Similarity and Prioritization of Disease Proteins using Path Length Measure

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

Hisham Al-Mubaid
University of Houston-Clear Lake
Houston, TX, 77058, USA
hisham@uhcl.edu

Abstract- Semantic similarity measures have been used successfully and extensively in the biomedical research with various applications. As the biomedical ontologies, which form the main ground for most of the similarity measures, are growing and progressing towards more completeness and higher accuracy, the results and outcomes of these semantic similarity measures become more acceptable and more reliable in the field. In this paper, we investigate a path length based measure for prioritization of disease proteins and for computing the similarity between diseases and proteins. Our measure is based on the GO annotation terms of the proteins and uses a simple exponential transfer function to convert the path length to similarity score. The evaluation results prove that this similarity measure is fairly effective in assessing the closeness of proteins and diseases in the disease protein ranking and protein prioritization experiments.

1. Introduction

Biomedical ontologies have received increasing research attention in the recent years in medical informatics and computational biology. In the studies related to biomedical entities, ontologies are becoming key component in research that involves similarity, comparison and analysis of various kinds of biomedical entities [1-4]. Biomedical ontologies provide a structured and unified way to study genes and proteins from different aspects like prediction of gene functions, disease protein prediction, and protein-protein interactions [1, 5, 6]. For example, two biological entities can be compared using their annotations from certain ontology by comparing and analyzing their annotation information from the ontology [6]. Moreover, the biomedical ontologies are progressing over time and advancing towards more coverage, completeness, and accuracy which prompts for more research and utilization of the annotations and information derived from these ontologies [6].

Semantic similarity measures have been used effectively and extensively in the biomedical research with wide range of applications [1 – 7]. Furthermore, computing semantic similarity using similarity measures are now considered more reliable means to estimate and predict various aspects of gene products and other entities, e.g. disease proteins, drug targets, and interactions.

A semantic similarity measure is a function, *e.g. sim(p, q)*, that attempts to estimate the similarity or *closeness* between two given samples or entities (*p* and *q*) as a numeric value based on the available information on the given pair of entities (p and q).

Semantic similarity measures have been studied for long time in different disciplines and applications including natural language processing, information retrieval, and bioinformatics [8]. Pesquita et al. (2009) [6] presents a review of several semantic similarity measures applied to biomedical ontologies. They classify these measures based on various aspects such as edge based versus node based, or pair-wise versus group-wise, and so on [6].

In this paper, we examine a semantic similarity measure based on path length for computing the similarity between diseases proteins and for ranking disease proteins. Our measure is based on the annotation terms of the proteins from the gene ontology and uses a simple exponential transfer function to convert the path length to similarity score. The evaluation results proved that our similarity measure is fairly accurate and effective in assessing the similarity of proteins and diseases in the disease protein ranking and protein prioritization experiments.

Related work:- The volumes of research on disease protein ranking, disease protein similarity and gene prioritization have been growing in the past several years [1 - 4]. Gene prioritization methods rank the candidate genes based on matching the available information on these genes from multiple data sources against biological processes, pathway, or genes known and confirmed to be associated with the disease phenotype. [3].

Schlicker et al (2010) presents a gene prioritization approach using the similarity measure of the GO annotation terms of diseases and candidate genes [4]. In that work, the GO terms of the genes and proteins known to be related with the disease are considered as the functional profile of the disease [4]. They reported the results of ranking proteins from 78 OMIM phenotypes using various settings. Wang et al. (2010) examined the GO annotation length and its effect on the similarity scores between proteins [1]. They examined 14 semantic similarity measures to compute the semantic similarity between protein pairs. Their results indicated that there is a bias in the similarity scores as these similarity scores are significantly correlated with the number of GO annotation terms [1].

In [2], Chen et al. (2009) used methods from social and web networks for disease gene prioritization and candidate gene identification. They examined network based methods as well as functional annotations based techniques which found to outperform the network based methods [2]. The protein interactions along with GO annotations were also applied and utilized to identify genes related to immunedeficiencies [4].

2. A Similarity Measure

Quite a few similarity measures based on the Gene Ontology (GO) annotation terms have been proposed and adopted in the past several years for the disease gene discovery and gene prioritization [1, 4, 5, 6, 8, 9]. However, none of these measures uses the simple path length [7] as a metric of similarity between genes or proteins. We use a simple path length measure with GO annotation terms of proteins to assess the similarity of disease proteins and to rank proteins. In this work, we use a similarity method proposed in our previous work [7, 8] which computes the similarity *sim(p1, p2)* between two proteins *p1* and *p2* as follows:

$$\mathrm{Sim(p1,\ p2)} = e^{-f*PL(p1,p2)} \quad \ldots\ldots(1)$$

where *PL(p1,p2)* is the path length between the two proteins p1, p2 based on their GO annotation terms and *f* is a tuning parameter (*f=0.20* in this research). The path length between two proteins is computed as follows:

$$PL\ (P_p,\ P_q) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} PL\ \ (go_p^i, go_q^j)}{n \times m} \qquad \ldots\ldots(2)$$

where go^i_p and go^j_q are annotation terms of proteins P_p and P_q respectively. And the path length between two GO terms is as follows:

$\mathrm{PL}(go_{\mathrm{x}}, go_{\mathrm{y}}) =$ *the minimum path length in the GO graph between the terms* go_{x} *and* go_{y} …(3)

For a given disease phenotype that is known to be related with proteins p_i we assign the GO annotation terms of these proteins p_i to the disease. Thus, measuring the similarity between a disease and a protein is then the similarity between two sets of annotation terms [4, 7]. For example, if a disease *Di* is known to be related with 3 proteins p_x, p_y, p_z then let us randomly select one of these 3 proteins (say p_z) for a prioritization experiment and keep the other two proteins (p_x and p_y) for the disease *Di*. Thus, *Di* is then assigned the GO annotation terms of p_x and p_y. We randomly select *n-1* proteins related to other diseases and not related to *Di*; we add protein p_z to this set. The ranking experiment is then conducted by measuring the similarity between *Di* and the *n* proteins (n-1 non-*Di* proteins and p_z). The protein that receives the highest similarity with *Di* is ranked as #1 and so on.

3. Evaluation and Experiments

The diseases and proteins data used in our evaluations are extracted from the OMIM database [www.ncbi.nlm.nih.gov/omim] and UniprotKB [www.uniprot.org/help/uniprotkb]. The GO annotation terms of proteins are taken from Human UniProtKB-GOA database (www.ebi.ac.uk/GOA/human_release.html).

Firstly, we examined the method in ranking 10 disease proteins using the similarity measure explained in Section 2. This test was conducted for 50 times (50 experiments) and the results are shown in Table 1. In each one of these 50 experiments (Table 1), we selected ten proteins to rank them with our method for similarity with the disease in the second column. Of these ten randomly selected proteins, only one (protein-2 shown in the fourth column) is taken from the same disease in the second column. The fifth column shows the rank given to protein-2 by our method. We used the *biological process* (BP) sub-ontology of GO in this evaluation (Table 1). Moreover, the detailed results of the first experiment in Table 1 are shown in Table 2. These results in Table 2 show the ranks assigned by our methods to 10 proteins based on their closeness to *Obesity Leanness* disease.

In another evaluation, we examined how our method will rank the protein *Amyloid beta A4 protein* (*UniProtKB accession # P05067*) which is known to be related with the *Alzheimer* disease (disease OMIM #104300) among 50 proteins in which 49 proteins are selected randomly from other diseases. So we used our method to measure the similarity of *Alzheimer* (OMIM # 104300) represented by the two proteins UniProtKB # P78380 and # P49810 with the 50 proteins. The test is repeated three times with the *biological process* (BP), *cellular component* (CC), and *molecular function* (MF) sub-ontologies of GO. The results are shown in Table 3.

Table 4 shows the results of measuring the similarity between proteins taken randomly from OMIM diseases. In this test, 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 (*set-same*) while each pair in the second set contains two proteins taken from two different diseases (*set-diff*). The second column in Table 4 shows the mean similarity values computed by our method to the 50 pairs of *set-same*; and the third column shows the mean similarity value for the 50 pairs of *set-diff*. The detailed results of the 50 same disease protein pairs *set-same* with BP ontology are shown in Table 5.

4. Discussion and Conclusion

In general, the evaluation and experimental results in this paper support the effectiveness of the path length semantic similarity measure for disease protein similarity and prioritization. The first evaluation of 50 protein ranking experiments produced fairly impressive results as shown in Table 1. In each one of the 50 experiments, we record the rank assigned by our method to one protein selected randomly from the same disease as well as to the 9 other proteins selected from other diseases. In 33 cases (out of 50; or 66%) the target protein was ranked #1 (best) by achieving the highest similarity with the disease (Table 1). And in 74% of the cases the protein was ranked as #1 or #2. The mean value of all 50 ranks is 2.48. Of course, the ranks range from 1 (best) to 10 (worst). In the results in Table 3, 50 proteins were ranked based on similarity with the *Alzheimer* disease (OMIM#104300); of these 50 proteins, only one protein (UniprotKN #*P05067*) is known to be related with Alzheimer. This protein was ranked 3 (out of 50) when BP sub-ontology is used and ranked #1 when CC sub-ontology is used (Table 3). When MF is used this protein is ranked #32. This indicates that the MF GO annotation profile of this protein is not as highly correlated with MF annotation profile of the disease as compared to BP or CC annotations. Table 4 and Table 5 illustrate the similarity values computed by our method to 2 data sets of randomly selected protein pairs where each set includes 50 protein pairs. The first set include pairs such as each pair consists of 2 proteins selected from the same disease (*set-same*) whereas in the second set, each pair consists of 2 proteins taken from 2 different diseases (*set-diff*); see Table 4. As shown by the results, the mean similarity values of the same disease proteins (*set-same*) are significantly higher than for *set-diff* with the three sub-ontologies. The highest difference achieved is when BP is used. These again are encouraging results. Overall, this measure, as the results shown and asserted, is fairly accurate in estimating the similarity and prioritization of disease proteins.

References

[1] J. Wang, X. Zhou, J. Zhu, C. Zhou and Z. Guo. Revealing and avoiding bias in semantic similarity scores for protein pairs. *BMC Bioinformatics*, 11:290, 2010.

[2] J. Chen, B.J. Aronow, and A.G. Jegga. Disease candidate gene identification and prioritization using protein interaction networks. BMC Bioinformatics, 10:73, 2009.

[3] D. Nitsch1 , J P Gonçalves , F Ojeda, B de Moor, and Y Moreau. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics* 11:460, 2010.

Experiment	Disease	Protein-1	Protein-2	Rank
1	OBESITY LEANNESS	O00253	P41159	1
2	OBESITY LEANNESS	P32245	P41159	1
3	RETINITIS PIGMENTOSA	P29973	P82279	1
4	RETINITIS PIGMENTOSA	Q03395	P82279	1
5	LEBER OPTIC ATROPHY	P00156	P00395	1
6	LEBER OPTIC ATROPHY	P00414	P00395	2
7	PARKINSON DISEASE	O43464	O60260	1
8	PARKINSON DISEASE	P04062	O43186	4
9	FANCONI ANEMIA	O15287	O15360	1
10	FANCONI ANEMIA	P51587	O15360	1
11	NONINSULIN-DEPENDENT DIABETES MELLITUS	O15357	P06213	1
12	NONINSULIN-DEPENDENT DIABETES MELLITUS	P14672	P06213	3
13	BARDET-BIEDL SYNDROME	Q3SYG4	Q6ZW61	10
14	BARDET-BIEDL SYNDROME	Q8IWZ6	Q6ZW61	9
15	SEVERE COMBINED IMMUNODEFICIENCY	P04234	P08575	1
16	SEVERE COMBINED IMMUNODEFICIENCY	P16871	P08575	1
17	JUVENILE MYELOMONOCYTIC LEUKEMIA	P01111	P01116	1
18	JUVENILE MYELOMONOCYTIC LEUKEMIA	P21359	P01116	1
19	LACRIMOAURICULODENTODIGITAL SYNDROME	O15520	P21802	1
20	LACRIMOAURICULODENTODIGITAL SYNDROME	O15520	P21802	1
21	PROSTATE CANCER	O96017	P29323	5
22	PROSTATE CANCER	P50539	P29323	9
23	PROGRESSIVE EPIDERMOLYSIS BULLOSA	P16144	Q13751	1
24	PROGRESSIVE EPIDERMOLYSIS BULLOSA	Q9UMD9	Q13751	1
25	HYPOKALEMIC PERIODIC PARALYSIS	P35499	Q13698	1
26	HYPOKALEMIC PERIODIC PARALYSIS	Q9Y6H6	Q13698	5
27	PEROXISOME BIOGENESIS DISORDERS	O00623	O00628	2
28	PEROXISOME BIOGENESIS DISORDERS	O43933	O00628	1
29	HOMOCYSTEINEMIA	P35520	P42898	1
30	HOMOCYSTEINEMIA	Q99707	P42898	1
31	PROTOCADHERIN-BETA GENE CLUSTER	Q9NRJ7	Q9UN66	1
32	PROTOCADHERIN-BETA GENE CLUSTER	Q9UN67	Q9UN66	1
33	ESCC	P04637	P37173	3
34	ESCC	Q9NZC7	P37173	3
35	HEPATOCELLULAR CARCINOMA	O15169	P08581	1
36	HEPATOCELLULAR CARCINOMA	Q16667	P08581	2
37	OMENN SYNDROME	P15918	P55895	1
38	OMENN SYNDROME	Q96SD1	P55895	1
39	PAPILLARY CARCINOMA OF THYROID	O15164	P04629	8
40	PAPILLARY CARCINOMA OF THYROID	P06753	P04629	5
41	MITOCHONDRIAL COMPLEX IV DEFICIENCY	O43819	O75880	1
42	MITOCHONDRIAL COMPLEX IV DEFICIENCY	P00395	O75880	2
43	ZELLWEGER SYNDROME	O00623	O60683	1
44	ZELLWEGER SYNDROME	O75381	O60683	1
45	HERMANSKY-PUDLAK SYNDROME	O00203	Q6QNY0	1
46	HERMANSKY-PUDLAK SYNDROME	Q86YV9	Q6QNY0	1
47	WILLIAMS-BEUREN SYNDROME	O43709	O75344	9
48	WILLIAMS-BEUREN SYNDROME	P15502	O75344	10
49	HIRSCHSPRUNG DISEASE	P07949	P14138	1
50	HIRSCHSPRUNG DISEASE	P24530	P14138	1

Table 1: 50 protein ranking experiments; the fifth column shows the rank of protein-2 in the fourth column when ranked among ten proteins for similarity (using our method and BP ontology) with the disease in second column. Protein-1 represents the disease in the second column.

[4] A Schlicker, T Lengauer and M Albrecht. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. Bioinformatics, Vol. 26 ECCB 2010, pp. i561–i567.

[5] C. Ortutay and M. Vihinen. Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies. Nucleic Acids Res., 37, pp.622–628, 2009.

[6] C. Pesquita, D. Faria1, A.O. Falca, P. Lord, and F.M. Couto1. Semantic Similarity in Biomedical Ontologies. PLoS Computational Biology vol.5, no.7, July 2009.

[7] Al-Mubaid H. and Nagar A., "A New Path Length Measure Based on GO for Gene Similarity with Evaluation Using SGD Pathways". Proceedings of The 2st IEEE International Symposium on Computer-Based Medical Systems CBMS, 2008.

[8] H. Al Mubaid and A. Nagar. Comparison of four similarity measures based on GO annotations for gene clustering. In proceedings of IEEE Symposium on Computers and Communications ISCC 2008, pp. 531–536, July 2008.

[9] A. Gefen, R. Cohen and O.S. Birk. Syndrome to Gene (S2G): In-Silico Identification of Candidate Genes for Human Diseases. Human Mutation, Vol. 31, No. 3, 229–236, 2010.

Disease 1	Protein 1	Disease 2	Protein 2	Sim	Rank
OBESITY LEANNESS	**O00253**	OBESITY LEANNESS	**P41159**	**0.48**	1
OBESITY LEANNESS	O00253	FAMILIAL HYPERCHOLANEMIA	Q9UDY2	0.42	2
OBESITY LEANNESS	O00253	BLADDER CANCER	P01112	0.41	3
OBESITY LEANNESS	O00253	BARDET-BIEDL SYNDROME	Q8NFJ9	0.41	4
OBESITY LEANNESS	O00253	MULTIPLE SULFATASE DEFICIENCY	P15289	0.41	5
OBESITY LEANNESS	O00253	PARKINSON DISEASE	Q99497	0.41	6
OBESITY LEANNESS	O00253	ISCHEMIC STROKE	P24723	0.37	7
OBESITY LEANNESS	O00253	ALZHEIMER DISEASE	P78380	0.33	8
OBESITY LEANNESS	O00253	JUVENILE MYOCLONIC EPILEPSY	O00305	0.32	9
OBESITY LEANNESS	O00253	MITOCHONDRIAL COMPLEX IV DEFICIENCY	P00414	0.20	10

Table 2: Ranking experiment of disease Obesity Leanness (protein O00253) for similarity with 10 proteins

Disease	Disease protein	Ontology	Sim	Rank
Alzheimer (OMIM #103400) represented by (P78380 & P49810)	*Amyloid beta A4 protein* (*UniProtKB accession #: P05067*)	BP	0.52	3
		CC	0.71	1
		MF	0.57	32

Table 3: 50 proteins are ranked by our method for similarity with *Alzheimer* disease. Only one protein (shown in the second column) is taken from the Alzheimer disease and 49 proteins are selected randomly for different diseases.

	Mean sim (*set-same*)	Mean sim (*set-diff*)
Number of protein pairs	50	50
BP	0.581	0.351
CC	0.692	0.519
MF	0.606	0.478

Table 4: The mean *sim* values of two sets of proteins measured by our method using the three ontologies BP, CC, and MF

Disease 1	Protein 1	Disease 2	Protein 2	Sim
ABDOMINAL BODY FAT DISTRIBUTION	P01189	ABDOMINAL BODY FAT	P37231	0.432
ADENOCARCINOMA OF LUNG	P00533	ADENOCARCINOMA OF LUNG	P15056	0.548
ALZHEIMER DISEASE	P49810	ALZHEIMER DISEASE	P78380	0.502
ANGELMAN SYNDROME	O60312	ANGELMAN SYNDROME	P51608	0.192
BETHLEM MYOPATHY	P12111	BETHLEM MYOPATHY	P12109	0.86
BLADDER CANCER	P22607	BLADDER CANCER	P06400	0.401
BLADDER CANCER	P06400	BLADDER CANCER	P22607	0.401
BREAST CANCER	Q9BX63	BREAST CANCER	P38398	0.579
BREAST CANCER	O60934	BREAST CANCER	P38398	0.605
ENDOMETRIAL CANCER	P52701	ENDOMETRIAL CANCER	P12830	0.381
ESCC	Q9NZC7	ESCC	Q9Y238	0.35
FAMILIAL ATYPICAL MYCOBACTERIOSIS	P38484	FAMILIAL ATYPICAL	P42701	0.555
FAMILIAL HYPERTROPHIC CARDIOMYOPATHY	P09493	FAMILIAL HYPERTROPHIC	P56539	0.468
FAMILIAL HYPERTROPHIC CARDIOMYOPATHY	P45379	FAMILIAL HYPERTROPHIC	P56539	0.397
GLYCINE ENCEPHALOPATHY	P23434	GLYCINE ENCEPHALOPATHY	P23378	0.842
HYPOGONADOTROPIC HYPOGONADISM	Q969F8	HYPOGONADOTROPIC	P11362	0.419
HYPOKALEMIC PERIODIC PARALYSIS	Q9Y6H6	HYPOKALEMIC PERIODIC	P35499	0.717
IDIOPATHIC HYDROPS FETALIS	P08236	IDIOPATHIC HYDROPS FETALIS	P04062	0.57
INFLAMMATORY BOWEL DISEASE 5	Q9HC29	INFLAMMATORY BOWEL DISEASE	Q9UIG0	0.363
ISCHEMIC STROKE	P05112	ISCHEMIC STROKE	P12821	0.376
LACRIMOAURICULODENTODIGITAL SYNDROME	O15520	LACRIMOAURICULODENTODIGITAL	P21802	0.693
LEBER OPTIC ATROPHY	P03923	LEBER OPTIC ATROPHY	P00846	0.518
LEBER OPTIC ATROPHY	P03891	LEBER OPTIC ATROPHY	P00846	0.621
LEIGH SYNDROME	P03897	LEIGH SYNDROME	P00846	0.464
MATURITY-ONSET DIABETES OF THE YOUNG	Q13562	MATURITY-ONSET DIABETES OF	P19835	0.312
MOLYBDENUM COFACTOR DEFICIENCY	Q9NZB8	MOLYBDENUM COFACTOR	Q9NQX3	1
MYASTHENIC SYNDROME, CONGENITAL, SLOW-	Q07001	MYASTHENIC SYNDROME,	P02708	0.924
MYASTHENIC SYNDROME, CONGENITAL, SLOW-	Q07001	MYASTHENIC SYNDROME,	P11230	0.764
NONINSULIN-DEPENDENT DIABETES MELLITUS	Q9HC96	NONINSULIN-DEPENDENT	P14672	0.389
OMENN SYNDROME	P55895	OMENN SYNDROME	Q96SD1	0.618
PAPILLARY CARCINOMA OF THYROID	Q8TBA6	PAPILLARY CARCINOMA OF	Q16204	0.403
PAPILLARY CARCINOMA OF THYROID	P06753	PAPILLARY CARCINOMA OF	Q16204	0.593
PARKINSON DISEASE	P04062	PARKINSON DISEASE	O43464	0.499
PROTOCADHERIN-BETA GENE CLUSTER	Q9UN67	PROTOCADHERIN-BETA GENE	Q9Y5F3	0.784
PROTOCADHERIN-BETA GENE CLUSTER	Q9Y5F0	PROTOCADHERIN-BETA GENE	Q9Y5F3	0.784
RENAL CELL CARCINOMA, PAPILLARY	Q92733	RENAL CELL CARCINOMA,	Q9BZE9	1
RENAL CELL CARCINOMA, PAPILLARY	Q9BZE9	RENAL CELL CARCINOMA,	Q92733	1
RENAL TUBULAR DYSGENESIS	P30556	RENAL TUBULAR DYSGENESIS	P12821	0.538
RETINITIS PIGMENTOSA	P82279	RETINITIS PIGMENTOSA	P29973	0.641
RETINITIS PIGMENTOSA	P12271	RETINITIS PIGMENTOSA	P29973	0.7
RETINITIS PIGMENTOSA	P08100	RETINITIS PIGMENTOSA	P29973	0.584
SQUAMOUS CELL CARCINOMA	P04637	SQUAMOUS CELL CARCINOMA	Q9UK53	0.409
STREPTOMYCIN OTOTOXICITY	O75648	STREPTOMYCIN OTOTOXICITY	Q969Y2	0.769
TETRALOGY OF FALLOT	P52952	TETRALOGY OF FALLOT	Q8WW38	0.599
TURCOT SYNDROME	P40692	TURCOT SYNDROME	P25054	0.429
USHER SYNDROME, TYPE I	Q96QU1	USHER SYNDROME, TYPE I	Q9H251	0.932
WAARDENBURG-SHAH SYNDROME	P14138	WAARDENBURG-SHAH	P24530	0.661
WILLIAMS-BEUREN SYNDROME	P35250	WILLIAMS-BEUREN SYNDROME	Q9UIG0	0.448
WILLIAMS-BEUREN SYNDROME	Q9GZY6	WILLIAMS-BEUREN SYNDROME	Q9UIG0	0.19
ZELLWEGER SYNDROME	Q7Z412	ZELLWEGER SYNDROME	O60683	0.802
			average	**0.581**

Table 5: The full similarity results of the 50 same-disease protein pairs *set-same* (each pair contain two proteins taken from the same disease) and BP ontology is used.