SCIENTIFIC REPORTS

OPEN

Gene Ontology Enrichment Improves Performances of Functional Similarity of Genes

Received: 23 November 2017
Accepted: 25 July 2018
Published online: 14 August 2018

Wenting Liu[1], Jianjun Liu[1] & Jagath C. Rajapakse[2]

There exists a plethora of measures to evaluate functional similarity (*FS*) between genes, which is a widely used in many bioinformatics applications including detecting molecular pathways, identifying co-expressed genes, predicting protein-protein interactions, and prioritization of disease genes. Measures of *FS* between genes are mostly derived from Information Contents (IC) of Gene Ontology (GO) terms annotating the genes. However, existing measures evaluating IC of terms based either on the representations of terms in the annotating corpus or on the knowledge embedded in the GO hierarchy do not consider the enrichment of GO terms by the querying pair of genes. The enrichment of a GO term by a pair of gene is dependent on whether the term is annotated by one gene (i.e., partial annotation) or by both genes (i.e. complete annotation) in the pair. In this paper, we propose a method that incorporate enrichment of GO terms by a gene pair in computing their *FS* and show that GO enrichment improves the performances of 46 existing *FS* measures in the prediction of sequence homologies, gene expression correlations, protein-protein interactions, and disease associated genes.

Gene ontology (GO) provides a controlled vocabulary of gene functions and molecular attributes that are arranged in a directed acyclic graph (DAG) representing semantic relations among GO terms. GO is often used to interpret results and make inferences of biological experiments. Functional similarity (*FS*) between two genes is inferred from GO terms annotating the genes and is widely adopted in detecting and interpreting genetic interactions, functional interactions, protein-protein interactions[1], biological pathways[2,3], prioritization of disease genes[4], and disease similarities[5]. Most *FS* measures in the literature are computed using information contents (IC) of GO terms annotating the querying pair of genes. The IC of a GO term is evaluated using either the representation of GO terms in the annotation corpus associated with a species (corpus-based methods) or the structure of the DAG (structure-based methods).

Functional similarity between two genes is given by the common information or semantic similarity of the GO terms annotating the two genes. Semantic similarity (SS) among GO terms is generally evaluated by the ICs of common ancestor terms as ancestors subsume semantic concepts of descendants due to the hierarchical nature of the DAG. Corpus-based SS measures such as Resnik[6], Lin[7], Nunivers[8], and Schlicker[9] are based on the ICs of the most informative common ancestor, and XGraSM[10] and TopoICSim[11] have extended Lin and Nunivers measures to include IC of all the common ancestors. Structure-based methods such SORA[12] and WIS[13] determine the ICs of a GO term based on the number of descendants and/or the depth of the term in the DAG. Semantic similarity of a GO term set is derived from the ICs of (i) individual terms, (ii) pairs of terms, or (iii) the term set. The ICs of a term set is derived combining individual or pair of terms by using statistical averaging measures or Tversky's ratios[14].

Functional similarity of two genes are measured by ICs of common GO terms annotating the genes, which are evaluated purely based on the annotations of the corpus or the expert knowledge embedded in the DAG. Figure 1 illustrates how two genes are represented in a DAG by their GO terms and as seen, some GO terms in the DAG annotate only one gene while others annotate both genes or none. However, existing IC measures ignore the local context or how GO terms are represented in the querying gene pair. For example, when measuring *FS* of a gene pair, GO terms annotating both genes are more likely to be enriched than those annotating only one gene. To overcome this drawback of existing *FS* measures, we propose to incorporate GO-enrichment by querying gene

[1]Human Genetics, Genome Institute of Singapore, Singapore, Singapore. [2]School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore. Correspondence and requests for materials should be addressed to W.L. (email: liuwenting@ucla.edu) or J.L. (email: liuj3@gis.a-star.edu.sg) or J.C.R. (email: asjagath@ntu.edu.sg)

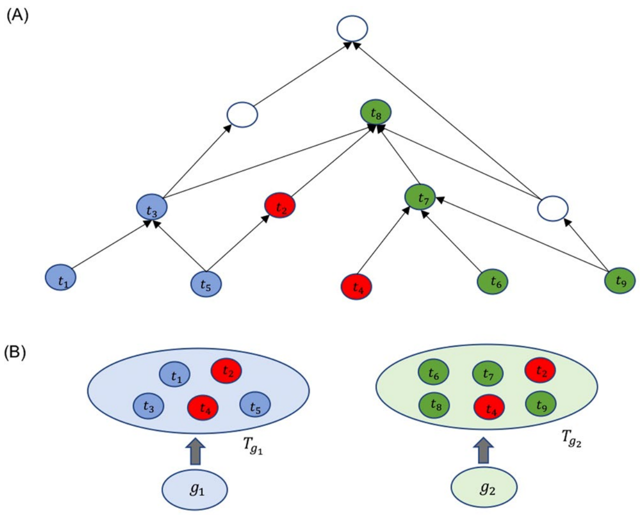

Figure 1. Illustration of a DAG representing GO terms annotating two genes. (**A**) The DAG representing GO terms that annotates the two genes, and (**B**) the two genes g_1 and g_2 and their GO term sets T_{g_1} and T_{g_2}. The *FS* is derived as the semantic similarity or the common information contents (IC) in the two term sets. Our approach takes care of differential enrichment of GO terms by the querying gene pair: for example, blue terms annotates only g_1, green terms annotates only gene g_2, and red terms annotates both genes g_1 and g_2.

Data Type	Data Sets	#Protein pairs	ontology	#Experiments	*p*-value
Disease Genes	DG_BP; DG_MF; DG_CC	6084	BP; MF; CC	138	5.619e-08
Yeast PPI	PPI_BP; PPI_MF; PPI_CC	8654; 7166; 8852	BP; MF; CC	138	2.885e-07
Human PPI	PPI_BP; PPI_MF; PPI_CC	2408; 2576; 2108	BP; MF; CC	138	3.528e-03
Yeast GE	GE_BP; GE_MF; GE_CC	4800	BP; MF; CC	138	6.912e-09
CESSM	ECC; Pfam; SeqSim	13430	BP; MF	276	4.94e-16
Total	DG; PPI; GE; ECC; Pfam; SeqSim	35376; 34056; 35274	BP; MF; CC	828	<2.2e-16

Table 1. Details of five datasets and statistical significances of the improvement of performances by FS^* over corresponding *FS* measures on predicting disease genes, protein interactions on yeast PPI dataset and yeast GE dataset, gene co-expressions on yeast GE dataset in different ontological domains (BP, MF, and CC), and sequence similarities on CESSM datasets (ECC, Pfam, and SeqSim).

pair in the computation of IC of a GO term. Specifically, in the context of two genes, the probability of a GO term is defined as the joint probability of the term as inferred by background corpus and as annotated by two querying genes. We investigate the effect of introducing GO enrichment on 46 existing *FS* measures and demonstrate that enriched *FS* (*FS**) measures outperform the prediction of sequence homologies, gene expression correlations, protein-protein interactions, and disease-associated genes in majority of the cases.

Results

Performance on *FS** on all datasets.

We assessed performances of *FS** measures on benchmark datasets for predicting sequence similarities, gene expression (GE) correlations, protein-protein interactions (PPI), disease genes (DG) and compared with those of corresponding *FS* measures. Table 1 shows one-sided *p*-values of the improvement of performances of all the experiments on five benchmark datasets, by using Wilcoxon signed rank tests[15]. As seen, *FS** measures showed significant improvement over *FS* measures in the prediction of disease genes on 138 experiments, protein interactions on 138 experiments of yeast PPI data and 138 experiments of human PPI data, gene co-expressions on 138 experiments of yeast GE data, and sequence similarities on 276 experiments on CESSM dataset; and on all 828 experiments. Irrespective of the ontology (BP, MF, or CC) and the type of *FS* measure, incorporation of GO enrichment significantly improved the prediction of sequence similarities, gene co-expression patterns, protein-protein interactions, and disease associated genes.

Performance of *FS** measures on individual datasets.

Table 2 lists 46 FS measures that are computed based on ICs of individual terms, pairs of terms, and the set of terms annotating the genes. Lin[7] (L), Nunivers[8] (N), Schlicker[9] (C), and extended (XGraSM[10]) Lin (XL) and Nunivers (XN), Zhang[16] (Z) GO-universal[8] (U) and Wang[17] (W) are corpus-based methods using IC son individual terms or pairs of terms (i.e., SS). Functional similarities of the gene pair are computed by combining ICs of annotating GO terms by using GIC[1] (Jaccard index), DIC[8] (dice index), and UIC[8] (universal index) on individual terms, and Average (AVG), Maximum (MAX), Best-Match Average (BMA) and Average Best-Matches (ABM) on pairs of terms; SORA[12] (R) and WIS[13] (I) are

Acronyms	IC/SS	*FS* measures
U	GO-universal[8]	U_{ABM}, U_{BMA}, U_{MAX}, U_{AVG}, U_{DIC}, U_{GIC}, U_{UIC}
Z	Zhang[16]	Z_{ABM}, Z_{BMA}, Z_{MAX}, Z_{AVG}, Z_{DIC}, Z_{GIC}, Z_{UIC}
W	Wang[17]	W_{ABM}, W_{BMA}, W_{MAX}, W_{AVG}, W_{DIC}, W_{GIC}, W_{UIC}
N	Nunivers[8]	N_{ABM}, N_{BMA}, N_{MAX}, N_{AVG}
XN	Extended Nunivers[10]	XN_{ABM}, XN_{BMA}, XN_{MAX}, XN_{AVG}
L	Lin[7]	L_{ABM}, L_{BMA}, L_{MAX}, L_{AVG}
XL	Extended Lin[10]	XL_{ABM}, XL_{BMA}, XL_{MAX}, XL_{AVG}
S	Schlicker[9]	S_{ABM}, S_{BMA}, S_{MAX}, S_{AVG}
D	Direct-term based[18]	D_{DIC}, D_{GIC}, D_{UIC}
R	SORA[12]	R_{OR}
I	WIS[13]	I_{IUR}

Table 2. Details of 46 FS measures. The types of IC/SS and methods used to compute *FS* measures: GIC[1] (Jaccard index), DIC[8] (dice index), and UIC[8] (universal index) for individual terns; Average (AVG), Maximum (MAX), Best-Match Average (BMA) and Average Best-Matches (ABM) for measures based on pairs of terms; and Overlap Ratio (OR) and Intersection to Union Ratio (IUR) for measures based on sets of terms.

Datasets	Methods	Correlation	Datasets	Methods	Correlation	Datasets	Methods	Correlation
ECC_BP	$\mathbf{XN_{BMA}}^*$	**0.4748**	Pfam_BP	$\mathbf{W_{ABM}}^*$	**0.5261**	SeqSim_BP	$\mathbf{I_{IUR}}^*$	**0.8028**
	$\mathbf{XN_{BMA}}$	**0.4748**		W_{BMA}^*	0.5223		I_{IUR}	0.7927
	XL_{BMA}^*	0.4748		R_{OR}^*	0.5199		R_{OR}	0.7884
	XL_{BMA}	0.4708		I_{IUR}^*	0.5005		W_{ABM}^*	0.7741
	N_{BMA}^*	0.4651		R_{OR}	0.4933		R_{OR}^*	0.7738
	$\mathbf{R_{OR}}^*$	**0.7828**		$\mathbf{I_{IUR}}^*$	**0.6961**		$\mathbf{I_{IUR}}^*$	**0.7217**
	W_{BMA}^*	0.7665		R_{OR}^*	0.6829		I_{IUR}	0.7165
ECC_MF	XN_{BMA}^*	0.7567	Pfam_MF	I_{IUR}	0.6627	SeqSim_MF	R_{OR}	0.6505
	XN_{BMA}	0.7525		R_{OR}	0.6565		D_{GIC}^*	0.6358
	N_{BMA}^*	0.7525		W_{ABM}^*	0.6283		D_{GIC}	0.6285

Table 3. Top five performers of *FS* and *FS** measures on predicting ECC, Pfam, and SeqSim similarities of protein pairs of CESSM datasets, using BP and MF ontologies.

structure-based methods that define ICs for individual terms from information from DAN and compute ICs on the whole term set by using Overlap Ratio (OR) and Intersection to Union Ratio (IUR).

Tables 3, 4, 5 and 6 lists the top five performers of *FS*/*FS** measures on each dataset. As seen from the tables, the best performers (ranked by AUC values of prediction) are mostly *FS** measures: (i) I_{IUR}* on predicting sequence similarity on both BP and MF ontology, Pfam similarity on MF ontology, and PPIs of yeast on CC ontology; (ii) R_{OR}* on predicting ECC similarity on MF ontology, gene co-expressions of yeast on both BP and MF ontology, and disease genes on both BP and MF ontology; (iii) W_{ABM}* on predicting Pfam similarity on BP ontology and PPIs of human on CC ontology; Z_{BMA}* on predicting PPIs of human on BP ontology; and (iv) XN_{BMA}* on predicting ECC similarity on BP ontology, PPIs of yeast on BP ontology, and disease genes on BP ontology. For predicting PPIs of yeast on MF ontology, both D_{UIC} and D_{UIC}* performed best with quite similar AUC score of 0.6930 and 0.6928, respectively, followed by D_{DIC}* and D_{GIC}* with AUC score both of 0.6926. S_{ABM} performed best on predicting human PPIs from MF ontology.

We notice that the best performers are mostly *FS** measures derived from structure-based IC measures (I_{IUR}* and R_{OR}*), corpus-based IC measures that considered ancestors of the terms (W_{ABM}* and Z_{BMA}*) or extended corpus-based measures (XN_{BMA}*). This indicates that using only the information of annotating corpus is insufficient and both the structure of DAG and GO enrichment by the querying gene pair are essential for determining *FS* between genes.

Supplementary Tables 1–6 give the details of performances of *FS** over all 46 FS measures in predicting sequence similarities, gene co-expressions, protein-protein interactions and disease genes. The tables list correlations or AUC scores of the measures, percentages improvement of *FS** over *FS*, and statistical significances of improvement on every dataset. The significant improvements at FDR < 0.01 are indicated in bold and the significant drops in performance are marked in red; for each dataset, top *FS* and top *FS** performers are indicated in bold.

Our results show that GO enrichment improves on almost all 46 *FS* measures on all the datasets, except inferring human PPIs on MF ontology. The corpus-based measures (Lin[7], Nunivers[8], GO-universal[8], Wang[17], Zhang[16]) and graph-based extensions of corpus-based measures (XGraSM[10] of Lin[7] and Nunivers[8]) were improved significantly with GO enrichment on most datasets. In general, BMA and ABM methods provided best performances and performed equally well on most semantic similarity measures. Adaptation of efficient correction factors improved the performance on some measures: Schlicker[9] uses the IC value of MICA and does not significantly

Datasets	Methods	AUC	Datasets	Methods	AUC	Datasets	Methods	AUC
human PPI_BP	**$\mathbf{Z_{BMA}}^{*}$**	**0.8750**	human PPI_MF	S_{ABM}	0.7787	human PPI_CC	**$\mathbf{W_{ABM}}^{*}$**	**0.7775**
	Z_{BMA}	0.8747		L_{ABM}	0.7777		W_{BMA}^{*}	0.7697
	N_{BMA}^{*}	0.8739		S_{ABM}^{*}	0.7771		I_{IUR}^{*}	0.7678
	N_{BMA}	0.8737		L_{ABM}^{*}	0.7762		U_{ABM}^{*}	0.7658
	S_{BMA}^{*}	0.8721		N_{ABM}	0.7718		U_{ABM}	0.7657
yeast PPI_BP	**$\mathbf{XN_{BMA}}^{*}$**	**0.8565**	yeast PPI_MF	D_{UIC}	0.6930	yeast PPI_CC	**$\mathbf{I_{IUR}}^{*}$**	**0.8248**
	XN_{MAX}^{*}	0.8563		D_{UIC}^{*}	0.6928		I_{IUR}	0.8158
	XL_{MAX}^{*}	0.8561		D_{DIC}^{*}	0.6926		R_{OR}^{*}	0.8143
	XN_{MAX}	0.8559		D_{GIC}^{*}	0.6926		U_{ABM}^{*}	0.8072
	XN_{BMA}	0.8559		D_{GIC}	0.6916		N_{ABM}^{*}	0.8068

Table 4. Top five performers of *FS* and *FS** measures predicting protein-protein interactions of human and yeast PPI datasets, using three ontologies: BP, MF, and CC.

Datasets	Methods	Correlation	Datasets	Methods	Correlation	Datasets	Methods	Correlation
yeast GE_BP	**$\mathbf{R_{OR}}^{*}$**	**0.2927**	yeast GE_MF	**$\mathbf{R_{OR}}^{*}$**	**0.2138**	yeast GE_CC	**$\mathbf{Z_{DIC}}^{*}$**	**0.4263**
	$\mathbf{D_{GIC}}^{*}$	**0.2877**		R_{OR}	0.2087		Z_{DIC}	0.4253
	R_{OR}	0.2876		D_{GIC}^{*}	0.2023		Z_{GIC}^{*}	0.4236
	Z_{GIC}^{*}	0.2875		D_{GIC}	0.2022		Z_{GIC}	0.4233
	D_{GIC}	0.2873		D_{DIC}^{*}	0.2008		Z_{UIC}	0.4229

Table 5. Top five performers of *FS* and *FS** measures predicting gene co-expressions on yeast GE dataset, using three ontologies: BP, MF, and CC.

Datasets	Methods	AUC	Datasets	Methods	AUC	Datasets	Methods	AUC
Disease Genes_BP	**$\mathbf{XN_{BMA}}^{*}$**	**0.8065**	Disease Genes_MF	**$\mathbf{R_{OR}}^{*}$**	**0.7541**	Disease Genes_CC	**$\mathbf{R_{OR}}^{*}$**	**0.7064**
	R_{OR}	0.8062		U_{BMA}^{*}	0.7357		U_{BMA}	0.7032
	XN_{BMA}	0.8058		U_{BMA}	0.7357		R_{OR}	0.7031
	I_{IUR}	0.8030		N_{BMA}^{*}	0.7344		U_{BMA}^{*}	0.7029
	N_{BMA}^{*}	0.8019		N_{BMA}	0.7330		W_{BMA}	0.7006

Table 6. Top five performers of *FS* and *FS** measures predicting disease genes on benchmark dataset, using three ontologies: BP, MF, and CC.

improve the performance of the Lin[7] approach; XGraSM[10] uses all common informative ancestors to correct Lin[7] and Nunivers[8] approaches in order to improve their performances. Thus, including common informative ancestors in the conception of SS improves performance, especially for approaches including only the features of child terms in the computation of IC such as Zhang[16], Wang[17], SORA[12] and WIS[13] measures.

As *FS** measures differently treats GO terms uniquely annotating (i.e., annotating one gene) and GO terms commonly annotating (i.e., annotating both genes) the querying genes, measures including both types of terms are significantly improved with GO enrichment: for example, Lin[7], Nunivers[8], and Direct-term based[18] measures consider both common terms and individual terms; GO-universal[8] measure considers all children terms (common or individual terms); and Zhang[16], Wang[17], SORA[12] and WIS[13] measure consider all ancestors (common terms) and children terms (common or individual terms). Especially, Wang[17] measures (W_{ABM}, W_{BMA}) improved significantly on capturing sequence homology with a correlation improvement of 8% of ECC, 25% of Pfam, 34% of SeqSim on MF ontology; and 13% of Pfam, 16% of SeqSim on BP ontology while Wang[17] measures (W_{ABM}, W_{BMA}, W_{MAX}) improved significantly for GE correlations on BP and MF with a correlation improvement of 3.5% on BP, and 16% on MF. Wang[17] measure (W_{AVG}) also improved most significantly for inferring human PPIs on CC with 3% AUC improvement, yeast PPIs on BP and CC with AUC improvement 4% and 3%, respectively. SORA[12] approach (R_{OR}) improved most significantly for predicting yeast PPIs on MF ontology with AUC improvement 2%, disease genes on MF ontology with AUC improvement 4%. WIS[13] approach (I_{IUR}) improved most significantly for GE correlations on BP, MF, and CC with correlation improvement of 5%, 6%, and 5%, respectively. GO-universal approach (U_{AVG}) improved most significantly (labelled as green) for GE correlations on BP and CC with correlation improvement of 3% and 10%, respectively; and inferred human PPIs on BP, yeast PPIs on CC with AUC improvement both 1%.

Direct term-based D_{GIC}[18] and D_{DIC}[18] are improved most significantly for inferring human PPIs on BP and MF, yeast PPIs on BP and CC, GE correlations on BP and CC, disease genes on CC. Out of all 46 FS measures, the performance of measure related to UIC measure didn't improve with GO enriched *FS** measures. This is because the UIC measure does not discriminate common terms and unique terms while the enrichment is manifested by the differences between common and unique terms. UIC is defined as the sum of IC of the terms annotated by both genes, divided by the maximum of the sum of ICs annotated individual genes and GIC is defined as the sum of ICs of GO terms annotated by both genes, divided by the sum of ICs of terms annotated by individual genes. The enriched IC term leads to increase the ICs of the terms that are annotated by both genes more than those annotated by one gene. When there are only a few terms annotated by both genes, GIC*/UIC* do not perform as good as UIC. Therefore, the *FS* with GIC*/UIC* lead to quality loss when predicting sequence similarity as seen in Supplementary Table 2. The loss in predicting with GIC* is smaller than the loss in predicting with UIC*.

Supplementary Figs 1–3 show ROC curves of overall best performers (I_{IUR}, R_{OR}, XN_{BMA}, Z_{BMA} and D_{GIC}) of *FS* to *FS**, predicting human and yeast PPIs, disease genes on three ontologies, respectively. As seen, most of top performers are *FS**, underscoring that GO enrichment indeed improves performance of best *FS* performers on all the datasets.

Conclusions

Many *FS* measures have been proposed using GO annotation to quantify similarities between genes for exploitation and validation of biological knowledge embedded in omics data. These measures were derived based on the topological structure of GO semantics and/or the GO annotations of the genes/proteins annotating (background) corpus. However, the representativeness of GO terms in the two querying genes has not been considered in evaluating *FS* measures. In other words, differential enrichment of the terms annotating one gene and the terms annotating two genes in the querying pair has been ignored in the existing *FS* measures. We proposed an enriched *FS* measure, *FS**, that can be used to incorporates the enrichment of GO terms by querying genes in the existing *FS* measures and demonstrated improved performances of *FS** measures in the prediction of sequence similarities, protein-protein interactions, gene co-expressions, and disease associated genes.

We tested GO enrichment on 46 *FS* measures on five benchmark datasets including sequence similarities of the CESSM dataset, yeast GE data, human and yeast PPI data, and disease genes, and presented comparison of performances of *FS* and *FS** measures. Results indicate that *FS** outperforms *FS* measures in a vast majority of the experiments. We conclude that consideration of GO enrichment by the querying genes is an essential step for computation of *FS* between genes. As seen, *FS* measures including both commonly and individually annotating terms, the performances of *FS** between genes improved much significantly over *FS* measures. We also noticed that *FS** significantly outperformed on datasets containing a lot of uniquely annotated genes (i.e., those annotated by the terms in the low levels of GO hierarchy).

Enriched *FS** of structure-based measures (I_{IUR} and R_{OR}), corpus-based methods using DAG structure (W_{ABM} and Z_{BMA}), or graph-based extensions of corpus-based measures (XN_{BMA}) achieved best performances on all the benchmark datasets except predicting human and yeast PPIs on MF ontology. This indicate that introducing the annotations of the corpus and the querying pair of genes in the structure-based IC measures gives accurate *FS* measures, underscoring the need for incorporating the representativeness of querying genes.

Enriched *FS** is easily adapted to and generally improves the performance of any *FS* measure that uses ICs of GO terms. On the other hand, the accuracy of GO annotation naturally limits the performance of existing *FS* measures as they do not consider both the local context of two genes and the background distribution of terms in the annotating corpus. Our experiments suggest that the local context of querying genes is sensitive to the missing and spurious terms in the GO annotating corpus. One could extend our method to evaluate the functional coherence of gene sets, which will have applications in the detection of functional modules or pathways. *FS** measures more accurately identify functionally similar genes than *FS* measures and will provide more reliable computational evidences for finding new pathways and disease genes. We conclude that the GO enrichment is an essential step when assessing *FS* of two genes and a set of genes.

Methods

Data Sets.

Molecules with sequence similarities show similar functions or MF ontology, and with similar gene expressions are likely to belong to same pathway or so have similar BP ontology. Interacting proteins are located in the same cellular location and so have the similar CC ontology. Therefore, we evaluated the performances of *FS** measures by their correlations with sequence similarities, gene co-expressions, protein-protein interactions, and disease association of genes on benchmark datasets.

Correlation with sequence similarity.

Various studies have shown that molecules with sequence similarities have similar ontological annotations[19], so we used sequence similarities to demonstrate the goodness of *FS* measures[20,21]. For BP and MF, we downloaded sequence similarities of selected human proteins with known relationships from CESSM[22] online tool (http://xldb.di.fc.ul.pt/tools/cessm/) and compared the performances of different *FS* measures predicting sequence similarities. The CESSM website provides a list of protein pairs and similarities between pairs of proteins, using three distinct evaluations: sequence similarity (SeqSim), Pfam domain similarity, and enzyme commission class (ECC) similarity. The goodness of prediction was evaluated by the correlations between protein similarities captured by SeqSim, Pfam similarity, and ECC similarity and the *FS* measures.

Correlation with gene co-expressions.

Genes involved in the same biological process, sharing similar functions or cellular components, tend to exhibit similar expression patterns, so a good correlation ought to exist between co-expressed genes and their *FS*. We used the *S.cerevisiae* gene-expression dataset used in earlier studies[20,21], which contains co-expression values of 4800 pairs of genes for each ontology, downloaded from GeneMANIA[23] and other microarray experiments. We computed Pearson's correlations between gene co-expressions and *FS* values of BP, MF and CC ontologies.

Correlations with AUC on predicting protein-protein interactions. Two interacting proteins have same CC ontology, share similar functions, and are likely to belong to same BP, so the *FS* between two proteins is an indicative of an interaction[24,25]. The prediction of protein-protein interaction (PPI) was formulated as a classification problem where *FS* exceeding a certain threshold indicated an interaction between two proteins. We used yeast PPI datasets from the Jain and Davis's database[21,26], which contain 4385 PPIs on BP, 3858 PPIs on MF, and 4469 PPIs on CC. The human PPIs was downloaded from the Database of Interacting Proteins (DIP)[27] that contains 1435 PPIs on BP, 1441 PPIs on MF, and 1431 PPIs on CC. The same numbers of negative interactions generated by randomly choosing annotated protein pairs in BP, MF, and CC ontology. The area under the curve (AUC) values of receiver operating characteristic (ROC) curves of the predictor were used to evaluate performances. A ROC curve plots true positive rates (sensitivity) against false positive rates (1-specificity) of prediction at different thresholds.

Correlations with AUC on predicting disease genes. Recent studies[28,29] have combined gene-gene similarities, disease gene associations, and disease–disease similarities in order to predict disease associated genes. For example, Zeng *et al.*[28] used a path-based similarity measure HeteSim[30] to calculate the similarities between nodes in heterogeneous networks constructed using protein-protein interactions, gene-phenotype associations, and phenotype-phenotype similarity, to prioritize candidate disease genes; and Zou *et al.*[29] constructed a microRNA-disease network from microRNA–microRNA and disease–disease networks. Inspired by these works, we formulate the prediction of disease associated genes as a *FS* between a known set of disease genes[31] and the candidate gene. The *FS* between disease-associated genes and the candidate gene were computed using 46 *FS* measures and the performances were evaluated using AUC values of the prediction of the candidate gene as a disease associated gene. The dataset for disease associated gene prediction was constructed from a preliminary set of 78 OMIM (http://www.omim.org) disease phenotypes collected by Schlicker *et al.*[31]. For each of the phenotypes, one disease protein was randomly selected and predicted as a disease candidate by using functional profiles of other genes. Thereby, a disease genes benchmark set consisting of 78 phenotypes and 78 randomly selected known disease proteins as candidates was constructed.

Significance test for correlation improvement. To determine statistical significance of an improvement of correlations between *FS* and *FS** measures and sequence similarities, gene co-expressions, and AUCs of prediction of PPI and disease associated genes, we adopted Williams test[32] for correlations between two metrics[33,34]. Specifically, to test whether the population correlation between X_1 and X_3 equals the population correlation between X_2 and X_3, we computed the following *t*-test:

$$t(n-3)=\frac{(r_{13}-r_{12})\sqrt{(n-1)(1+r_{12})}}{\sqrt{\frac{2K(n-1)}{(n-3)}+\frac{(r_{23}+r_{13})^{2}}{4}(1-r_{12})^{3}}} \tag{1}$$

where $K=1-r_{12}{}^{2}-r_{13}{}^{2}-r_{23}{}^{2}+2r_{12}r_{13}r_{23}$.

The higher the correlation between the metric scores, the greater the statistical power of this test than the Fisher *r* to *z*-transformation test on independent correlations is. As *FS* and *FS** are highly correlated, we used this Williams test[32] and adopted FDR for multiple test correction.

To determine whether correlations or AUC values are significantly improved for all *FS* measures to *FS** measures on each dataset (CESSM, yeast GE, yeast PPI, and the combination of the three datasets), we implemented the Wilcoxon signed rank test with continuity correction[35], which tests repeated measurements on a single sample to assess whether their population mean ranks differ. This test is suggested as an alternative for *t*-test for dependent samples when the population cannot be assumed to be normally distributed. We used one-sided Wilcoxon signed rank test to show whether *FS** significantly improves the performance of *FS* irrespective of the *FS* measure and the type of ontology.

Funsim measures. *Information content of a gene ontology term.* Gene ontology (GO) is an ontology of terms describing how gene products behave in a cellular context in a species-independent manner and comprises of three ontological domains: biological process (BP), molecular function (MF), and cellular component (CC)[36]: BP is a collection of molecular events, MF defines gene functions in biological processes, and CC describes gene localizations inside a cell. There are three semantic relations between two GO terms: *is-a* is used when one GO term is a subtype of another GO term, *part-of* is used to represent part-whole relationship in the GO terms, and *regulate* is used when the occurrence of one biological process directly affects the manifestation of another process or quality[37]. GO terms and their semantic relations form a hierarchical directed acyclic graph (DAG) where three domains, BP, MF and CC, are represented as the root terms. The ancestor terms in the hierarchy subsume the semantics of descendant terms.

A gene is associated (or annotated) with GO terms describing the properties of its products (i.e., proteins) and with a corpus consisting of GO annotations (GOA) of all genes in an organism. GOA data of species can be readily downloaded from the GO annotation database (http://www.geneontology.org/GO.downloads.annotations.shtml). Functional similarity of a gene pair or a set is determined by the semantic similarities of the GO terms annotating the gene pair or set. Semantic similarity defines a distance between terms in the semantic space of GO and is quantified by the information contents (IC) of the terms. The information content (IC) of a GO term *t* is defined by negative log-likelihood:

$$IC(t)=-\log p(t) \tag{2}$$

where term probability $_{p}(t)$ of term t is determined from the annotations of the corpus (corpus-based) or from the structure of the DAG (structure-based). The intuition is that terms in lower levels of DAG, that is, the terms with lower probability carry more specific information than the terms at higher levels in the hierarchy. Corpus-based methods evaluate the term probability as

$$p(t) = \frac{M}{N} \tag{3}$$

where M is the number of genes annotated by term t and N is the total number of genes in the annotating corpus.

Structure-based methods evaluate term probability based on the location and the number of children and ancestors of the term. For example, SORA[12] method defines a term IC as

$$IC(t) = depth(t)\ *\ \left(1 - \frac{\log(|C(t)| + 1)}{\log(T_{total})}\right) \tag{4}$$

where $depth\ (t)$ is the depth and $C\ (t)$ is the set of children of term t and T_{total} is the total number of terms in the DAG. The WIS[13] method extends the idea from SORA[12] and defines the IC of term t not only based on its children but also its depth, as

$$IC(t) = depth(t)\ *\ \log(|A(t)|)\ *\ \left(1 - \frac{\log\left(\sum_{x\in C(t)\cup\{t\}}\frac{1}{depth(x)} + 1\right)}{\log(T_{total})}\right) \tag{5}$$

where $A\ (t)$ the set of the ancestor (parent) terms of term t. The GO-universal method[8] combines information from both the corpus and the DAG and defines term probability as

$$p(t) = \begin{cases} 1, & if\ \ t\ \ is\ root \\ \prod_{x\in P(t)} \frac{p(x)}{|C(x)|} & otherwise \end{cases} \tag{6}$$

Semantic similarity measures between GO terms. A semantic similarity measure defines a semantic distance between a pair or a set of GO terms and is evaluated as the IC of the gene pair or set. Since the ancestor terms in the DAG subsume the concepts of descendent terms, semantic similarities are mostly computed based on the ICs of common ancestors of the terms. Let $A(T) = \{a_0,\ a_1,\ \cdots a_{n-1}\}$ denote the set of ancestor terms of the GO-terms in the set T where a_0 denotes the most informative common ancestor (i.e., the ancestor with the largest IC). The measures by Resnik[6], Lin[7], Jiang & Conrath[38], Nunivers[8], Schlicker[9], and XGraSM[10] (Extended Lin and Nunivers) define semantic similarities or $IC(\{t_1,\ t_2\})$ of a pair of terms, t_1 and t_2, based on the IC of their most informative common ancestor:

Resnik[6]:

$$IC(\{t_1,\ t_2\}) = IC(a_0) = \ \max\{IC(x)\colon x \in A(\{t_1,\ t_2\})\} \tag{7}$$

Lin[7]:

$$IC(\{t_1,\ t_2\}) = \frac{2 \times IC(a_0)}{IC(t_1) + IC(t_2)} \tag{8}$$

Nunivers[8]:

$$IC(\{t_1, t_2\})\ \ = \frac{IC(a_0)}{\max\ \{IC(t_1), IC(t_2)\}} \tag{9}$$

Zhang[16]:

$$IC(\{t_1,\ t_2\}) = \exp(-IC(a_0)) \tag{10}$$

Schlicker[9]:

$$IC(\{t_1,\ t_2\}) = \frac{2 \times IC(a_0)}{IC(t_1) + IC(t_2)}(1 - \exp(-IC(a_0))) \tag{11}$$

Extended Lin[10]:

$$IC(\{t_1,\ t_2\}) = \frac{2 \times IC(a_0)}{IC(t_1) + IC(t_2)}\frac{1}{n}\left(1 + \sum\nolimits_{j=1}^{n-1}\frac{IC(a_j)}{IC(a_0)}\right) \tag{12}$$

Extended Nunivers[10]:

$$IC(\{t_1,\ t_2\}) = \frac{IC(a_0)}{\max\{IC(t_1),\ IC(t_2)\}}\frac{1}{n}\left(1 + \sum\nolimits_{j=1}^{n-1}\frac{IC(a_j)}{IC(a_0)}\right) \tag{13}$$

All the above similarity measures assign unit weight to every edge in the DAG. Edge-based similarity measures such as Wang[17], SORA[12] and WIS[13] incorporate weights to the edges of the DAG. Wang[17] defined a semantic value s_t of term t by assigning semantic weight w of 0.8 and 0.6 for *is-a* and *part-of* associations, respectively, and

$$s_t(x) = \begin{cases} 1, & \mathit{if}\ x = t \\ \max\{w(x')\text{: } x' \in C(x)\}, & \mathit{otherwise} \end{cases} \tag{14}$$

And the IC of a term and semantic similarity between terms are computed from semantic values of the ancestors as respectively given by

$$IC(t) = \sum_{x\in A(\{t\})\bigcup\{t\}} s_t(x) \tag{15}$$

$$IC(\{t_1,\ t_2\}) = \frac{\sum_{t\in A(\{t_1,\ t_2\})} s_{t_1}(t) + s_{t_2}(t)}{IC(t_1) + IC(t_2)} \tag{16}$$

In methods of SORA[12] and of WIS[13], the IC of a term is considered to be composed of an inherited IC from the parent and an extended IC intrinsic to the term; and the extended IC is expressed as a weighted inherited IC from the parent. For a term t and its parent a, the IC of the pair is given by

$$IC(\{a,\ t\}) = IC(a) + IC_{extended}(a \to t) = IC(a) + IC(t) - wIC(a) \tag{17}$$

where w denotes the weight of the association between parent and the term, In SORA[12] $w = 1$ and in WIS[13], the weight is computed from the numbers of their children as $w = \frac{|C(t)|}{|C(a)|}$. Eq. (17) provides a means of computing the IC of a given term set without repeatedly summing the shared ICs of ancestors; and SORA and WIS methods define the semantic similarity of a term set T as the IC of the term set.

Functional similarity measures between two genes. Functional similarity (*FS*) between two genes is computed using the ICs of individual terms (term-based) or the semantic similarities between the pairs of terms (term pair-based) or among the set of terms (term set-based). Let T_{g_1} and T_{g_2} be the set of GO terms annotating genes g_1 and g_2, respectively. Term-based measures such as GIC[1] (Jaccard index), DIC[39] (dice index), and UIC[39] (universal index) are defined using ICs of individual terms:

GIC[1]:

$$FS(g_1,\ g_2) = \frac{\sum_{t\in T_{g_1}\cap T_{g_2}} IC(t)}{\sum_{t\in T_{g_1}\cup T_{g_2}} IC(t)} \tag{18}$$

DIC[39]:

$$FS(g_1,\ g_2) = \frac{2 \times \sum_{t\in T_{g_1}\cap T_{g_2}} IC(t)}{\sum_{t\in T_{g_1}} IC(t) + \sum_{t\in T_{g_2}} IC(t)} \tag{19}$$

UIC[39]:

$$FS(g_1,\ g_2) = \frac{\sum_{t\in T_{g_1}\cap T_{g_2}} IC(t)}{\max\{\sum_{t\in T_{g_1}} IC(t),\ \sum_{t\in T_{g_2}} IC(t)\}} \tag{20}$$

Term pair-based methods are defined by pairwise gene similarities and use statistical closeness measures such as Average (AVG) and Maximum (MAX), Best-Match Average (BMA) and Average Best-Matches (ABM):

AVG:

$$FS(g_1,\ g_2) = \frac{1}{|T_{g_1}||T_{g_2}|}\sum\nolimits_{t_1\in T_{g_1},\ t_2\in T_{g_2}} IC\ (\{t_1,\ t_2\}) \tag{21}$$

MAX:

$$FS(g_1,\ g_2) = \ \max\{IC(\{t_1,\ t_2\})\text{: } t_1 \in T_{g_1},\ t_2 \in T_{g_2}\} \tag{22}$$

BMA:

$$FS(g_1, g_2) = \frac{1}{2}\left(\frac{1}{|T_{g_1}|}\sum_{t_1 \in T_{g_1}} IC(\{t_1, t_2\}) + \frac{1}{|T_{g_2}|}\sum_{t_2 \in T_{g_2}} IC(\{t_1, t_2\})\right) \tag{23}$$

AMB:

$$FS(g_1, g_2) = \frac{1}{|T_{g_1}||T_{g_2}|}\left(\sum_{t_1 \in T_{g_1}} IC(\{t_1, t_2\}) + \sum_{t_2 \in T_{g_2}} IC(\{t_1, t_2\})\right) \tag{24}$$

Term set-based measures Overlap Ratio (OR) and Intersection to Union Ratio (IUR) were introduced by SORA[12] and WIS[13] methods, respectively, and use ICs of term sets:

OR:

$$FS(g_1, g_2) = \frac{1}{2}\left(\frac{IC(T_{g_1} \cap T_{g_2})}{IC(T_{g_1})} + \frac{IC(T_{g_1} \cap T_{g_2})}{IC(T_{g_2})}\right) \tag{25}$$

IUR:

$$FS(g_1, g_2) = \frac{IC(T_{g_1} \cap T_{g_2})}{IC(T_{g_1} \cup T_{g_2})} \tag{26}$$

Table 2 lists 46 *FS* measures itemized based on nine corpus-based and two structure-based semantic similarities of GO terms, which are combined using direct-term based measures (DIC, GIC, and UIC), pair-based measures (MAX, AVG, BMA, and ABM), and set-based operations (OR and IUR). There exist several packages implementing some of these measures: for example, GOSemSim[40], an R package implementing five measures, Resnik, Lin, Jiang, Schlicker and Wang's similarity, and A-DaGO-Fun[41], a python package, implementing 44 corpus-based *FS* measures. We extend these works to include structure-based methods such as SORA and WIS and developed EnrichFunSim (https://gitlab.com/liuwt/EnrichFunSim), a python package that implements all 46 *FS* measures and corresponding *FS** measures discussed in this paper.

Association of a candidate gene with a disease. The association of a candidate gene with a specific disease is computed as the *FS* between a set of known disease associated genes and a candidate gene. Given the set G_d of disease-associated genes of disease *d* and a candidate gene *g*, the association of gene with the disease is given by *FS* (G_d, *g*), the functional similarity between G_d and *g*.

Enriched Functional Similarity. Generally, *FS* measures of genes are derived from ICs of GO terms annotating the genes and the ICs are computed by either how terms are annotated or represented in the corpus or how terms are represented in the DAG. However, how GO terms are represented in the querying gene pair or set has not been considered when evaluating the *FS* of a pair or a set of genes. In other words, the existing FS measures do not consider how GO terms are represented or enriched by the querying pair of genes. The enrichment of a GO term by the pair of gene is dependent on whether the term is annotated by one gene (i.e., partial annotation) or by both genes (i.e. complete annotation) in the pair. For example, consider the protein pair (P01906, P17693) shown in Fig. 2 and the GO DAG of their terms set. The terms *GO*:006955 and *GO*:0019882 are commonly annotated to both proteins; *GO*:0002504 is annotated to only P01906; and *GO*:002474 and *GO*:0006968 are annotated to only P17693. The enrichment of a term by the querying gene pair depends on whether the term annotates one gene or both the genes. Existing *FS* measures does not take note of whether the term annotates only one gene or both genes when computing their ICs.

We introduce GO term enrichment by the pair of genes in the computation of IC and propose enriched *FS* (FS^*) between two genes. The probability of term *t* annotating *k* genes in a gene set of size *n* is given by a hypergeometric distribution as

$$p(k, n|t) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \text{ where } k = \{0, \cdots, n\} \tag{27}$$

where *N* is the number of genes in the corpus and *M* is the number of genes that annotate term *t*. Figure 3 shows the Venn diagram depicting how a gene is enriched by the annotating corpus and the querying gene set. Note that in (27), the annotation of term *t* by the corpus is represented by set {*N*, *M*} and by the querying pair is by set {*n*, *k*}.

We define the enriched probability term $p^*(t)$ as the joint probability *p* (*k*, *n*, *t*) of *k* genes in a querying set of *n* genes, being annotated by term *t* as

$$p^*(t) = p(k, n, t) = p(k, n|t)p(t) \tag{28}$$

and enriched IC, $IC^*(t) = -\log(p^*(t))$ is given by

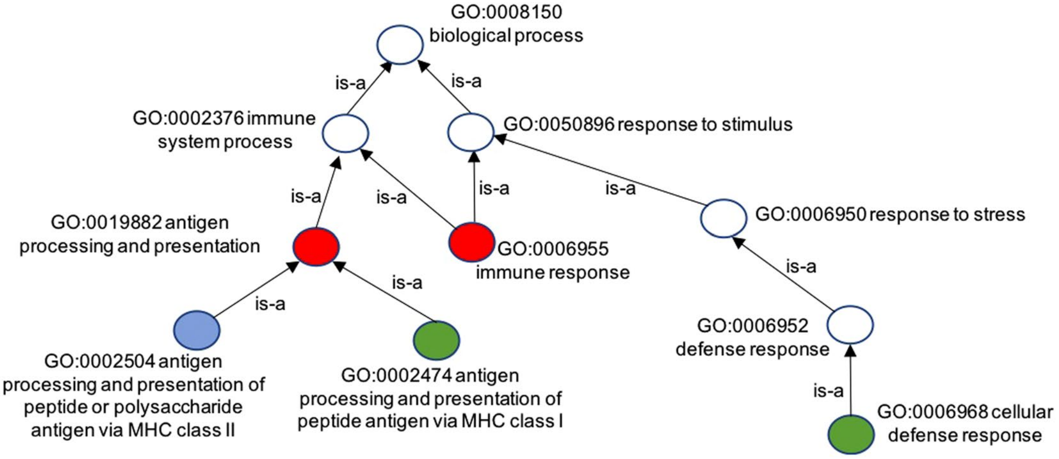
GO:0008150
biological process
is-a
is-a
GO:0002376 immune system process
GO:0050896 response to stimulus
is-a
is-a
is-a
is-a
GO:0019882 antigen processing and presentation
GO:0006955 immune response
GO:0006950 response to stress
is-a
is-a
is-a
GO:0002504 antigen processing and presentation of peptide or polysaccharide antigen via MHC class II
GO:0002474 antigen processing and presentation of peptide antigen via MHC class I
GO:0006952 defense response
is-a
GO:0006968 cellular defense response

Figure 2. The DAG representing the sets of GO terms annotating proteins: P01906 and P17693. Protein P01906 is annotated by terms *GO*:0006955, *GO*:0019882 and *GO*:0002504; and P17693 is annotated by terms *GO*:0006955, *GO*:0019882, *GO*:0002474 and *GO*:0006968. Blue terms denote terms annotating only protein P01906, green terms denote terms annotating only protein P17693, and red terms denote terms annotating both proteins.

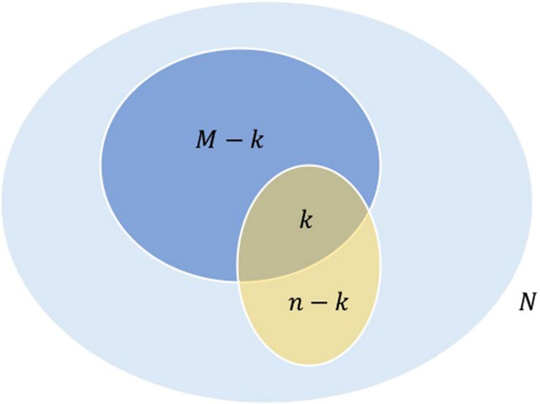

Figure 3. Venn diagram illustrating a GO term annotating the genes in the corpus (blue) and the querying set (yellow): *M* genes in the corpus of *N* genes and *k* genes in a querying set of *n* genes are annotated by the GO term.

$$IC^*(t) = -\log(p(k, n|t)) + IC(t) \tag{29}$$

where *IC* (*t*) is computed from the prior knowledge by annotations of corpus or by the expert knowledge embedded in the DAG. Note that by including {*n*, *k*} in the IC of term, IC^* (*t*) accounts for the enrichment of the term by querying gene set.

From (27) and (29), for a querying gene pair (*n* = 2), when term *t* is partially annotated (*k* = 1):

$$IC^*(t) = -\log\left(\frac{2M(N-M)}{N(N-1)}\right) + IC(t) \tag{30}$$

and when term *t* is fully annotated (*k* = 2):

$$IC^*(t) = -\log\left(\frac{M(M-1)}{N(N-1)}\right) + IC(t) \tag{31}$$

By substituting IC^* (*t*) for *IC* (*t*) in the computation of *FS* measures of gene pairs, we obtain corresponding enriched FS^* measures. The method is applicable to both corpus-based and structure-based methods evaluating IC. It also provides a means for incorporating information of querying gene pair and the annotating corpus in the structure-based methods of evaluating *FS*.

Software availability. The software (python code) and all the benchmark datasets evaluation (R script) are available at https://gitlab.com/liuwt/EnrichFunSim.

References

1. Pesquita, C. *et al*. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics* **9**, S4 (2008).
2. Bien, S. J. *et al*. Bi-directional semantic similarity for gene ontology to optimize biological and clinical analyses. *Am. Med. Informatics Assoc.* **19**, 765–774 (2012).
3. Guo, X., Liu, R., Shriver, C. D., Hu, H. & Liebman, M. N. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics* **22**, 967–973 (2006).
4. Moreau, Y. & Tranchevent, L.-C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.* **13**, 1–14 (2012).
5. Cheng, L., Li, J., Ju, P., Peng, J. & Wang, Y. SemFunSim: A New Method for Measuring Disease Similarity by Integrating Semantic and Gene Functional Association. *PLoS One* **9**, e99415 (2014).
6. Resnik, P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Artif. Intell. Res.* **11**, 95–130 (1999).
7. Lin, D. An Information-Theoretic Definition of Similarity. In *Proceedings of the 15th ICML* 296–304 (1998).
8. Mazandu, G. K. & Mulder, N. J. Information Content-Based Gene Ontology Semantic Similarity Approaches: Toward a Unified Framework Theory. *Biomed Res. Int.* **2013** (2013).
9. Schlicker, A., Domingues, F. S., Rahnenführer, J. & Lengauer, T. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* **7** (2006).
10. Couto, F. M. & Silva, M. J. Disjunctive shared information between ontology concepts: application to Gene Ontology. *Biomed. Semant.* **2**, 1–16 (2011).
11. Ehsani, R. & Drabløs, F. TopoICSim: a new semantic similarity measure based on gene ontology. *BMC Bioinformatics* **17**, 1–14 (2016).
12. Teng, Z. *et al*. Measuring gene functional similarity based on group-wise comparison of GO terms. *Bioinformatics* **29**, 1424–1432 (2013).
13. Tian, Z., Wang, C., Guo, M., Liu, X. & Teng, Z. An improved method for functional similarity analysis of genes based on gene ontology. *BMC Syst. Biol.* **10** (2016).
14. Tversky, A. Features of similarity. *Psychol. Rev.* **84**, 327 (1977).
15. Wilcoxon, F. Individual comparisons of grouped data by ranking methods. *Biometrics Bull.* **1**, 80–83 (1945).
16. Zhang, P. *et al*. Gene functional similarity search tool (GFSST). *BMC Bioinformatics* **7** (2006).
17. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23**, 1274–1281 (2007).
18. Mazandu, G. K. & Mulder, N. J. Information Content-Based Gene Ontology Functional Similarity Measures: Which One to Use for a Given Biological Data Type? *PLoS One* **9**, 1–20 (2014).
19. Lord, P. W., Stevens, R. D., Brass, A. & Goble, C. A. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* **19**, 1275–1283 (2003).
20. Yang, H., Nepusz, T. & Paccanaro, A. Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. *Bioinformatics* **28**, 1383–1389 (2012).
21. Pesaranghader, A., Matwin, S., Sokolova, M. & Beiko, R. G. simDEF: Definition-based Semantic Similarity Measure of Gene Ontology Terms for Functional Similarity Analysis of Genes. *Bioinformatics* 1–7 (2015).
22. Pesquita, C. *et al*. CESSM: Collaborative Evaluation of Semantic Similarity Measures. *Challenges Bioinforma* (2009).
23. Gillis, J. & Pavlidis, P. Assessing identity, redundancy and confounds in Gene Ontology annotations over time. *Bioinformatics* **29**, 476–482 (2013).
24. Chabalier, J., Mosser, J. & Burgun, A. A transversal approach to predict gene product networks from ontology-based similarity. *BMC Bioinformatics* **8**, 1–12 (2007).
25. Maetschke, S. R., Simonsen, M., Davis, M. J. & Ragan, M. A. Gene Ontology-driven inference of protein – protein interactions using inducers. *Bioinformatics* **28**, 69–75 (2012).
26. Jain, S. & Bader, G. D. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics* **11**, 562 (2010).
27. Salwinski, L. *et al*. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32**, 449–451 (2004).
28. Zeng, X., Liao, Y., Liu, Y. & Zou, Q. Prediction and validation of disease genes using HeteSim scores. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **14**, 687–695 (2017).
29. Zou, Q., Li, J., Song, L., Zeng, X. & Wang, G. Similarity computation strategies in the microRNA-disease network: A survey. *Brief. Funct. Genomics* **15**, 55–64 (2016).
30. Shi, C., Kong, X., Huang, Y., Yu, P. S. & Wu, B. HeteSim: A General Framework for Relevance Measure in Heterogeneous Networks. *IEEE Trans. Knowl. Data Eng.* **26**, 2479–2492 (2014).
31. Schlicker, A., Lengauer, T. & Albrecht, M. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. *Bioinformatics* **27**, i561–i567 (2011).
32. Williams, D. A. The comparison of several dose levels with a zero dose control. *Biometrics* **28**, 519–31 (1972).
33. Steiger., J. H. Tests for comparing elements of a correlation matrix. *Psychol. Bull.* **87**, 245–251 (1980).
34. Graham, Y. & Baldwin, T. Testing for Significance of Increased Correlation with Human Judgment. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)* 172–176 (2014).
35. Longnecker, M. T. A modified Wilcoxon rank sum test for paired data. *Biomelrika* **70**, 510–513 (1983).
36. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nat. Genet*. **25** (2000).
37. Gene, T. & Consortium, O. Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).
38. Jiang, J. J. & Conrath, D. W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. in *Proceedings of International Conference Research on Computational Linguistics* 1–15 (1997).
39. Mazandu, G. K. & Mulder, N. J. DaGO-Fun: tool for Gene Ontology-based functional analysis using term information content measures. *BMC Bioinformatics* **14**, 284 (2013).
40. Yu, G. *et al*. GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978 (2010).
41. Mazandu, G. K., Chimusa, E. R., Mbiyavanga, M. & Mulder, N. J. A-DaGO-Fun: An adaptable Gene Ontology semantic similarity based functional analysis tool. *Bioinformatics* 1–3 https://doi.org/10.1093/bioinformatics/btv590 (2015).

Acknowledgements

This research was partially supported by Tier-2 MOE2016-T2-1-029 grant by the Ministry of Education, Singapore.

Author Contributions

W.L. and J.C.R. conceived the project. W.L. designed and performed experiments. The analysis and interpretation of the results were done by W.L., J.L. and J.C.R. W.L. wrote the manuscript and J.L. and J.C.R. contributed to the manuscript and approved its final version.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-30455-0.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.