Article Title:	Gene Multifunctionally Scoring Using Gene Ontology
Author(s):	Hisham Al-Mubaid
DOI:	10.1142/S0219720018400188
Accepted:	18 September 2018
To be cited as:	Hisham Al-Mubaid, Gene Multifunctionally Scoring Using Gene Ontology, *Journal of Bioinformatics and Computational Biology*, doi: 10.1142/S0219720018400188
Link to final version:	https://doi.org/10.1142/S0219720018400188

World Scientific
www.worldscientific.com

Journal of Bioinformatics and Computational Biology
© Imperial College Press

GENE MULTIFUNCTIONALITY SCORING USING GENE ONTOLOGY

HISHAM AL-MUBAID*

Computer Science Department, University of Houston-Clear Lake
Houston, TX 77062, USA†
hisham@uhcl.edu

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Multifunctional genes are important genes because of their essential roles in human cells. Studying and analyzing multifunctional genes can help understand disease mechanisms and drug discovery. We propose a computational method for scoring gene multifunctionality based on functional annotations of the target gene from the Gene Ontology. The method is based on identifying pairs of GO annotations that represent semantically different biological functions and any gene annotated with two annotations from one pair is considered multifunctional. The proposed method can be employed to identify multifunctional genes in the entire human genome using solely the GO annotations. We evaluated the proposed method in scoring multifunctionality of all human genes using four criteria: gene-disease associations; protein-protein interactions; gene studies with PubMed publications; and published known multifunctional gene sets. The evaluation results confirm the validity and reliability of the proposed method for identifying multifunctional human genes. The results across all four evaluation criteria were statistically significant in determining multifunctionality. For example, the method confirmed that multifunctional genes tend to be associated with diseases more than other genes, with significance $p<0.01$. Moreover, consistent with all previous studies, proteins encoded by multifunctional genes, based on our method, are involved in protein-protein interactions significantly more ($p<0.01$) than other proteins.

***Keywords*:** multifunctional genes, functional genomics.

1. Introduction

Studying and understanding the functions of all genes in a genome is a central step in functional genomics [1, 2, 4, 3]. In particular, multifunctional genes are important to study as they convey essential roles in an organism and in the human genome in particular [1, 2, 3, 4]. A gene is multifunctional if it is involved in more than one distinct function in human body. Studying and uncovering multifunctional genes are important tasks for various fields like gene-disease associations, drug discovery, and functional genomics studies.

In this paper, we study human genes in the entire human genome to examine gene multifunctionality and identify the most likely multifunctional genes. Determining if a gene is multifunctional is not a trivial task as many genes conduct more than one functionality. A gene involved in two functions may not be a multifunctional if the two functions are not distinct (i.e., not semantically different) enough [1]. In this work, we use a computational methodology to determine whether or not a gene is multifunctional and involved in two or more (semantically different) functions. Specifically, we present a method based on the functional annotations of the gene from the Gene Ontology (GO) for examining gene multifunctionality. We use the GO annotations from the biological process (*bp*) and molecular function (*mf*) aspects of the GO. The proposed method extracts and examines all functions and processes that a gene is annotated with.

The method is based on identifying pairs of GO annotations such that each pair represents two semantically different biological functions; hence, any gene annotated with two annotations of one pair is considered multifunctional. We examined the proposed methods in estimating the multifunctionality of all genes in the human genome. We evaluated the results with four different criteria as compared with previous related work in this problem. The four evaluation criteria are: –gene-disease association; –protein-protein interactions *PPI*; –gene studies and *PubMed* publications; and –using published sets of confirmed multifunctional genes. The evaluation results are encouraging and prove that both scoring methods are valid and reliable indicators of gene multifunctionality across the four evaluation criteria. For example, the proposed method confirms that multifunctional genes tend to be associated with diseases more than other genes in the same annotation population, with significance *p<0.01*, as also proved by previous studies. Moreover, consistent with all previous studies, proteins encoded by multifunctional genes, based on our method, are involved in PPI interactions significantly more (p<0.01, *hypergeometric test*) than other proteins.

2. Background and Related Work

One of the important characteristics of multifunctional genes that motivate more investigations is the gene-disease association. This kind of association is significantly higher in multifunctional genes compared to all genes as confirmed by all previous studies in this domain [1, 2, 8, 5-7]. Therefor the relationships between diseases and multifunctional genes are significant and proved [1, 7].

A multifunctional gene is a gene that is involved with several functions and activities, including molecular and cellular tasks, inside the cell [1, 2, 8, 5]. Typically, studying multifunctional genes can disclose more knowledge about diseases associated with the multifunctional genes. In this paper, we rely on the gene ontology (GO) which is the most popular repository of functional information about human genes [9, 10]. Pritykin, Ghersi, and Singh (2015) presented a comprehensive study of genome-wide multifunctional genes in human [1]. They found that multifunctional genes are significantly more likely to be involved in human disorders [1]. Also, they found that 32% of all multifunctional genes produced by their method are involved in at least one OMIM disorder (http://omim.org), whereas the fraction of other annotated genes involved in at least one OMIM disorder is 21% [1, 11].

Ballouz, Pavlidis, and Gil (2017) studied various gene sets for functional genomics and enrichment [3]. They found that heavily functional genes are highly likely to appear in many genomic study results [12]. They leave it as an ‘open question’ to biologist to assess if their finding of gene multifunctionality is a true biological property. Khan and Kihara (2016) extracts a domain of features including GO, protein-protein interaction, and more, to classify protein into moonlighting (i.e. multifunctional) versus non-moonlighting proteins [13]. Kim et al.

(2017) in their system, DigSee, found that genes that interact with more genes in a PPI network are involved in more disease categories than those with fewer neighbors in the protein interaction network [14].

Salathe et al. (Salathe et al., 2006), investigated the multifunctionality of yeast genes and proteins for a different goal [15]. They found a positive correlation between how many biological process (bp) GO terms a gene is annotated with and its evolutionary conservation in yeast; that is, they found highly significant negative correlation between number of bp GO terms and rate of change of yeast genes [15]. Also, Pritykin et al. observed that multifunctional genes tend to be more evolutionarily conserved [1]. A method for identifying novel moonlighting proteins from current functional annotations in public databases was proposed by [8]. They identified potential moonlighting proteins in the Escherichia coli K-12 genome by examining clusters of GO term annotations taken from *UniProt* and constructed three datasets of experimentally confirmed moonlighting proteins (Khan et al., 2014) [8]. In another study of multifunctional genes, Clark et al. (2011) [16] discovered a statistically significant positive correlation between the number of GO biological process leaf terms a gene has and its number of *Pfam* domains and are usually longer [16].

3. Materials and Methods

The proposed gene multifunctionality method is based on the annotation terms from the GO. The GO is highly regarded as the main source for gene functional information [4, 29]. It is the largest and most comprehensive source of information on gene functions with contents growing and becoming more accurate every day.

We can utilize the GO along with the functional genomics data sets for human to induce the relationships among the functions encoded in the ontology. For example, the path length between two GO terms has been extensively used as a metric in computing semantic similarity among genes [4, 17]. Moreover, many gene similarity measures use the depth of the lowest common subsumer (LCS) in computing gene similarity [18, 17]. In our previous work, we investigated and explained the relationship between GO annotation terms of a gene and gene-disease associations [4]. This paper proposes a new method derived from the gene ontology for identifying multifunctional genes in the entire human genome. The proposed method is based on identifying each pair of GO terms that represents a multifunctionality (semantically different biological functions).

Typically, the similarity between two genes is computed as a function of the similarity of their annotations mainly using the *bp* and the *mf* aspects. That is, the similarity $Sim_g(g_1, g_2)$ between two genes g_1 and g_2 can be a similarity function between the annotations of g_1 and g_2:

$$Sim_g(g_1, g_2) = Sim_t(t_{1i}, t_{2i}) \ldots\ldots\ldots\ldots\ldots (1)$$

where $Sim_g(g_1, g_2)$ is the gene similarity between $\mathrm{g}_1$and g_2; and $\mathrm{Sim}_\mathrm{t}(\mathrm{t}_{1\mathrm{i}}, \mathrm{t}_{2\mathrm{i}})$ is a similarity function between GO terms $\mathrm{t}_{1\mathrm{i}}, \mathrm{t}_{2\mathrm{i}}$ annotating $\mathrm{g}_1$and g_2 respectively.

The gene ontology consists of 3 aspects: Molecular Function *mf*, Biological Process *bp* and Cellular Component *cc*. Each one of these aspects $\{mf, bp, cc\}$ is a complete ontology in itself (www.geneontology.org; [10, 4, 17]). For gene multifunctionality, it is normal to rely only on the *bp* and *mf* aspects.

Let $\mathrm{MaxPL}_\mathrm{p}(\mathrm{g}_\mathrm{x})$ be the maximum path length between all pairwise *bp* annotation terms of gene g_x; that is:

$$MaxPL_p(g_x) = \max_{t_x, t_y \in GOT_p(g_x)} PL(t_x, t_y) \;\; \ldots\ldots\ldots (2)$$

where $\mathrm{PL}(\mathrm{t_x}, \mathrm{t_y})$ is the *shortest* path length between the two annotations t_x and t_y, and $GOT_p(g_x)$ is the set of all *bp* annotations of gene $\mathrm{g_x}$ (and none them subsumes any term; i.e., none of the annotations in $GOT_p(\quad)$ is a descendant of any other term in the same set). For example, in Figure 1, there are two different paths shown between GO:0000001 and GO:0006996 one of them is of length 2 (through GO:0048308) and the second path is of length 3 (through the two GO terms GO:0048311 and GO:0007005); we take 2 as the shortest path length between them. The multifunctionality of a gene increases with the increase in the distinctiveness (i.e., diversity) of the functions that the gene in involved in [1]. The path length between two bp annotations of a target gene can be utilized as an indicator of the distinctiveness of the functions of that gene. Based on this, we employ $MaxPL_p$ in a multifunctionality method based on the maximum shortest path length between the *bp* annotation terms.

In the biological process (*bp*) aspect of GO, each annotation term is basically a node in the ontology graph (which is a directed acyclic graph *DAG*) and is a biological functionality upheld by certain genes [28]. When two *bp* annotations (i.e., graph nodes) are far apart with relatively large path length between them (exceeds a threshold) then we can consider that these two terms represent two distinct (semantically different) biological functionalities. That is, our hypothesis is that, two highly far apart bp annotation terms can be considered as two distinct functions given that neither of these two terms is subsuming the other. Therefore, a gene annotated with two such terms can be considered multifunctional. The computations of multifunctionality scores with bp annotations for human genes go through the algorithm shown in Figure A1.

```
Algorithm 1: Compute multifunctionality scores for all human genes using bp
             annotations
Input: - GOA_human: set of all human gene annotations.
       - GO.obo: set of all gene ontology annotation terms with their parents
Output: - Set {$\mathrm{MaxPL_p(g_x)}$ }: multifunctionality score for every human gene $\mathrm{g_x}$ based
        on bp annotations.
Algorithm:
 (1) Create the set G
     1a) G = $\emptyset$ : let G be the set of all genes annotated in GOA_human
     1b) For each annotated gene $\mathrm{g_i}$ from the set GOA_human:
         i) G = G $\cup$ $\mathrm{g_i}$  : add $\mathrm{g_i}$ to G
 (2) Create the set BP
     2a) BP = $\emptyset$ : let BP be the set of all bp annotation terms in GO.obo
     2b) For each bp annotation term $\mathrm{t_i}$ in GO.obo:
         i) BP = BP $\cup$ $\mathrm{t_i}$  :add $\mathrm{t_i}$ to BP along with its parents
 (3) Create the set GOA_human_bp
     3a) Extract all bp annotations from GOA_human and add them to
         GOA_human_bp
 (4) For each gene $\mathrm{g_x}$ in the set G
     4a) Extract the set $\mathrm{GOT_p(g_x)}$ of all annotations of $\mathrm{g_x}$ from GOA_human_bp
     4b) Set $\mathrm{MaxPL_p(g_x)} = 0$
     4b) If $|\mathrm{GOT_p(g_x)}| < 2$ go to step (4)   :go to step (4) up from beginning
     4c) For each pair $\mathrm{t_i, t_j}$ of annotation terms in $\mathrm{GOT_p(g_x)}$:
         i) Compute the shortest path length $\mathrm{PL(t_i, t_j)}$ between pair $\mathrm{t_i, t_j}$ using the set
            BP
         ii) If $\mathrm{PL(t_i, t_j)} > \mathrm{MaxPL_p(g_x)}$ then set
```

$$\mathrm{MaxPL_p}(\mathrm{g_x}) = \mathrm{PL}(\mathrm{t_i}, \mathrm{t_j})$$

Fig. A1. Algorithm for *MaxPLp()* for all human genes

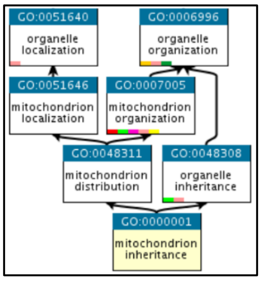

Fig. 1. A small part of the GO.

This algorithm (Algorithm A1) explains the steps of the method.
Now among shortest path length of all-pairwise annotations of a target gene g, we take the maximum one as an indicator of multifunctionality. Considering the bp aspect of GO, the maximum value of path length in the bp graph is 21 which is indicated for two genes (gene Ids: 672 and 5071) as follows:

Gene Id	Gene MIM	UniProtKB	Gene Symbol	No. of Phenotypes	No. of *bp* terms	Max *bp* path length
672	113705	P38398	BRCA1	2	59	21
5071	602544	O60260	PRKN	4	100	21

To further investigate and utilize functional gene GOA datasets along with semantic distance in GO we used mf annotations to compute multifunctionality based on mf annotation terms as:

$$MaxPL_f(g_x) = \max_{t_x, t_y \in GOT_f(g_x)} PL(t_x, t_y) \quad \text{........ (3)}$$

where $\mathrm{GOT_f(g_x)}$ is the set of all mf annotations of gene $\mathrm{g_x}$. We conducted the evaluations based on both bp and mf separately; *details in the next section.*

In an empirical computational work, we tried to vary the contribution of bp and mf by employing a contribution factor α and adopting as a multifunctionality { α .$\mathrm{MaxPL}_\mathbf{p}$ (.) + (1 – α).$\mathrm{MaxPL}_\mathbf{f}$ (.)}. Then, we experimented with a number of values for α between 0.3 to 0.7 with 0.1 increase. We found there is no significant influence in the results from the middle value of α=0.5 which allowed us to eliminate it. In the evaluation and results section, we report the results of using both mf and bp separately. Moreover, we estimate that a gene is considered multifunctional if the *MaxPLp* is exceeding certain threshold (specificity level). This threshold is typically >10. We found that when the threshold is >10 (*i.e., MaxPLp>10*) the fine/coarse grained of the ontology branches do not matter and make no difference, just like the depth of the nodes. The depth of the (*LCS of*) two nodes makes difference only when the path length is relatively small (e.g., <5); and similarly when considering path length > 10, it becomes irrelevant whether coarse-grained (towards the top of the ontology tree) or fine-grained (towards leaves) branches are used.

Contribution: There is a real need for computational methods to provide insights in understating gene functions in the human genome. And identifying multifunctional genes is a central step because of their essentiality as major players in most functionalities in human cells. The Gene Ontology along with the functional genomics databases have not been extensively investigated in computational methods within the domain of multifunctional genes. This work is based on the Gene Ontology which is the most comprehensive, and perhaps the main, source for gene functional information. Gene Ontology is a structure and vocabulary of semantic understanding of the various functions that genes can perform. This work utilizes this semantic structure of gene functions that has been built carefully over the years to induce insight in understanding and identifying genes functions.

4. Evaluation and Results

For all human genes, we extracted all annotation terms from the Gene Ontology Annotation (GOA) database for human [4]. By considering only bp annotations, we found a total of ~35,700 genes annotated with at least one bp terms. Overall, there are ~5.2 bp annotations per gene. By considering mf, there are on average 4.3 mf annotations per gene with a total of ~35,800 genes annotated with at least one mf term. Among all genes with mf annotations (~35,800 genes) in GOA database, almost 42% of them (or 15,142 genes) are annotated with only one mf terms. Each gene with only one mf annotation will have $\mathrm{MaxPL_f} = 0$. Therefore, in mf we have 42% of the genes do not count in the computations of the multifunctionality scoring. For all genes with two or more bp terms, we extracted all bp annotations for each gene from the *GOA* database. In the human annotation dataset GOA_human, ~80% of the genes (= ~29,000 genes) have 4 or fewer mf annotations. We computed the maximum all-pair shortest path length among all terms for every gene as per our method. For evaluation, we would like to verify the reliability of our multifunctionality scoring techniques, $\mathrm{MaxPL_p}$ and $\mathrm{MaxPL_f}$, in estimating the whether or not a gene is multifunctional. We could not find any gold standard dataset to evaluate our methods. So, we used four criteria for multifunctionality [1, 13]. These four criteria are: (1) Gene-disease association is more in multifunctional genes compared with other non-multifunctional genes; (2) Multifunctional genes are more evolutionary conserved; (3) Multifunctional genes tend to be highly studied with relatively higher number of publications; and (4) Using previously tested and published multifunctional gene sets as criteria to test our method.

We analyzed all human genes having bp or mf annotations using the proposed system. After computing $\mathrm{MaxPL_f}(g_x)$ value for each gene, we grouped all genes into clusters of 1000 genes in each cluster after being sorted based on $\mathrm{MaxPL_f}(g_x)$; shown in Table 0 (in the appendix). For example, the top 1000 genes have an average $\mathrm{MaxPL_f}(g_x)$ of 13.99 whereas the next cluster (next 1000 genes) have $\mathrm{MaxPL_f}$ average of 12.189; (*see Table 0 in the appendix*).

Criteria 1. Gene-disease association:

Multifunctional genes are more highly likely to be associated with human diseases than non-multifunctional genes [1, 2, 8, 4, 14, 19]. We analyzed all human genes from the GOA database and from *OMIM* morbid map for disease information (http://omim.org/) [11]. Also, for identifying number of phenotypes per gene, we used the disease ontology (DO) to check whether the two phenotypes are distinct [20]. We wanted to investigate if the number of phenotypes, according to *morbid map*, exhibits any meaningful relationship with our multifunctionality method. We firstly examined the components of the multifunctionality scoring method in equation (4), namely $\mathrm{MaxPL_p}$ and $\mathrm{MaxPL_f}$ independently.

The results in Table 1 show the correlation between MaxPLp and average number of phenotypes for all human genes; these results are also illustrated in Figure 2. Next, we examined $\mathrm{MaxPL_p}$ for each group of genes associated with the same number of phenotypes and the results are in Table 2 and Figure 3. For example, there are 2,572 genes associated with only one phenotype and their average *MaxPLp* is 11.22 whereas the group of genes associated with exactly two phenotypes (648 genes) have an average *MaxPLp* 12.33; Table 2. We mention here that groups of genes associated with ≥7 phenotypes are very small and do not affect the results. For example, there are only 11 genes associated with 7 diseases, and only 7 genes associated with 8 diseases.

We repeated the same evaluation for MaxPL$_f$ (i.e., using *mf* annotation terms) and the results are in Table 3 and also illustrated in Figure 4. As it is shown in both Table 3 and Figure 4, the MaxPL$_f$ increases as the average number of associated phenotypes increases; thus, there is a clear strong correlation between MaxPL$_f$ and average number of phenotypes. Hence, our MaxPL$_f$ is a reliable indicator of multifunctionality of genes. Next, we examined the behavior of MaxPL$_f$ with the increase of phenotypes (i.e., number of phenotypes is independent variable x-axis) for all human genes and the results are shown in Table 4.

Table 1. For each value of MaxPLp this table shows how many genes and the average number of phenotypes

MaxPLp	No. of genes	Avg. of No. of phenotypes
0	1399	0.15
1	71	0.32
2	141	0.21
3	158	0.19
4	204	0.26
5	266	0.21
6	356	0.35
7	497	0.31
8	578	0.31
9	733	0.31
10	919	0.35
11	1319	0.32
12	1490	0.36
13	1541	0.45
14	1512	0.48
15	1166	0.58
16	725	0.64
17	455	0.85
18	217	0.76
19	113	0.73
20	13	1.00
21	2	3.00

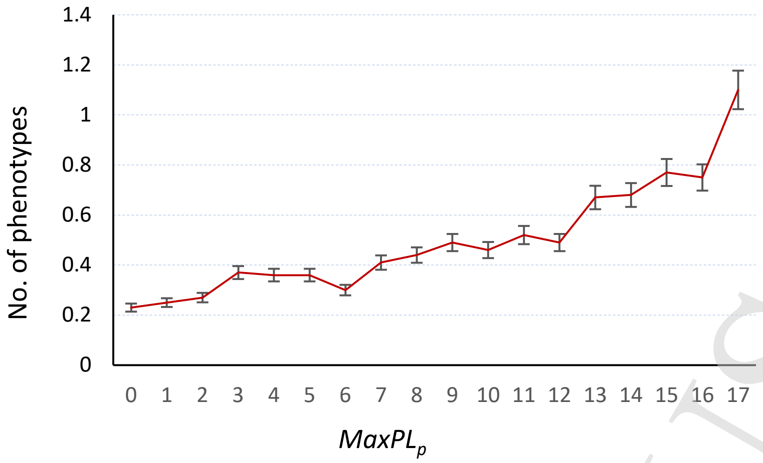

Fig. 2. The relationship between MaxPL$_p$ and number of diseases for all human genes.

Criteria 2. Protein-protein interactions:

Multifunctional genes are typically involved more than normal in protein-protein interactions PPI's [1, 21, 22, 14, 3]. We used this criterion in evaluating our method. We retrieved and compiled PPI data from the Hippie database (http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/) [22]. The obtained data include PPI data involving ~14,800 genes with a total of ~250K experimentally documented P-P interactions [21, 22, 23]. We analyzed the average number of PPI's that a gene involved in with respect to MaxPL$_p$ (and *MaxPL$_f$*) and there is a clear relationship as illustrated in Figure 5. These results prove again that the proposed methods are in line and consistent with this criteria for gene multifunctionality.

Table 2. The average value of MaxPLp for six groups of genes where each group have the same number of phenotypes.

No. of phenotypes	No. of genes	Avg. of MaxPLp
0	10234	10.09
1	2572	11.22
2	648	12.33
3	226	12.96
4	84	13.67
5	51	13.86
6	29	15.45

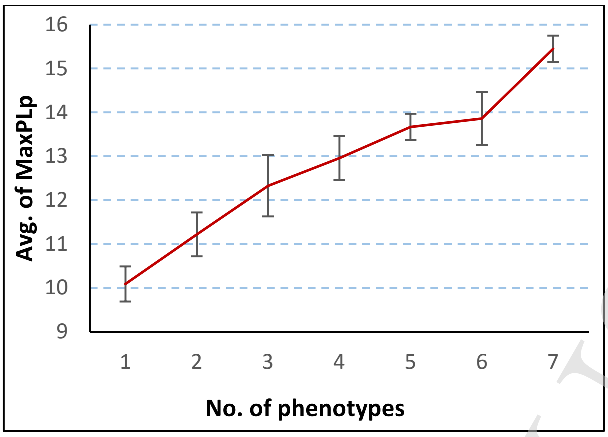

Fig. 3. The relationship of the MaxPLp for each value of average number of phenotype

Table 3. Avg No. of phenotypes for genes with each value of *MaxPL$_f$*

MaxPL$_f$	No. of genes	Avg. of No. of phenotypes
0	2630	0.23
1	244	0.26
2	490	0.27
3	467	0.37
4	557	0.36
5	578	0.36
6	890	0.30
7	831	0.41
8	1018	0.44
9	911	0.49
10	1148	0.46
11	1182	0.52
12	1057	0.49
13	512	0.68
14	475	0.68
15	126	0.77
16	53	0.66
17	10	1.10
18	11	0.73
19	1	0.00

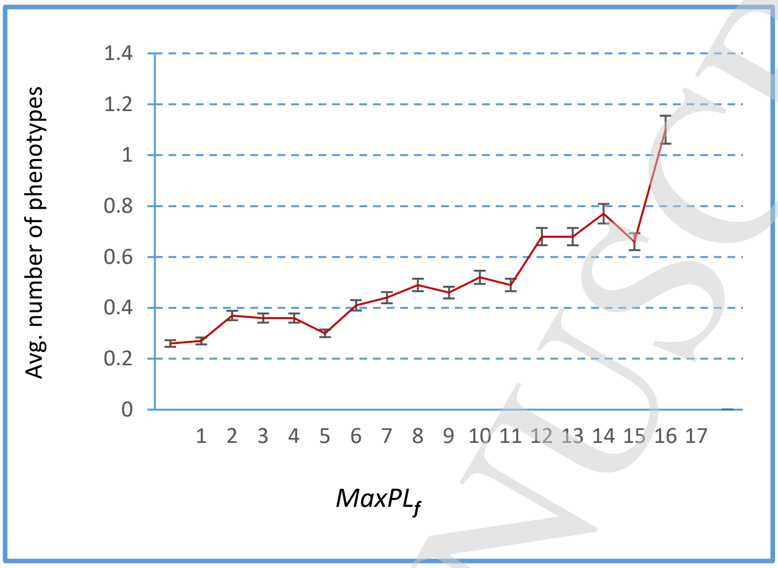

Fig. 4. Average number of phenotypes increases as a function of MaxPL$_f$

Table 4. For each number of phenotypes this table shows number of genes, average number of mf annotations, and average MaxPL$_f$ (e.g., the first row shows that there are 9720 genes having no phenotypes association (0) and having an average MaxPL$_f$ of 6.36)

No. of phenotypes	No. of genes	Avg. MaxPL$_f$
0	9720	6.36
1	2442	7.40
2	619	8.39
3	221	9.13
4	81	8.40
5	48	9.31
6	29	10.17
7	11	7.45
8	7	11.14
9	3	7.67
10	2	12.5
11	4	11.75
12	1	3
13	1	11
14	1	11
16	1	7

Criteria 3. Using *PubMed* publications as indicator of highly studied multifunctional genes: It has been shown that multifunctional genes are highly studied genes and have relatively more publications in the biomedical literature [1, 12]. So, we use publication counts of genes as a criteria of multifunctionality. That is, multifunctional genes tend to have relatively higher

number of publications compared to other annotated genes. We relied on *PubMed* since it is the most comprehensive repository of biomedical literature with more than 24 million citations and references to articles (with abstracts, and some

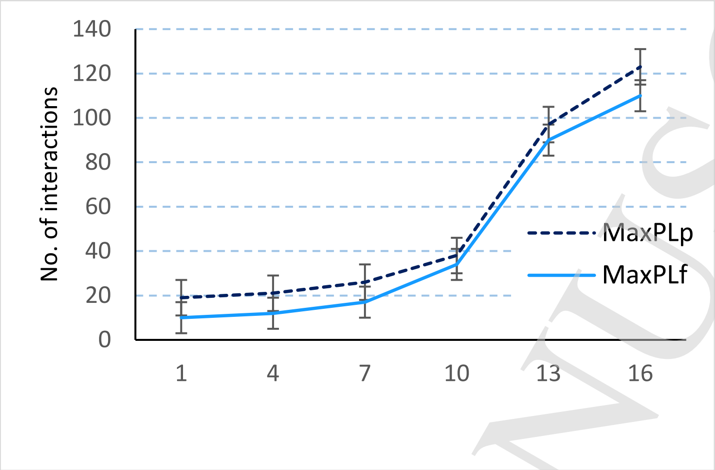

Fig. 5. Number of protein-protein interactions increases as the gene multifunctionality score increase. X-axis here represents MaxPL$_p$ (or MaxPL$_f$) value, and Y-axis represents average number of interactions.

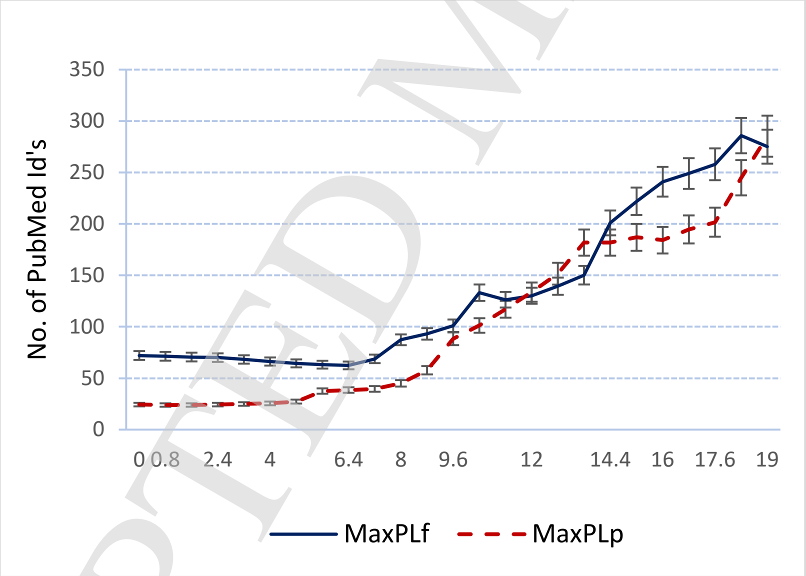

Fig. 6. Analyzing number of PubMed publications against multifunctionality scores with both MaxPL$_f$ and MaxPL$_p$ for all human genes show a direct proportional relationship. The X-axis here represents MaxPL$_p$ (or MaxPL$_f$) value.

with full texts). We analyzed number of publications related to each gene in *PubMed* as it is published by NCBI/PubMed and freely available with file name: gene2pubmed.gz, (link: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA; downloaded Sept.2017) [26, 27].

We examined genes with our multifunctionality scores versus number of publications. The analysis results show a clear straightforward proportionality between number of publications and both scoring methods ($MaxPL_f$ and MaxPLp) for all human genes as illustrated in Figure_6.

Criteria 4. Using published multifunctional genes:

We retrieved two lists of experimentally tested and known multifunctional genes [1].

Source 1: http://moonlightingproteins.org/proteins/ which includes 361 moonlighting proteins (74 for human).

Source 2: http://wallace.uab.es/multitask/ which includes 288 proteins (88 of them for human).

This test was not reliable as we are considering only 162 (74 from set1 and 88 from set2) human genes (out of ~35000 annotated genes); however, these genes exhibit higher $MaxPL_f$ and $MaxPL_p$ values than expected by chance with significance ($p<0.01$) confirming multifunctionality.

5. Discussion

The results in Tables 1 and 2 confirm that there is a direct proportional relationship between gene multifunctionality scores and the number of diseases associated with gene. For example, genes with $MaxPL_f$ =2 (490 genes) have on average 0.27 associated diseases whereas genes with $MaxPL_f$ =3 (467 genes) are on average associated with 0.37 diseases; and this is significant $p<0.01$ (*using hypergeometric test*) as shown in the following excerpt of Table 3:

$MaxPL_f$	No. of genes	Avg. No. of phenotypes
2	490	0.27
3	467	0.37

Also, the average number of phenotypes for 1182 genes with $MaxPL_f$ of 11 is 0.52, and when the $MaxPL_f$ increases to 13 the average number of phenotypes increases to 0.68 which is significant result ($p<0.01$) as shown in the following excerpt of Table 2:

$MaxPL_f$	No. of genes	Avg. of No. of phenotypes
11	1182	0.52
13	512	0.68

From Table 3, we can see for genes with one phenotype (2442 genes) the average $MaxPL_f$ is 7.40 whereas for those genes with 2 phenotypes the average $MaxPL_f$ increases to 8.39 and this is significant with $p<0.01$ as shown in the following excerpt of Table 3:

No. of phenotypes	No. of genes	Avg. $MaxPL_f$
1	2442	7.40
2	619	8.39

Similarly, for MaxPLp, we see that the 10234 genes associated with 0 phenotypes have MaxPLp average of 10.09 and for the genes associated with one phenotype (2572 genes) the value of MaxPLp increases to 11.22 which is significant $p<0.01$ as shown in the following excerpt of Table 5:

No. of phenotypes	No. of genes	Avg. MaxPLp
0	10234	10.09
1	2572	11.22

Other results: we analyzed some fairly well known disease genes to examine the behavior of our proposed scoring method with these particular genes. Clearly all of them are multifunctional (equations (4) – (6)) as shown below:

Gene	Gene Id	MaxPL_f	MaxPL_p	mfs
SNCA - Parkinson disease	6622	13	15	0.70
BRCA2 - breast cancer gene	675	15	14	0.73
TP53 - tumor protein	7157	12	19	0.78
BRCA1 - tumor protein	672	11	21	0.80
APP - Alzheimer disease AD	351	11	17	0.70

Table 5. The multifunctionality scores based on mf annotations, MaxPL_f, show clear difference between genes associated with phenotypes (MaxPL_f =9.02) versus gene not associated with any phenotype (MaxPL_f = 6.36)

With mf annotations only		
Genes with 0 phenotype	Avg. MaxPL_f 6.36	No. of genes: 9720
Genes with ≥1 phenotypes	Avg. MaxPL_f 9.02	No. of genes: 3471
Genes with 0 or 1 phenotypes	Avg. MaxPL_f 6.88	No. of genes: 12162
Genes ≥ 2 phenotypes	Avg. MaxPL_f 9.14	No. of genes: 1029
Overall MaxPL_f Avg: 6.733		

Table 6. MaxPLp shows clear differentiation between genes associated with phenotypes versus those genes not associated with any phenotypes (similar to Table 6 above).

With bp annotations only		
Genes with 0 phenotype	Avg. MaxPLp 10.09	No. of genes: 10234
Genes with ≥1 phenotypes	Avg. MaxPLp 14.23	No. of genes: 3641
Genes with 0 or 1 phenotypes	Avg. MaxPLp 10.65	No. of genes: 12806
Genes ≥ 2 phenotypes	Avg. MaxPLp 14.45	No. of genes: 1069
Overall MaxPLp Avg: 10.50		

Further, the results in the last two tables (Table 5 and Table 6) prove the significant direct proportionality relationship between our multifunctionality methods and disease association. Regarding number of publications, criteria 3, we confirmed that as our multifunctionality score of a gene tend to increase the number of PubMed publications related to the gene also increases as illustrated in Figure 6. We should mention here that higher number of publications implies that the gene is highly studied [1]. One of the main reason of being highly studied is the gene is highly likely associated with one or more diseases. We should mention here that there are genes with fairly high number of publications but with low (≤ 7) multifunctionality score for which reason we relied on the aggregate averages. For example, considering genes with MaxPLp of 12; their average number of PubMed publications is ~133; when we increase the score to 14 the average increases to ~181 and this is significant ($p<0.01$). Finally, we assume multifunctional are genes with MaxPLp≥15, we got 2691 multifunctional genes (genes having MaxPLp of 15 or more). Among these 2691 genes, we found 46% (or 1231 genes) of them are also *mf* multifunctional with mf annotations only using threshold T_f =10 (i.e., $MaxPL_f \geq 10$), and this is significant ($p<0.01$ with hypergeometric test).

Gene pathways: The proposed method is founded on the fact that the ontology structure reflects semantic relationships among gene functions. Gene ontology is basically a semantic understanding of the functions and processes performed by genes of various organisms. The structure of GO relates semantic, each edge is an *is-a* relationship. Moreover, the gene ontology was developed incrementally over the years by adding new functions (*ontology terms*) in positions that can be the most *semantically* appropriate, for each term, within the ontology structure to maintain the meaningful *is-a* buildup. Therefore, we can utilize it to infer semantically similar concepts and semantically distant concepts. Then, two functions encoded in the ontology (GO) with the shortest distance between them exceeding some (*specificity level*) threshold can be considered diverse functions and any gene annotated with them can be classified as multifunctional gene. For a given gene *g*, our method relies on the fact that if the shortest distance between two functions of *g* is greater than some specificity level threshold, T_s , (*e.g.*, we examined with T_s =14) then gene *g* is multifunctional so long as the two function do not have genes in common more than expected by chance. In [29], Wang et. al confirm that only ontology structure of the GO can be an indicator of the semantic similarity of gene functions [29]. With that, we would like to attempt a different kind of evaluation using gene pathways. We want to examine our method in a different way. We use *Kegg* database for gene pathways [34] for this analysis of the semantic distance among gene functional annotations and gene multifunctionality as follows:

(1) Genes that belongs to more pathways are more highly likely to be multifunctional compared with genes that belong to only one pathway in *Kegg*. We extracted two sets of genes: The first set consists of genes participating (each) in only one *Kegg* pathway, and the second set contains genes that each gene participates in at least three pathways. We analyzed the shortest path length among all *bp* annotations of each gene in both sets and the results are shown in Figure_7.

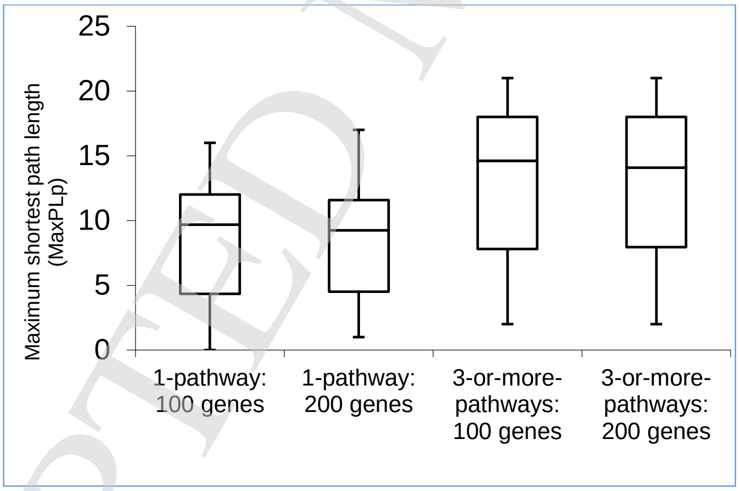

Fig. 7: The results of maximum shortest path length among the *bp annotations* (MaxPLp()) of four sets of genes. Each test is done with 100 or 200 genes have either only one pathway or at least three pathways.

(2) Secondly, we extracted pairs of genes that participate in highly diverse pathways in Kegg. Each gene pair *p* consists of two genes g_1 and g_2 (i.*e*., $p=(g_1, g_2)$) such that g_1 and g_2 belongs to two highly different pathways. Thus we consider the pair *p* representing a true multifunctionality. We extract all *bp* annotations of both genes g_1 and g_2 of each pair and

analyze their maximum path length and compare this to normally annotated genes; the results are shown in Figure 8. Finally, these evaluations using gene pathways from *Kegg* also are significant (p<0.01) for all sets of 100 and 200 genes with one pathways and with three pathways shown in both Figures 7 and 8.

In conclusion, this paper presents a well-defined study of multifunctional genes in the entire human genome using gene functional annotations and the GO. The work in this paper relies solely on the ontology structure of the GO irrespective of number of annotations per gene. This direction is to increase our knowledge and understanding of gene and protein functions in human cells. The proposed method was verified and evaluated against several criteria used commonly for this task as predictors for gene multifunctionality including gene pathways.

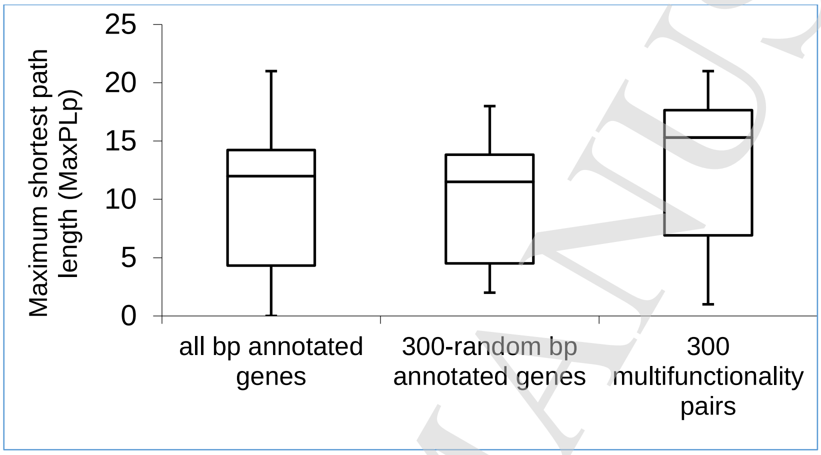

Fig. 8: Distribution of the maximum shortest path length (MaxPLp()) for the three sets of genes including the set of all human bp annotated genes.

References

1. Pritykin,Y. et al. (2015) Genome-Wide Detection and Analysis of Multifunctional Genes. PLOS Computational Biology, 11, e1004467.
2. Peppel,J.v-de and Holstege,F. CP. (2005) Multifunctional genes. Molecular Systems Biology, doi:10.1038/msb4100006
3. Gillis,J. and Pavlidis,P. (2011) The impact of multifunctional genes on 'guilt by association' analysis. PloS One, 6, e17258.
4. Al-Mubaid,H. et al. (2016) Assessing Gene-Disease Relationship with Multifunctional Genes Using GO. Proc. of IEEE AICCSA.
5. Hernandez,S. et al. (2014) MultitaskProtDB: a database of multitasking proteins. Nucleic Acids Research 42, D517–D520.
6. Mani M., et al. (2015) Moonprot: a database for proteins that are known to moonlight. Nucleic Acids Research 43, D277–D282.
7. Day, A. et al. (2009) Disease Gene Characterization through Large-Scale Co-Expression Analysis. PLoS ONE, 4.
8. Khan, I., et al. (2014) Genome-scale identification and characterization of moonlighting proteins. Biology Direct 9, 30.
9. Ashburner et al. (2000) Gene Ontology: tool for the unification of biology Nat Genet 25, 25-9.
10. The Gene Ontology Consortium. (2015) Gene Ontology Consortium: going forward. Nucl Acids Res 43, D1049–D1056.

11. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (2016) Online Mendelian Inheritance in Man, OMIM. World Wide Web URL: http://omim.org/

12. Ballouz,S. et al. (2017). Using predictive specificity to determine when gene set analysis is biologically meaningful. Nucleic Acids Res, 45, e20.

13. Khan,I.K. and D. Kihara, D. (2016). Genome-scale prediction of moonlighting proteins using diverse protein association information. Bioinformatics, 32, 2281-8..

14. Kim,J. et al. (2017) An analysis of disease-gene relationship from Medline abstracts by DigSee. Scientific Reports, 7, 40154.

15. Salathe,M. et al. (2006) The effect of multifunctionality on the rate of evolution in yeast. Molecular Biology and Evolution 23, 721–722.

16. Clark,W.T. and Radivojac,P. (2011) Analysis of protein function and its prediction from amino acid sequence. Proteins: Structure, Function, and Bioinformatics 79, 2086–2096.

17. Nagar,A. and Al-Mubaid,H. (2015) A Hybrid Semantic Similarity Measure for Gene Ontology Based On Offspring and Path Length. Proc. of IEEE CIBCB-2015 IEEE Conf. on Comp. Intelligence in Bioinformatics and Comp. Biology.

18. Glass,K. and Girvan,M. (2015) Finding New Order in Biological Functions from the Network Structure of Gene Annotations. PLoS Comput Biol., 11.

19. Singh-Blom,U.M.et al. (2013) Prediction and Validation of Gene-Disease Associations Using Methods Inspired by Social Network Analyses. PLOS ONE, 8 (9).

20. Schriml,L.M. and Mitraka,E. (2015) The Disease Ontology: fostering interoperability between biological and clinical human disease-related data. Mamm Genome, 26, 584–589.

21. Zhou and J. Skolnick. (2016) A knowledge-based approach for predicting gene–disease associations. Bioinformatics; 32, 2831–2838.

22. Schaefer M.H.. et al. (2012) HIPPIE: Integrating protein interaction networks with experiment based quality scores. PLoS One, 7.

23. Hippie PPI database: http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/

24. Becker E. et al. (2012) Multifunctional proteins revealed by overlapping clustering in protein interaction network. Bioinformatics 28, 84–90.

25. Hernandez,S. et al. (2011) Do moonlighting proteins belong to the intrinsically disordered protein class? Proteomics Bioinformatics, 5, 262–264.

26. The National Center for Biotechnology Information, NIH, USA: https://www.ncbi.nlm.nih.gov

27. NCBI. Clearing Up Confusion with Human Gene Symbols & Names Using NCBI Gene Data. NIH, USA. (2016) https://ncbiinsights.ncbi.nlm.nih.gov/2016/11/07/clearing-up-confusion-with-human-gene-symbols-names-using-ncbi-gene-data/.

28. QuickGO; http://www.ebi.ac.uk/QuickGO

29. Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007; 23(10):1274–1281.

30. H. Al-Mubaid et. al. Determining Multifunctional Genes and Diseases in Human Using Gene Ontology. Proceedings of 9th Int'l Conf on Bioinformatics and Computational Biology BICOB-2017, Honolulu, HI, USA, March 2017.

Appendix:

Table 0. Avg $MaxPL_f$ with clusters of 1000 genes in each cluster.

After sorting all genes based on MaxPL$_f$() (descending order)	**mean MaxPL$_f$()**

Top 1000	13.987
1001 – 2000	12.189
2001 – 3000	11.250
3001 – 4000	10.427
4001 – 5000	9.576
5001 – 6000	8.487
6001 – 7000	7.505
7001 – 8000	6.336
8001 – 9000	2.030
9001 – 10000	3.189
10001 – 11000	0.880
11001 – 12000	0.498
Lowest 1191	0
Total number of genes with mf annotations: 13,191. Number of genes with only 1 mf term: 2630 (i.e., MaxPL$_f$ = 0)	

Journal of Bioinformatics and Computational Biology
© Imperial College Press

GENE MULTIFUNCTIONALITY SCORING USING GENE ONTOLOGY

HISHAM AL-MUBAID*

Computer Science Department, University of Houston-Clear Lake
Houston, TX 77062, USA†
hisham@uhcl.edu

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Multifunctional genes are important genes because of their essential roles in human cells. Studying and analyzing multifunctional genes can help understand disease mechanisms and drug discovery. We propose a computational method for scoring gene multifunctionality based on functional annotations of the target gene from the Gene Ontology. The method is based on identifying the pairs of GO annotations that represent semantically different biological functions and any gene annotated with these two annotations is considered multifunctional. The proposed method can be employed to identify multifunctional genes in the entire human genome using solely the GO annotations. We evaluated the proposed method in scoring multifunctionality of all human genes using four criteria: gene-disease associations; protein-protein interactions; gene studies with PubMed publications; and published known multifunctional gene sets. The evaluation results confirm the validity and reliability of the proposed method for identifying multifunctional human genes. The results across all four evaluation criteria were statistically significant in determining multifunctionality. For example, the method confirmed that multifunctional genes tend to be associated with diseases more than other genes, with significance $p<0.01$. Moreover, consistent with all previous studies, proteins encoded by multifunctional genes, based on our method, are involved in protein-protein interactions significantly more ($p<0.01$) than other proteins.

Keywords: multifunctional genes, functional genomics.

1. Introduction

Studying and understanding the functions ~~that a gene is involved in~~of all genes in a genome is a central step in functional genomics [1, 2, 4, 3]. In particular, multifunctional genes are important to study as they convey essential roles in an organism and in the human genome in particular [1, 2, 3, 4]. A gene is multifunctional if it is involved in more than one distinct function in human body. Studying and uncovering multifunctional genes ~~is~~ are important tasks for various fields like gene-disease associations, drug discovery, and functional genomics studies.

In this paper, we study human genes in the entire human genome to examine gene multifunctionality and identify the most likely multifunctional genes. Determining if a gene is multifunctional is not a trivial task as many genes conduct more than one functionality. A gene involved in two functions may not be a multifunctional if the two functions are not distinct (i.e., not ~~diverse~~semantically different) enough [1]. In this work, we use a computational methodology to determine whether or not a gene is multifunctional ~~with distinct~~and involved in two or more (semantically different) functions. Specifically, we present a method based on the functional annotations of the gene from the Gene Ontology (GO) for examining gene multifunctionality. We use the GO annotations from the biological process (*bp*) and molecular function (*mf*) aspects of the GO. The proposed ~~gene multifunctionality~~ method extracts and examines all ~~paths between all *pb* and *mf*~~ functions and processes that a gene is annotated with. The method is based on identifying pairs of GO annotations such that each pair represents two semantically different biological functions; hence any gene annotated with two annotations of one pair is considered multifunctional. We examined the proposed methods in ~~scoring and~~ estimating the multifunctionality of all genes in the human genome. We evaluated the results with four different criteria as compared with previous related work in this problem. The four evaluation criteria are: –gene-disease association; –protein-protein interactions *PPI*; –gene studies and *PubMed* publications; and –using published sets of confirmed multifunctional genes. The evaluation results are encouraging and prove that both scoring methods are valid and reliable indicators of gene multifunctionality across the four evaluation criteria. For example, the proposed method~~s~~ confirms that multifunctional genes tend to be associated with diseases more than other genes in the same annotation population, with significance $p<0.01$, as also proved by previous studies. Moreover, consistent with all previous studies, proteins encoded by multifunctional genes, based on our method, are involved in PPI interactions significantly more (p<0.01, *hypergeometric test*) than other proteins.

2. Background and Related Work

One of the important characteristics of multifunctional genes that motivate more investigations is the gene-disease association. This kind of association is significantly higher in multifunctional genes compared to all genes as confirmed by all previous studies in this domain [1, 2, 8, 5-7]. Therefor the relationships between diseases and multifunctional genes are significant and proved [1, 7].

A multifunctional gene is a gene that is involved with several functions and activities, including molecular and cellular tasks, inside the cell [1, 2, 8, 5]. Typically, studying multifunctional genes can disclose more knowledge about diseases associated with the multifunctional genes. In this paper, we rely on the gene ontology (GO) which is the most popular repository of functional information about human genes [9, 10]. Pritykin, Ghersi, and Singh (2015) presented a comprehensive study of genome-wide multifunctional genes in human [1]. They found that multifunctional genes are significantly more likely to be involved in human disorders [1]. Also, they found that 32% of all multifunctional genes produced by their method are involved in at least one OMIM disorder (http://omim.org), whereas the fraction of other annotated genes involved in at least one OMIM disorder is 21% [1, 11].

Ballouz, Pavlidis, and Gil (2017) studied various gene sets for functional genomics and enrichment [3]. They found that heavily functional genes are highly likely to appear in many genomic study results [12]. They leave it as an 'open question' to biologist to assess if their finding of gene multifunctionality is a true biological property. Khan and Kihara (2016) extracts a domain of features including GO, protein-protein interaction, and more, to classify protein into moonlighting (i.e. multifunctional) versus non-moonlighting proteins [13]. Kim et al.

(2017) in their system, DigSee, found that genes that interact with more genes in a PPI network are involved in more disease categories than those with fewer neighbors in the protein interaction network [14].

Salathe et al. (Salathe et al., 2006), investigated the multifunctionality of yeast genes and proteins for a different goal [15]. They found a positive correlation between how many biological process (bp) GO terms a gene is annotated with and its evolutionary conservation in yeast; that is, they found highly significant negative correlation between number of bp GO terms and rate of change of yeast genes [15]. Also, Pritykin et al. observed that multifunctional genes tend to be more evolutionarily conserved [1]. A method for identifying novel moonlighting proteins from current functional annotations in public databases was proposed by [8]. They identified potential moonlighting proteins in the Escherichia coli K-12 genome by examining clusters of GO term annotations taken from *UniProt* and constructed three datasets of experimentally confirmed moonlighting proteins (Khan et al., 2014) [8]. In another study of multifunctional genes, Clark et al. (2011) [16] discovered a statistically significant positive correlation between the number of GO biological process leaf terms a gene has and its number of *Pfam* domains and are usually longer [16].

3. Materials and Methods

The proposed gene multifunctionality method is based on the ~~set of~~ annotation terms from the GO ~~for each target gene~~. The GO is highly regarded as the main source for gene functional information [4, 29]. It is the largest and most comprehensive source of information on gene functions with contents growing and becoming more accurate every day.

We can utilize the GO along with the functional genomics data sets for human to induce the relationships among the functions encoded in the ontology. For example, the path length between two GO terms has been extensively used as a metric in computing semantic similarity among genes [4, 17]. Moreover, many gene similarity measures use the depth of the lowest common subsumer (LCS) in computing gene similarity [18, 17]. In our previous work, we investigated and explained the relationship between GO annotation terms of a gene and gene-disease associations [4]. This paper proposes a new method derived from the gene ontology for identifying multifunctional genes in the entire human genome. The proposed method is based on identifying each pair of GO terms that represents a multifunctionality (semantically different biological functions).

Typically, the similarity between two genes is computed as a function of the similarity of their annotations mainly using the *bp* and the *mf* aspects. That is, the similarity $Sim_g(g_1, g_2)$ between two genes g_1 and g_2 can be a similarity function between the annotations of g_1 and g_2:

$$Sim_g(g_1, g_2) = Sim_t(t_{1i}, t_{2i}) \ldots\ldots\ldots\ldots\ldots (1)$$

where $Sim_g(g_1, g_2)$ is the gene similarity between g_1 and g_2; and $Sim_t(t_{1i}, t_{2i})$ is ~~the~~ a similarity function between GO terms t_{1i}, t_{2i} annotating g_1 and g_2 respectively.

The gene ontology consists of 3 aspects: Molecular Function *mf*, Biological Process *bp* and Cellular Component *cc*. Each one of these aspects {*mf*, *bp*, *cc*} is a complete ontology in itself (www.geneontology.org; [10, 4, 17]). For gene multifunctionality, it is normal to rely only on the *bp* and *mf* aspects.

Let $MaxPL_p(g_x)$ be the maximum path length between all pairwise *bp* annotation terms of gene g_x; that is:

$$MaxPL_p(g_x) = \max_{t_x, t_y \in GOT_p(g_x)} PL(t_x, t_y) \quad \ldots\ldots\ldots (2)$$

where $\mathrm{PL}(\mathrm{t_x}, \mathrm{t_y})$ is the *shortest* path length between the two annotations t_x and t_y, and $GOT_p(g_x)$ is the set of all *bp* annotations of gene $\mathrm{g_x}$ (and none them subsumes any term; i.e., none of the annotations in $GOT_p(\quad)$ is a descendant of any other term in the same set). For example, in Figure 1, there are two different paths shown between GO:0000001 and GO:0006996 one of them is of length 2 (through GO:0048308) and the second path is of length 3 (through the two GO terms GO:0048311 and GO:0007005); we take 2 as the shortest path length between them. The multifunctionality of a gene increases with the increase in the distinctiveness (i.e., diversity) of the functions that the gene in involved in [1]. The path length between two bp annotations of a target gene can be utilized as an indicator of the distinctiveness of the functions ~~that the gene is part of~~of that gene. Based on this, we employ $MaxPL_p$ in a multifunctionality method based on the maximum shortest path length between the *bp* annotation terms.

In the biological process (*bp*) aspect of GO, each annotation term is basically a node in the ontology graph (which is a directed acyclic graph *DAG*) and is a biological functionality upheld by certain genes ~~and proteins~~ [28]. When two *bp* annotations (i.e., graph nodes) are far apart with relatively large path length between them (exceeds a threshold) then we can consider that these two terms represent two distinct (~~diverse~~semantically different) biological functionalities. That is, our hypothesis is that, two highly far apart bp annotation terms can be considered as two distinct ~~gene~~ functions given that neither of these two terms is subsuming the other. Therefore, a gene annotated with two such terms can be considered ~~as~~ multifunctional. The computations of multifunctionality scores with bp annotations for human genes go through the algorithm shown in Figure A1.

```
Algorithm 1: Compute multifunctionality scores for all human genes using bp
             annotations
Input: - GOA_human: set of all human gene annotations.
       - GO.obo: set of all gene ontology annotation terms with their parents
Output: - Set {$\mathrm{MaxPL_p(g_x)}$ }: multifunctionality score for every human gene $\mathrm{g_x}$ based
          on bp annotations.
Algorithm:
  (1) Create the set G
      1a) $\mathrm{G} = \emptyset$ : let G be the set of all genes annotated in GOA_human
      1b) For each annotated gene $\mathrm{g_i}$ from the set GOA_human:
          i) $\mathrm{G = G \cup g_i}$  : add $\mathrm{g_i}$ to G
  (2) Create the set BP
      2a) $\mathrm{BP} = \emptyset$ : let BP be the set of all bp annotation terms in GO.obo
      2b) For each bp annotation term $\mathrm{t_i}$ in GO.obo:
          i) $\mathrm{BP = BP \cup t_i}$  :add $\mathrm{t_i}$ to BP along with its parents
  (3) Create the set GOA_human_bp
      3a) Extract all bp annotations from GOA_human and add them to
          GOA_human_bp
  (4) For each gene $\mathrm{g_x}$ in the set G
      4a) Extract the set $\mathrm{GOT_p(g_x)}$ of all annotations of $\mathrm{g_x}$ from GOA_human_bp
      4b) Set $\mathrm{MaxPL_p(g_x)} = 0$
      4b) If $|\mathrm{GOT_p(g_x)}| < 2$ go to step (4)   :go to step (4) up from beginning
      4c) For each pair $\mathrm{t_i, t_j}$ of annotation terms in $\mathrm{GOT_p(g_x)}$:
          i) Compute the shortest path length $\mathrm{PL(t_i, t_j)}$ between pair $\mathrm{t_i, t_j}$ using the set
             BP
          ii) If $\mathrm{PL(t_i, t_j) > MaxPL_p(g_x)}$ then set
```

$$\mathrm{MaxPL_p}(\mathrm{g_x}) = \mathrm{PL}(\mathrm{t_i}, \mathrm{t_j})$$

Fig. A1. Algorithm for *MaxPLp()* for all human genes

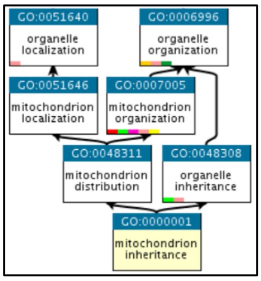

Fig. 1. ~~a~~ A small part of the GO.

This algorithm (Algorithm A1) explains the steps of the method.

Now among shortest path length of all-pairwise annotations of a target gene *g*, we take the maximum one as an indicator of multifunctionality. Considering the *bp* aspect of GO, the maximum value of path length in the *bp* graph is 21 which is indicated for two genes (gene Ids: 672 and 5071) as follows:

Gene Id	Gene MIM	UniProtKB	Gene Symbol	No. of Phenotypes	No. of *bp* terms	Max *bp* path length
672	113705	P38398	BRCA1	2	59	21
5071	602544	O60260	PRKN	4	100	21

To further investigate and utilize functional gene GOA datasets along with semantic distance in GO we used *mf* annotations to compute multifunctionality based on *mf* annotation terms as:

$$MaxPL_f(g_x) = \max_{t_x, t_y \in GOT_f(g_x)} PL(t_x, t_y) \quad \text{........ (3)}$$

where $\mathrm{GOT_f(g_x)}$ is the set of all *mf* annotations of gene $\mathrm{g_x}$. We conducted the evaluations based on both *bp* and *mf* separately; *details in the next section*.

In an empirical computational work, we tried to vary the contribution of *bp* and *mf* by employing a contribution factor α and adopting as a multifunctionality { α .$\mathrm{MaxPL_p}$ (.) + (1 – α).$\mathrm{MaxPL_f}$ (.)}. Then, we experimented with a number of values for α between 0.3 to 0.7 with 0.1 increase. We found there is no significant influence in the results from the middle value of α=0.5 which allowed us to eliminate it. In the evaluation and results section, we report the results of using both *mf* and *bp* separately. Moreover, we ~~mention~~ estimate that a gene is considered ~~(or can be considered)~~ multifunctional if the *MaxPLp* is exceeding certain threshold (specificity level). This threshold is typically >10. We found that when the threshold is >10 (*i.e., MaxPLp>10*) the fine/coarse grained of the ontology branches do not matter and make no difference, just like the depth of the nodes. The depth of the (*LCS of*) two nodes makes difference only when the path length is relatively small (e.g., <5); and similarly when considering path length > 10, it becomes irrelevant whether coarse-grained (towards the top of the ontology tree) or fine-grained (towards leaves) branches are used.

Contribution: There is a real need for computational methods to provide insights in understating gene functions in the human genome. And identifying multifunctional genes is a central step because of their essentiality as major players in most ~~processes, and~~ functionalities in human cells. The Gene Ontology along with the functional genomics ~~extensive~~ databases have not been extensively investigated in computational methods within the domain of multifunctional genes. This work is based on the Gene Ontology which is the most comprehensive, and perhaps the main, source for gene functional information. Gene Ontology is a structure and vocabulary of semantic understanding of the various functions that genes can perform. This work utilizes this semantic structure of gene functions that has been built carefully over the years to induce insight in understanding and identifying ~~multifunctional~~ genes functions.

4. Evaluation and Results

For ~~each~~ all human genes, we extracted all ~~its~~ annotation terms from the Gene Ontology Annotation (GOA) database for human [4]. By considering only *bp* annotations, we found a total of ~35,700 genes annotated with at least one *bp* terms. Overall, there are ~5.2 bp annotations per gene. By considering *mf*, there are on average 4.3 mf annotations per gene with a total of ~35,800 genes annotated with at least one *mf* term. Among all genes with *mf* annotations (~35,800 genes) in GOA database, almost 42% of them (or 15,142 genes) are annotated with only one *mf* terms. Each gene with only one *mf* annotation will have MaxPL$_{f}$ = 0. Therefore, in *mf* we have 42% of the genes do not count in the computations of the multifunctionality scoring. For all genes with two or more *bp* terms, we extracted all *bp* annotations for each gene from the *GOA* database. In the human annotation dataset GOA_human, ~80% of the genes (= ~29,000 genes) have 4 or fewer *mf* annotations. We computed the maximum all-pair shortest path length among all terms for every gene as per our method. For evaluation, we would like to verify the reliability of our multifunctionality scoring techniques, MaxPL$_{p}$ and MaxPL$_{f}$, in estimating the whether or not a gene is multifunctional. We could not find any gold standard dataset to evaluate our methods. So, we used four criteria for multifunctionality [1, 13]. These four criteria are: (1) Gene-disease association is more in multifunctional genes compared with other non-multifunctional genes; (2) Multifunctional genes are more evolutionary conserved; (3) Multifunctional genes tend to be highly studied with relatively higher number of publications; and (4) Using previously tested and published multifunctional gene sets as criteria to test our method.

We analyzed all human genes having *bp* or *mf* annotations using the proposed system. After computing MaxPL$_{f}$(g$_{x}$) value for each gene, we grouped all genes into clusters of 1000 genes in each cluster after being sorted based on MaxPL$_{f}$(g$_{x}$); shown in Table 0 (in the appendix). For example, the top 1000 genes have an average MaxPL$_{f}$(g$_{x}$) of 13.99 whereas the next cluster (next 1000 genes) have MaxPL$_{f}$ average of 12.189; (*see Table 0 in the appendix*).

Criteria 1. Gene-disease association:

Multifunctional genes are more highly likely to be associated with human diseases than non-multifunctional genes [1, 2, 8, 4, 14, 19]. We analyzed all human genes from the GOA database and from *OMIM* morbid map for disease information (http://omim.org/) [11]. Also, for identifying number of phenotypes per gene, we used the disease ontology (DO) to check whether the two phenotypes are distinct [20]. We wanted to investigate if the number of phenotypes, according to *morbid map*, exhibits any meaningful relationship with our multifunctionality method. We firstly examined the components of the multifunctionality scoring method in equation (4), namely MaxPL$_{p}$ and MaxPL$_{f}$ independently.

The results in Table 1 show the correlation between MaxPLp and average number of phenotypes for all human genes; these results are also illustrated in Figure 2. Next, we examined MaxPL$_p$ for each group of genes associated with the same number of phenotypes and the results are in Table 2 and Figure 3. For example, there are 2,572 genes associated with only one phenotype and their average *MaxPLp* is 11.22 whereas the group of genes associated with exactly two phenotypes (648 genes) have an average *MaxPLp* 12.33; Table 2. We mention here that groups of genes associated with ≥7 phenotypes are very small and do not affect the results. For example, there are only 11 genes associated with 7 diseases, and only 7 genes associated with 8 diseases.

We repeated the same evaluation for MaxPL$_f$ (i.e., using *mf* annotation terms) and the results are in Table 3 and also illustrated in Figure 4. As it is shown in both Table 3 and Figure 4, ~~as~~ the MaxPL$_f$ increases as the average number of associated phenotypes increases; thus, there is a clear strong correlation between MaxPL$_f$ and average number of phenotypes. Hence, our MaxPL$_f$ is a reliable indicator of multifunctionality of genes. Next, we examined the behavior of MaxPL$_f$ with the increase of phenotypes (i.e., number of phenotypes is independent variable x-axis) for all human genes and the results are shown in Table 4.

Table 1. For each value of MaxPLp this table shows how many genes and the average number of phenotypes

MaxPLp	No. of genes	Avg. of No. of phenotypes
0	1399	0.15
1	71	0.32
2	141	0.21
3	158	0.19
4	204	0.26
5	266	0.21
6	356	0.35
7	497	0.31
8	578	0.31
9	733	0.31
10	919	0.35
11	1319	0.32
12	1490	0.36
13	1541	0.45
14	1512	0.48
15	1166	0.58
16	725	0.64
17	455	0.85
18	217	0.76
19	113	0.73
20	13	1.00
21	2	3.00

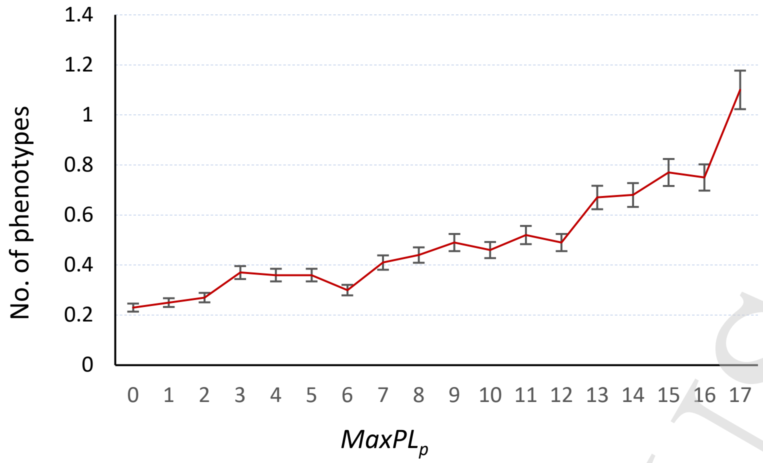

Fig. 2. The relationship between MaxPL$_p$ and number of diseases for all human genes.

Criteria 2. Protein-protein interactions:

Multifunctional genes are typically involved more than normal in protein-protein interactions PPI's [1, 21, 22, 14, 3]. We used this criterion in evaluating our method. We retrieved and compiled PPI data from the Hippie database (http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/) [22]. The obtained data include PPI data involving ~14,800 genes with a total of ~250K experimentally documented P-P interactions [21, 22, 23]. We analyzed the average number of PPI's that a gene involved in with respect to MaxPL$_p$ (and *MaxPL$_f$*) and there is a clear relationship as illustrated in Figure 5. These results prove again that the proposed methods are in line and consistent with this criteria for gene multifunctionality.

Table 2. The average value of MaxPLp for six groups of genes where each group have the same number of phenotypes.

No. of phenotypes	No. of genes	Avg. of MaxPLp
0	10234	10.09
1	2572	11.22
2	648	12.33
3	226	12.96
4	84	13.67
5	51	13.86
6	29	15.45

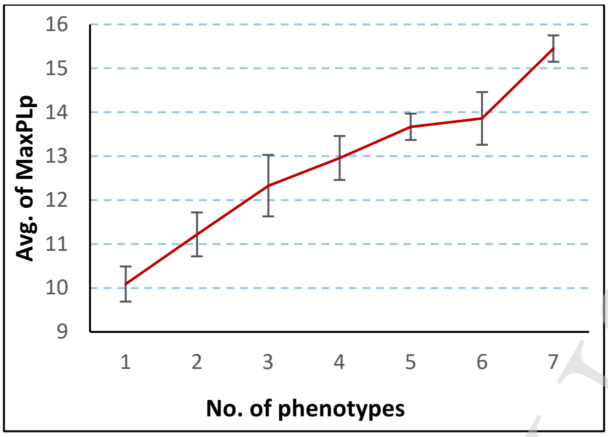

Fig. 3. The relationship of the MaxPLp for each value of average number of phenotype

Table 3. Avg No. of phenotypes for genes with each value of $MaxPL_f$

MaxPL$_f$	No. of genes	Avg. of No. of phenotypes
0	2630	0.23
1	244	0.26
2	490	0.27
3	467	0.37
4	557	0.36
5	578	0.36
6	890	0.30
7	831	0.41
8	1018	0.44
9	911	0.49
10	1148	0.46
11	1182	0.52
12	1057	0.49
13	512	0.68
14	475	0.68
15	126	0.77
16	53	0.66
17	10	1.10
18	11	0.73
19	1	0.00

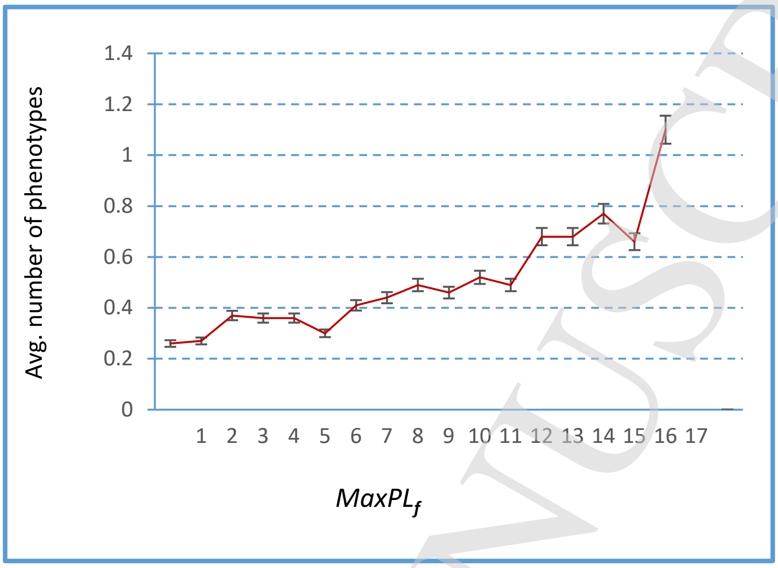

Fig. 4. Average number of phenotypes increases as a function of MaxPL$_f$

Table 4. For each number of phenotypes this table shows number of genes, average number of mf annotations, and average MaxPL$_f$ (e.g., the first row shows that there are 9720 genes having no phenotypes association (0) and having an average MaxPL$_f$ of 6.36)

No. of phenotypes	No. of genes	Avg. MaxPL$_f$
0	9720	6.36
1	2442	7.40
2	619	8.39
3	221	9.13
4	81	8.40
5	48	9.31
6	29	10.17
7	11	7.45
8	7	11.14
9	3	7.67
10	2	12.5
11	4	11.75
12	1	3
13	1	11
14	1	11
16	1	7

Criteria 3. Using *PubMed* publications as indicator of highly studied multifunctional genes: It has been shown that multifunctional genes are highly studied genes and have relatively more publications in the biomedical literature [1, 12]. So, we use publication counts of genes as a criteria of multifunctionality. That is, multifunctional genes tend to have relatively higher

number of publications compared to other annotated genes. We relied on *PubMed* since it is the most comprehensive repository of biomedical literature with more than 24 million citations and references to articles (with abstracts, and some

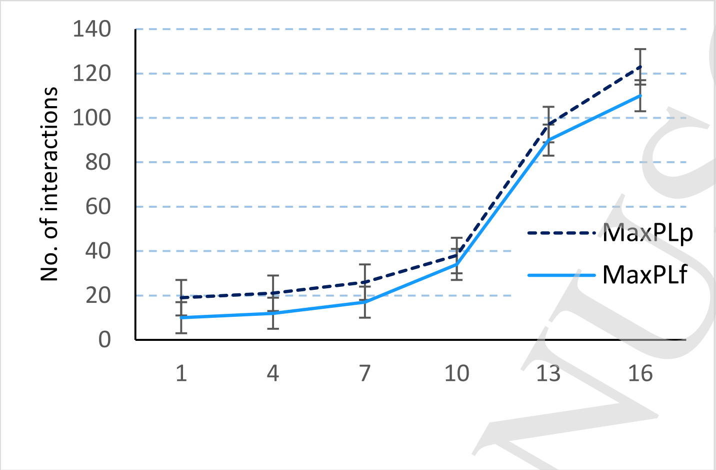

Fig. 5. Number of protein-protein interactions increases as the gene multifunctionality score increase. ~~Note:~~ X-axis here represents MaxPL$_p$ (or MaxPL$_f$) value, and Y-axis represents average number of interactions.

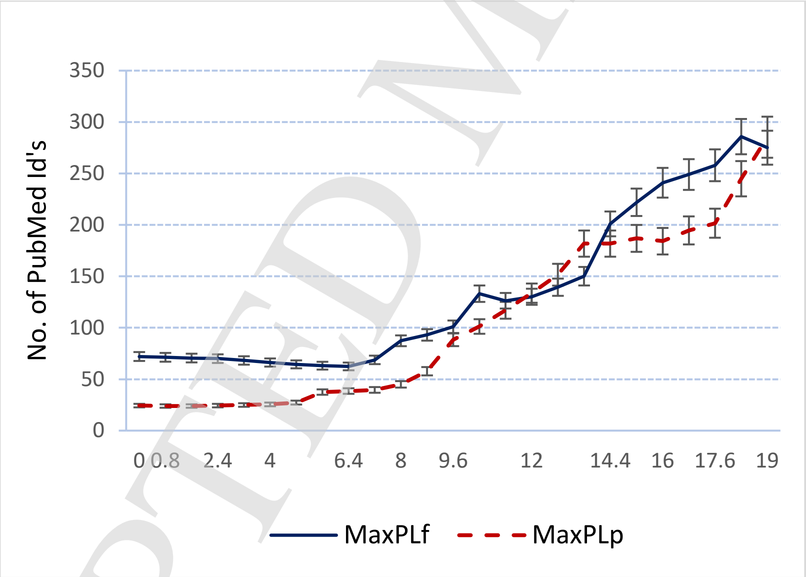

Fig. 6. Analyzing number of PubMed publications against multifunctionality scores with both MaxPL$_f$ and MaxPL$_p$ for all human genes show a direct proportional relationship. The X-axis here represents MaxPL$_p$ (or MaxPL$_f$) value.

with full texts). We analyzed number of publications related to each gene in *PubMed* as it is published by NCBI/PubMed and freely available with file name: gene2pubmed.gz, (link: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA; downloaded Sept.2017) [26, 27].

We examined genes with our multifunctionality scores versus number of publications. The analysis results show a clear straightforward proportionality between number of publications and both scoring methods (MaxPL_f and MaxPLp) for all human genes as illustrated in Figure_6.

Criteria 4. Using published multifunctional genes:

We retrieved two lists of experimentally tested and known multifunctional genes [1].

Source 1: http://moonlightingproteins.org/proteins/ which includes 361 moonlighting proteins (74 for human).

Source 2: http://wallace.uab.es/multitask/ which includes 288 proteins (88 of them for human).

This test was not reliable as we are considering only 162 (74 from set1 and 88 from set2) human genes (out of ~35000 annotated genes); however, these genes exhibit higher MaxPL_f and MaxPL_p values than expected by chance with significance ($p<0.01$) confirming multifunctionality.

5. Discussion

The results in Tables 1 and 2 confirm that there is a direct proportional relationship between gene multifunctionality scores and the number of diseases associated with gene. For example, genes with MaxPL_f =2 (490 genes) have on average 0.27 associated diseases whereas genes with MaxPL_f =3 (467 genes) are on average associated with 0.37 diseases; and this is significant $p<0.01$ (*using hypergeometric test*) as shown in the following excerpt of Table 3:

MaxPL_f	No. of genes	Avg. No. of phenotypes
2	490	0.27
3	467	0.37

Also, the average number of phenotypes for 1182 genes with MaxPL_f of 11 is 0.52, and when the MaxPL_f increases to 13 the average number of phenotypes increases to 0.68 which is significant result ($p<0.01$) as shown in the following excerpt of Table 2:

MaxPL_f	No. of genes	Avg. of No. of phenotypes
11	1182	0.52
13	512	0.68

From Table 3, we can see for genes with one phenotype (2442 genes) the average MaxPL_f is 7.40 whereas for those genes with 2 phenotypes the average MaxPL_f increases to 8.39 and this is significant with $p<0.01$ as shown in the following excerpt of Table 3:

No. of phenotypes	No. of genes	Avg. MaxPL_f
1	2442	7.40
2	619	8.39

Similarly, for MaxPLp, we see that the 10234 genes associated with 0 phenotypes have MaxPLp average of 10.09 and for the genes associated with one phenotype (2572 genes) the value of MaxPLp increases to 11.22 which is significant $p<0.01$ as shown in the following excerpt of Table 5:

No. of phenotypes	No. of genes	Avg. MaxPLp
0	10234	10.09
1	2572	11.22

Other results: we analyzed some fairly well known disease genes to examine the behavior of our proposed scoring method with these particular genes. Clearly all of them are multifunctional (equations (4) – (6)) as shown below:

Gene	Gene Id	MaxPL_f	MaxPL_p	mfs
SNCA - Parkinson disease	6622	13	15	0.70
BRCA2 - breast cancer gene	675	15	14	0.73
TP53 - tumor protein	7157	12	19	0.78
BRCA1 - tumor protein	672	11	21	0.80
APP - Alzheimer disease AD	351	11	17	0.70

Table 5. The multifunctionality scores based on mf annotations, MaxPL_f, show clear difference between genes associated with phenotypes (MaxPL_f =9.02) versus gene not associated with any phenotype (MaxPL_f = 6.36)

With mf annotations only		
Genes with 0 phenotype	Avg. MaxPL_f 6.36	No. of genes: 9720
Genes with ≥1 phenotypes	Avg. MaxPL_f 9.02	No. of genes: 3471
Genes with 0 or 1 phenotypes	Avg. MaxPL_f 6.88	No. of genes: 12162
Genes ≥ 2 phenotypes	Avg. MaxPL_f 9.14	No. of genes: 1029
Overall MaxPL_f Avg: 6.733		

Table 6. MaxPLp shows clear differentiation between genes associated with phenotypes versus those genes not associated with any phenotypes (similar to Table 6 above).

With bp annotations only		
Genes with 0 phenotype	Avg. MaxPLp 10.09	No. of genes: 10234
Genes with ≥1 phenotypes	Avg. MaxPLp 14.23	No. of genes: 3641
Genes with 0 or 1 phenotypes	Avg. MaxPLp 10.65	No. of genes: 12806
Genes ≥ 2 phenotypes	Avg. MaxPLp 14.45	No. of genes: 1069
Overall MaxPLp Avg: 10.50		

Further, the results in the last two tables (Table 5 and Table 6) prove the significant direct proportionality relationship between our multifunctionality methods and disease association. Regarding number of publications, criteria 3, we confirmed that as our multifunctionality score of a gene tend to increase the number of PubMed publications related to the gene also increases as illustrated in Figure 6. We should mention here that higher number of publications implies that the gene is highly studied [1]. One of the main reason of being highly studied is the gene is highly likely associated with one or more diseases. We should mention here that there are genes with fairly high number of publications but with low (≤ 7) multifunctionality score for which reason we relied on the aggregate averages. For example, considering genes with MaxPLp of 12; their average number of PubMed publications is ~133; when we increase the score to 14 the average increases to ~181 and this is significant ($p<0.01$). Finally, we assume multifunctional are genes with MaxPLp≥15, we got 2691 multifunctional genes (genes having MaxPLp of 15 or more). Among these 2691 genes, we found 46% (or 1231 genes) of them are also *mf* multifunctional with mf annotations only using threshold T_f =10 (i.e., MaxPL$_f$ ≥ 10), and this is significant (*p<0.01* with hypergeometric test).

Gene pathways: The proposed method is founded on the fact that the ontology structure reflects semantic relationships among gene functions. Gene ontology is basically a semantic understanding of the functions and processes performed by genes of various organisms. The structure of GO relates semantic, each edge is an *is-a* relationship. Moreover, the gene ontology was developed incrementally over the years by adding new functions (*ontology terms*) in positions that can be the most *semantically* appropriate, for each term, within the ontology structure to maintain the meaningful *is-a* buildup. Therefore, we can utilize it to infer semantically similar concepts and semantically distant concepts. Then, two functions encoded in the ontology (GO) with the shortest distance between them exceeding some (*specificity level*) threshold can be considered diverse functions and any gene annotated with them can be classified as multifunctional gene. For a given gene *g*, our method relies on the fact that if the shortest distance between two functions of *g* is greater than some specificity level threshold, T_s , (*e.g.*, we examined with T_s =14) then gene *g* is multifunctional so long as the two function do not have genes in common more than expected by chance. In [29], Wang et. al confirm that only ontology structure of the GO can be an indicator of the semantic similarity of gene functions [29]. With that, we would like to attempt a different kind of evaluation using gene pathways. We want to examine our method in a different way. We use ~~KEGG~~ *Kegg* database for gene pathways [34] for this analysis of the semantic distance among gene functional annotations and gene multifunctionality as follows:

(1) Genes that belongs to more pathways are more highly likely to be multifunctional compared with genes that belong to only one pathway in ~~KEGG~~*Kegg*. We extracted two sets of genes: The first set consists of genes participating (each) in only one *Kegg* pathway, and the second set contains genes that each gene participates in at least three pathways. We analyzed the shortest path length among all *bp* annotations of each gene in both sets and the results are shown in Figure 7.

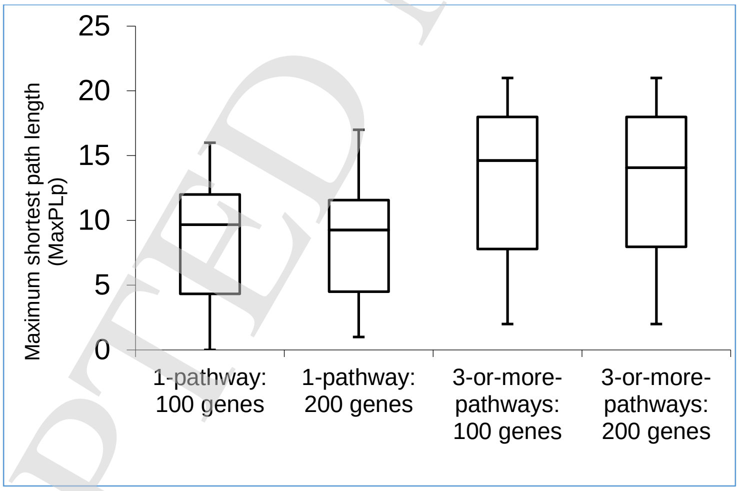

Fig. 7: The results of maximum shortest path length among the *bp annotations* (MaxPLp()) of four sets of genes. Each test is done with 100 or 200 genes have either only one pathway or at least three pathways.

(2) Secondly, we extracted pairs of genes that participate in highly diverse pathways in Kegg. Each gene pair *p* consists of two genes g_1 and g_2 (i.e., $p=(g_1, g_2)$) such that g_1 and g_2 belongs to two highly different pathways. Thus we consider the pair *p* representing a true

multifunctionality. We extract all *bp* annotations of both genes g_1 and g_2 of each pair and analyze their maximum path length and compare this to normally annotated genes; the results are shown in Figure 8. Finally, these evaluations using gene pathways from *Kegg* also are significant (p<0.01) for all sets of 100 and 200 genes with one pathways and with three pathways shown in both Figures 7 and 8.

In conclusion, this paper presents a well-defined study of multifunctional genes in the entire human genome using gene functional annotations and the GO. The work in this paper relies solely on the ontology structure of the GO irrespective of number of annotations per gene. This direction is ~~adding up to~~to increase our ~~cumulative~~ knowledge and understanding of gene and protein functions in ~~the~~ human cells. The proposed method was verified and evaluated against several criteria used commonly for this task as predictors for gene multifunctionality including gene pathways.

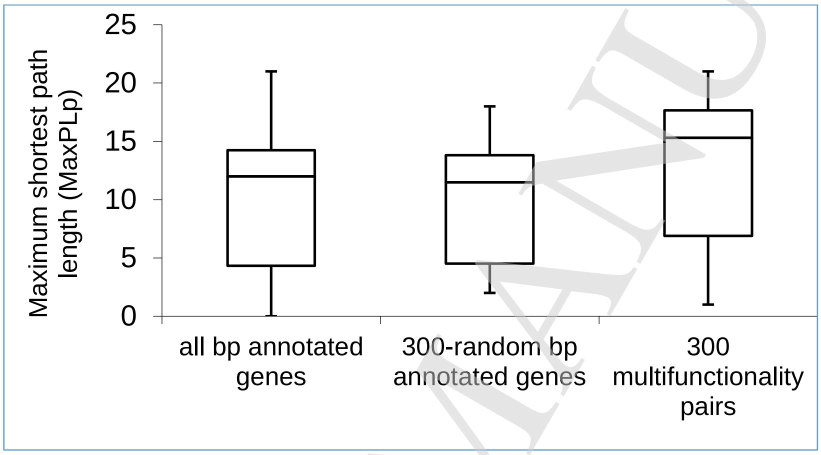

Fig. 8: Distribution of the maximum shortest path length (MaxPLp()) for the three sets of genes including the set of all human bp annotated genes.

References

1. Pritykin,Y. et al. (2015) Genome-Wide Detection and Analysis of Multifunctional Genes. PLOS Computational Biology, 11, e1004467.
2. Peppel,J.v-de and Holstege,F. CP. (2005) Multifunctional genes. Molecular Systems Biology, doi:10.1038/msb4100006
3. Gillis,J. and Pavlidis,P. (2011) The impact of multifunctional genes on ‘guilt by association’ analysis. PloS One, 6, e17258.
4. Al-Mubaid,H. et al. (2016) Assessing Gene-Disease Relationship with Multifunctional Genes Using GO. Proc. of IEEE AICCSA.
5. Hernandez,S. et al. (2014) MultitaskProtDB: a database of multitasking proteins. Nucleic Acids Research 42, D517–D520.
6. Mani M., et al. (2015) Moonprot: a database for proteins that are known to moonlight. Nucleic Acids Research 43, D277–D282.
7. Day, A. et al. (2009) Disease Gene Characterization through Large-Scale Co-Expression Analysis. PLoS ONE, 4.
8. Khan, I., et al. (2014) Genome-scale identification and characterization of moonlighting proteins. Biology Direct 9, 30.
9. Ashburner et al. (2000) Gene Ontology: tool for the unification of biology Nat Genet 25, 25-9.

10. The Gene Ontology Consortium. (2015) Gene Ontology Consortium: going forward. Nucl Acids Res 43, D1049–D1056.

11. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (2016) Online Mendelian Inheritance in Man, OMIM. World Wide Web URL: http://omim.org/

12. Ballouz,S. et al. (2017). Using predictive specificity to determine when gene set analysis is biologically meaningful. Nucleic Acids Res, 45, e20.

13. Khan,I.K. and D. Kihara, D. (2016). Genome-scale prediction of moonlighting proteins using diverse protein association information. Bioinformatics, 32, 2281-8..

14. Kim,J. et al. (2017) An analysis of disease-gene relationship from Medline abstracts by DigSee. Scientific Reports, 7, 40154.

15. Salathe,M. et al. (2006) The effect of multifunctionality on the rate of evolution in yeast. Molecular Biology and Evolution 23, 721–722.

16. Clark,W.T. and Radivojac,P. (2011) Analysis of protein function and its prediction from amino acid sequence. Proteins: Structure, Function, and Bioinformatics 79, 2086–2096.

17. Nagar,A. and Al-Mubaid,H. (2015) A Hybrid Semantic Similarity Measure for Gene Ontology Based On Offspring and Path Length. Proc. of IEEE CIBCB-2015 IEEE Conf. on Comp. Intelligence in Bioinformatics and Comp. Biology.

18. Glass,K. and Girvan,M. (2015) Finding New Order in Biological Functions from the Network Structure of Gene Annotations. PLoS Comput Biol., 11.

19. Singh-Blom,U.M.et al. (2013) Prediction and Validation of Gene-Disease Associations Using Methods Inspired by Social Network Analyses. PLOS ONE, 8 (9).

20. Schriml,L.M. and Mitraka,E. (2015) The Disease Ontology: fostering interoperability between biological and clinical human disease-related data. Mamm Genome, 26, 584–589.

21. Zhou and J. Skolnick. (2016) A knowledge-based approach for predicting gene–disease associations. Bioinformatics; 32, 2831–2838.

22. Schaefer M.H.. et al. (2012) HIPPIE: Integrating protein interaction networks with experiment based quality scores. PLoS One, 7.

23. Hippie PPI database: http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/

24. Becker E. et al. (2012) Multifunctional proteins revealed by overlapping clustering in protein interaction network. Bioinformatics 28, 84–90.

25. Hernandez,S. et al. (2011) Do moonlighting proteins belong to the intrinsically disordered protein class? Proteomics Bioinformatics, 5, 262–264.

26. The National Center for Biotechnology Information, NIH, USA: https://www.ncbi.nlm.nih.gov

27. NCBI. Clearing Up Confusion with Human Gene Symbols & Names Using NCBI Gene Data. NIH, USA. (2016) https://ncbiinsights.ncbi.nlm.nih.gov/2016/11/07/clearing-up-confusion-with-human-gene-symbols-names-using-ncbi-gene-data/.

28. QuickGO; http://www.ebi.ac.uk/QuickGO

29. Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007; 23(10):1274–1281.

30. H. Al-Mubaid et. al. Determining Multifunctional Genes and Diseases in Human Using Gene Ontology. Proceedings of 9th Int'l Conf on Bioinformatics and Computational Biology BICOB-2017, Honolulu, HI, USA, March 2017.

Appendix:

Table 0. Avg $MaxPL_f$ with clusters of 1000 genes in each cluster.

After sorting all genes based on $\text{MaxPL}_f()$ (descending order)	**mean $\text{MaxPL}_f()$**
Top 1000	13.987
1001 – 2000	12.189
2001 – 3000	11.250
3001 – 4000	10.427
4001 – 5000	9.576
5001 – 6000	8.487
6001 – 7000	7.505
7001 – 8000	6.336
8001 – 9000	2.030
9001 – 10000	3.189
10001 – 11000	0.880
11001 – 12000	0.498
Lowest 1191	0
Total number of genes with mf annotations: 13,191. Number of genes with only 1 mf term: 2630 (i.e., $\text{MaxPL}_f = 0$)	