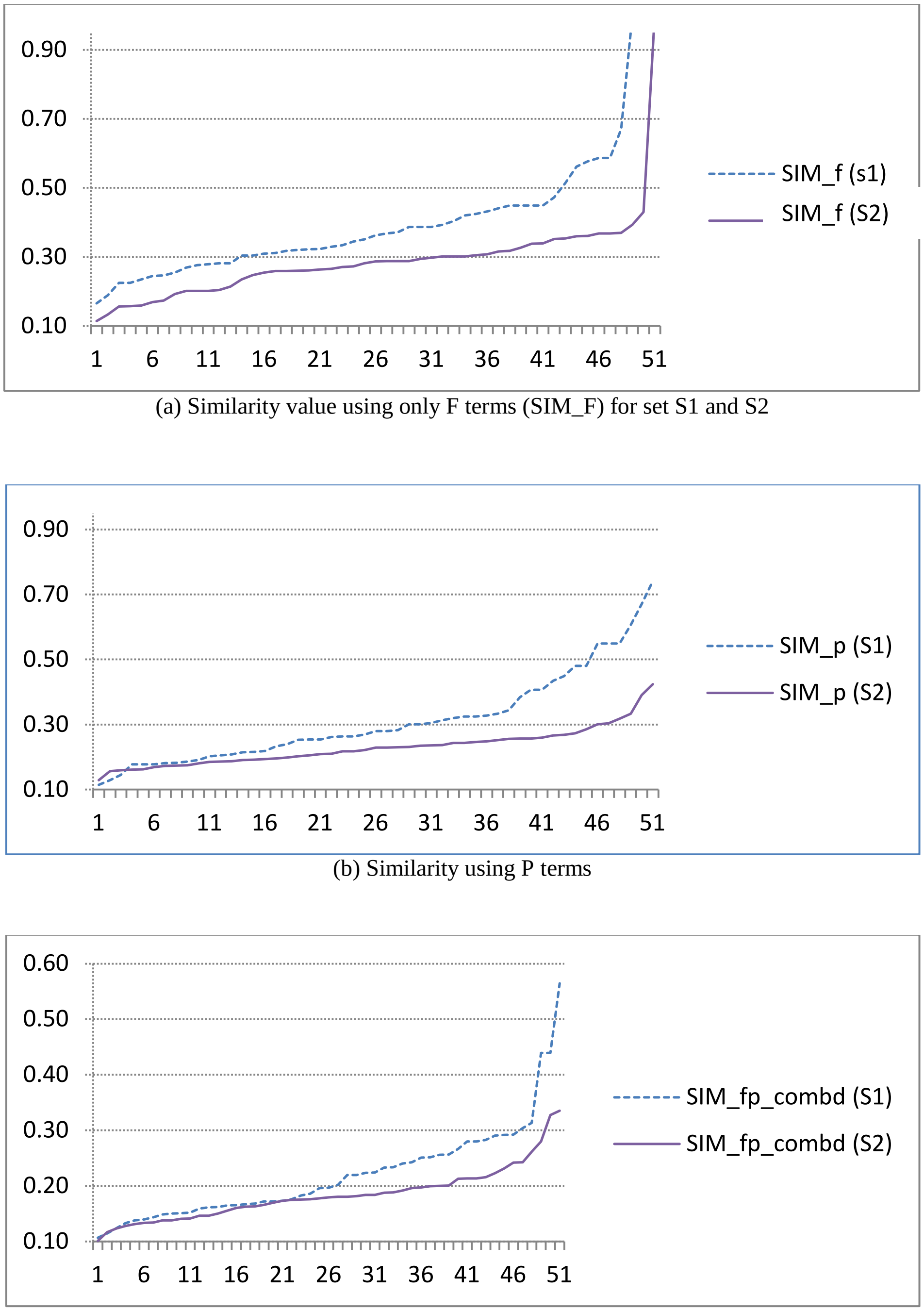
[14] Human UniProtKB GOA database: www.ebi.ac.uk/GOA/human_release.html

Disease	Protein 1	Protein 2	SIM_F	SIM_P	SIM_FP_combd
FANCONI ANEMIA	P51587	Q9BXW9	0.45	0.25	0.24
NEURAL TUBE DEFECTS, FOLATE-SENSITIVE	P42898	P11586	0.39	0.38	0.25
STREPTOMYCIN OTOTOXICITY	Q969Y2	Q8WVM0	0.22	0.67	0.17
PEROXISOME BIOGENESIS DISORDERS	P56589	O60683	0.58	0.55	0.20
ADENOCARCINOMA OF LUNG	Q9Y238	P15056	0.44	0.30	0.26
RENAL CELL CARCINOMA 2	P49789	P11362	0.25	0.11	0.12
LEBER OPTIC ATROPHY	P03915	P00846	0.17	0.28	0.17
FAMILIAL HYPERTROPHIC CARDIOMYOPATHY	Q9UM54	P56539	0.33	0.19	0.15
FANCONI ANEMIA	Q9NPI8	P51587	0.42	0.33	0.29
FAMILIAL HYPERTROPHIC CARDIOMYOPATHY	P10916	P56539	0.37	0.18	0.14
FAMILIAL HYPERTROPHIC CARDIOMYOPATHY	Q9H1R3	P56539	0.32	0.19	0.15
CFC SYNDROME	P01116	P15056	0.23	0.21	0.17
IDIOPATHIC HYDROPS FETALIS	P04062	P08236	0.39	0.30	0.22
PAPILLARY CARCINOMA OF THYROID	P06753	Q16204	0.56	0.45	0.31
MYASTHENIC SYNDROME, CONGENITAL, SLOW-	P02708	P11230	0.32	0.30	0.23
LEIGH SYNDROME	Q12887	P00846	0.19	0.18	0.11
MOLYBDENUM COFACTOR DEFICIENCY	O96033	Q9NQX3	0.43	0.61	0.22
BARDET-BIEDL SYNDROME	Q6ZW61	Q9H0F7	0.37	0.20	0.16
RETINITIS PIGMENTOSA	P29973	P82279	0.31	0.32	0.24
RENAL TUBULAR DYSGENESIS	P12821	P00797	0.30	0.22	0.17
NEONATAL ADRENOLEUKODYSTROPHY	O43933	Q92968	0.45	0.25	0.17
CATARACT, AUTOSOMAL DOMINANT	P02489	Q13515	0.51	0.22	0.22
MITOCHONDRIAL COMPLEX IV DEFICIENCY	Q15526	Q12887	0.39	0.26	0.15
STREPTOMYCIN OTOTOXICITY	O75648	Q969Y2	0.25	0.74	0.16
GLYCINE ENCEPHALOPATHY	P23378	P48728	0.36	0.55	0.23
MYASTHENIC SYNDROME, CONGENITAL,	Q04844	P11230	0.33	0.32	0.18
MATURITY-ONSET DIABETES OF THE YOUNG	Q13562	P19835	0.24	0.18	0.14
NEONATAL ADRENOLEUKODYSTROPHY	Q92968	O43933	0.45	0.25	0.16
CATARACT, AUTOSOMAL DOMINANT	P43320	Q13515	0.59	0.26	0.28
WILLIAMS-BEUREN SYNDROME	Q9Y4P3	Q9UIG0	0.45	0.28	0.29
PROTOCADHERIN-BETA GENE CLUSTER	Q9Y5E3	Q9Y5F3	1.00	0.48	0.44
WILLIAMS-BEUREN SYNDROME	P15502	Q9UIG0	0.32	0.18	0.20
PAPILLARY CARCINOMA OF THYROID	P07949	Q16204	0.28	0.43	0.29
MULTIPLE SULFATASE DEFICIENCY	P15289	Q8NBK3	0.28	0.32	0.27
SQUAMOUS CELL CARCINOMA	P04637	Q9UK53	0.35	0.33	0.20
PROTOCADHERIN-BETA GENE CLUSTER	Q9Y5E5	Q9Y5F3	1.00	0.48	0.44
EPIDERMOLYSIS BULLOSA LETALIS	Q13751	Q13753	0.32	0.41	0.28
NONINSULIN-DEPENDENT DIABETES MELLITUS	O15357	Q9HC96	0.28	0.21	0.17
MELAS SYNDROME	P03923	P03905	0.67	0.34	0.30
PROTOCADHERIN-BETA GENE CLUSTER	Q9Y5F1	Q9Y5F3	1.00	0.55	0.56
ADENOCARCINOMA OF LUNG	P00533	P15056	0.27	0.21	0.16
PHEOCHROMOCYTOMA	P40337	P21912	0.43	0.18	0.15
FAMILIAL ATYPICAL MYCOBACTERIOSIS	P42224	P42701	0.33	0.28	0.24
LACRIMOAURICULODENTODIGITAL SYNDROME	P21802	P22607	0.30	0.24	0.18
RHEUMATOID ARTHRITIS	Q9UBC1	Q9UM07	0.34	0.14	0.13
INFLAMMATORY BOWEL DISEASE 5	Q9HC29	Q9UM07	0.23	0.13	0.11
MITOCHONDRIAL COMPLEX IV DEFICIENCY	P00414	Q12887	0.39	0.27	0.14
PAPILLARY CARCINOMA OF THYROID	Q16204	Q8TBA6	0.31	0.41	0.25
WILLIAMS-BEUREN SYNDROME	Q9BQE9	Q9UIG0	0.28	0.28	0.26
BARDET-BIEDL SYNDROME	Q9H0F7	Q8TAM1	0.47	0.23	0.19
		Average	**0.40**	**0.31**	**0.22**

Table 2: Similarity values of 50 pairs of same-disease proteins (set S1).

Disease	Protein 1	Disease 2	Protein 2	SIM_F	SIM_P	SIM_FP_
RETINITIS PIGMENTOSA	O43186	CONGENITAL NONBULLOUS	O75342	0.27	0.27	0.18
PARKINSON DISEASE	O43464	CONGENITAL ONDINE	P07949	0.26	0.25	0.19
AUTOIMMUNE DISEASE	O43918	RETINITIS PIGMENTOSA	Q12866	0.29	0.30	0.23
AUTOIMMUNE DISEASE	O43918	BARDET-BIEDL SYNDROME	Q9H0F7	0.43	0.22	0.22
PROSTATE CANCER	O96017	USHER SYNDROME, TYPE I	Q96QU1	0.26	0.23	0.18
LEBER OPTIC ATROPHY	P00414	PITUITARY DWARFISM III	Q9UBX0	0.25	0.20	0.13
CFC SYNDROME	P01116	ZELLWEGER SYNDROME	Q7Z412	0.36	0.16	0.21
SHORT STATURE, IDIOPATHIC,	P01241	LEBER OPTIC ATROPHY	P03891	0.20	0.17	0.13
MELAS SYNDROME	P03886	MYASTHENIC SYNDROME,	P02708	0.20	0.26	0.19
SEVERE COMBINED	P04234	BLADDER CANCER	P22607	0.29	0.24	0.17
THROMBOPHILIA VENOUS	P05121	TRICHOTHIODYSTROPHY,	Q6ZYL4	0.35	0.17	0.15
CONGENITAL ONDINE CURSE	P07949	SUSCEPTIBILITY TO HUMAN	P41597	0.21	0.21	0.16
IDIOPATHIC HYDROPS FETALIS	P10746	ANGELMAN SYNDROME	P51608	0.30	0.19	0.20
RHEUMATOID ARTHRITIS	P11021	USHER SYNDROME, TYPE I	Q13402	0.34	0.17	0.17
FAMILIAL HYPERTROPHIC	P12883	IDIOPATHIC HYDROPS	P69905	0.29	0.20	0.14
DILATED CARDIOMYOPATHY 1A	P12883	MATURITY-ONSET DIABETES	Q13562	0.26	0.19	0.18
ISCHEMIC STROKE	P16109	FAMILIAL ATYPICAL	P29460	0.37	0.23	0.20
FAMILIAL HYPERTROPHIC	P19429	ISCHEMIC STROKE	P24723	0.30	0.21	0.20
RENAL CELL CARCINOMA,	P19532	PROTOCADHERIN-BETA	Q9UN67	0.30	0.26	0.24
PHEOCHROMOCYTOMA	P21912	FAMILIAL HYPERTROPHIC	Q9UM54	0.25	0.19	0.15
AUTOSOMAL RECESSIVE CUTIS	P28300	PAPILLARY CARCINOMA OF	O15164	0.27	0.18	0.16
MYASTHENIA GRAVIS	P28329	MEDULLOBLASTOMA	P25054	0.17	0.24	0.18
FAMILIAL ATYPICAL	P29460	PARKINSON DISEASE	Q9BXM7	0.26	0.19	0.16
SURFACTANT METABOLISM	P32927	PROTOCADHERIN-BETA	Q9Y5E9	0.23	0.39	0.26
ENDOMETRIAL CANCER	P43246	PHEOCHROMOCYTOMA	P40337	0.32	0.19	0.14
AMYLOIDOSIS	P61626	MITOCHONDRIAL COMPLEX I	O15239	0.20	0.26	0.17
MONILETHRIX	P78385	PAPILLARY CARCINOMA OF	O15164	0.37	0.16	0.20
MYASTHENIC SYNDROME,	Q04844	HYPOKALEMIC PERIODIC	Q13698	0.30	0.25	0.14
JUVENILE MYELOMONOCYTIC	Q06124	SHORT STATURE,	P10912	0.33	0.20	0.18
SUSCEPTIBILITY TO HEPATITIS	Q08334	AUTOSOMAL RECESSIVE	Q9UBX5	0.36	0.33	0.33
RETINITIS PIGMENTOSA	Q12866	LEBER OPTIC ATROPHY	P00846	0.16	0.25	0.15
LEIGH SYNDROME	Q12887	SUSCEPTIBILITY TO	P01903	0.16	0.16	0.12
LEIGH SYNDROME	Q12887	SEVERE COMBINED	P04234	0.19	0.16	0.12
CATARACT, AUTOSOMAL	Q13515	FAMILIAL HYPERTROPHIC	P10916	0.30	0.22	0.20
PSEUDOHYPOPARATHYROIDIS	Q5JWF2	WAARDENBURG-SHAH	P14138	0.26	0.22	0.19
BARDET-BIEDL SYNDROME	Q6ZW61	MYASTHENIA GRAVIS	Q04844	0.11	0.27	0.13
PAPILLARY CARCINOMA OF	Q8IUD2	MELAS SYNDROME	P03886	0.29	0.23	0.21
FANCONI ANEMIA	Q8IYD8	ENDOMETRIAL CANCER	P43246	0.20	0.32	0.14
WALKER-WARBURG	Q8WZA1	MATURITY-ONSET DIABETES	O14901	0.17	0.30	0.13
WILLIAMS-BEUREN SYNDROME	Q9BQE9	PAPILLARY CARCINOMA OF	Q8IUD2	0.39	0.42	0.34
BARDET-BIEDL SYNDROME	Q9BXC9	FANCONI ANEMIA	Q9NVI1	1.00	0.29	0.28
WILLIAMS-BEUREN SYNDROME	Q9UHL9	FAMILIAL HYPERTROPHIC	Q14896	0.34	0.26	0.22
RHEUMATOID ARTHRITIS	Q9UM07	JUVENILE MYOCLONIC	O00305	0.13	0.13	0.10
RHEUMATOID ARTHRITIS	Q9UM07	MOLYBDENUM COFACTOR	O96007	0.35	0.19	0.17
FAMILIAL HYPERTROPHIC	Q9UM54	FANCONI ANEMIA	O15287	0.31	0.24	0.18
MEDULLOBLASTOMA	Q9UMX1	ZELLWEGER SYNDROME	Q13608	0.32	0.17	0.16
PROTOCADHERIN-BETA GENE	Q9Y5E1	BARDET-BIEDL SYNDROME	Q8NFJ9	0.37	0.27	0.24
PROTOCADHERIN-BETA GENE	Q9Y5E7	FAMILIAL HYPERTROPHIC	P13533	0.16	0.24	0.18
PROTOCADHERIN-BETA GENE	Q9Y5E7	PHEOCHROMOCYTOMA	P21912	0.28	0.21	0.18
PROTOCADHERIN-BETA GENE	Q9Y5F0	FAMILIAL HYPERTROPHIC	O15273	0.27	0.24	0.21
			Average	**0.29**	**0.23**	**0.18**

Table 3: Similarity values of 50 pairs of different-disease proteins (Set S2).



(a) Similarity value using only F terms (SIM_F) for set S1 and S2

(b) Similarity using P terms

(c)Similarity using FP-combined terms

Figure 4: Illustration of the similarity values using F-term, P-terms and FP-combined