SWE: a novel method with semantic-weighted edge for measuring gene functional similarity

1st Zhen Tian
School of Information Engineering
Zhengzhou University
Zhengzhou, China
ieztian@zzu.edu.cn

2nd Haichuan Fang
School of Information Engineering
Zhengzhou University
Zhengzhou, China
hcfang@gs.zzu.edu.cn

3rd Yangdong Ye
School of Information Engineering
Zhengzhou University
Zhengzhou, China
ieydye@zzu.edu.cn

4th Zhenfeng Zhu
School of Information Engineering
Zhengzhou University
Zhengzhou, China
iezfzhu@zzu.edu.cn

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313495

***Abstract*—In recent years, functional similarity has played an independent role in some biological fields such as gene clustering, gene functional prediction, and evaluation for protein-protein interaction. In this premise, some effective methods have already been proposed based on Gene Ontology (GO). Although these mainstream methods achieve the purpose for measuring gene functional similarity, they may have some deficiency when calculating the Information Content (IC) of GO terms. Consequently, measuring the functional similarity accurately is still a meaningful objective of research. In this paper, a novel method called SWE, is proposed for measuring gene functional similarity based on the GO graph. Firstly, an algorithm to measure terms' semantics based on their information in the GO graph is put forward. The information of GO terms mainly contains their depth, ancestors and descendants. Secondly, we calculate the IC of a term set by means of retrieving the inherited relationship between terms in a term set. Finally, the functional similarity between two genes is computed based on the IC overlap ratio of term sets annotating two genes respectively. Results demonstrate that SWE is superior to existing methods in some experiments such as functional classification of genes in a biological pathway, protein-protein interaction and gene expression experiment. Further analysis demonstrates that SWE takes not only the specificity of terms into account, but their information in the GO graph, both of which are shown to be consistent with human perspectives.**

***Index Terms*—Gene Ontology, Information Content, Specificity of terms, Inherited relationship, Gene Functional Similarity**

I. Introduction

GO [1]–[3], a controlled vocabulary of terms, is a directed acyclic graph(DAG) and consists of three orthogonal ontologies: biological process (BP), cellular component (CC) and molecular function (MF). In the three ontologies, nodes represent terms and edges represent the relationship of two connected nodes. GO is a tree-like hierarchy but has a good deal of paths extending from the root node. The terms at higher levels are generic while the terms located in lower levels are more specific.

GO annotations (GOA) [4]–[6], an essential database for calculating gene functional similarity, are created by associating a gene or gene product with GO terms. Lines in GOA contain information about genes and their corresponding annotating terms.

In recent years, many gene functional similarity approaches have been proposed by researchers. Despite their usefulness, calculating gene functional similarity efficiently and accurately remains a challenging task. When measuring gene functional similarity, employing IC is a reasonable choice in the beginning because IC can measure how specific a term is. The methods of calculating the IC of a term could be generally classified two categories: corpus-based [7]–[12] and structured-based [13]–[16].

For corpus-based methods, such as Resnik [7], Jiang and Conrath [8] and Lin [9], they calculate the IC of a term t via the following definition:

$$IC_{corpus}(t) = -\log(p(t)) \tag{1}$$

where $p(t)$ is the occurrence probability of a term t and its descendants in a specific corpus such as GOA. As is shown in (1), the specificity of a term is fully dependent on the number of genes it annotates in a certain corpus. However, there is a fact that it's difficult to obtain the IC of a term correctly especially more than one corpus contains that term because the occurrence probability of the term may not different in different corpus. There is a evidence [17] that annotation corpus changes always have a high impact on IC.

Alternatively, IC can also be calculated based on the GO structure. In the GO graph, the lower a term's level, the more special the term. Thus, the IC of a term at high level should be less than the term located in the leaves. In this premise, David Sánchez [13] designed a model to compute the IC of a

term, which is defined as:

$$IC(c) = -\log\left(\frac{\frac{|leaves(c)|}{|subsumers(c)|}+1}{max\_leaves+1}\right) \quad (2)$$

where $subsumers(c)$ and max_leaves represent the ancestor terms of term c and total numbers of leaves respectively in the GO graph. In addition, $leaves(c)$ is the intersection of the descendants of term c and the leaves of GO graph. The fact that 1 is added to the numerator and denominator is to solve the probable circumstance that $log(0)$. Nevertheless, Sánchez's model not fully consider term's information such as their depth in the GO graph but only take their number into account. Besides, the descendants of a term which are not leaves are ignored. Furthermore, WIS [15] and Teng [14] considered that the IC of a term is proportional to its depth in the GO graph and they make some useful improvements for calculating IC as a result.

According to the GO graph and GO annotations, gene functional similarity approaches can generally be classified two categories: pairwise methods [7]–[9], [18] and groupwise methods [14], [15], [19]–[22].

Resnik [7] proposed a pairwise method based on the nodes of Go graph, where the semantic similarity between two terms can be computed by:

$$Sim_{Resnik}(t_1, t_2) = IC(LCA(t_1, t_2)) \quad (3)$$

where $LCA(t_1, t_2)$ is the lowest common ancestor of term t_1 and t_2. Besides, some researchers have improved this method such as Wang, Lin and Jiang. When calculating the semantic similarity of two terms, they considered the IC of terms themselves except the IC of LCA.

Groupwise methods measure gene functional similarity through integrating the terms annotating two genes into a group first. Then, they employ different methods to process the term set and calculating gene functional similarity according to the processed result last. Gentleman [21] raised a method called simUI and the formula of this method is defined as:

$$simUI(G_1, G_2) = \frac{|S_{G_1} \cap S_{G_2}|}{|S_{G_1} \cup S_{G_2}|} \quad (4)$$

where S_{G_1} and S_{G_2} represent the term set annotating gene G_1 and G_2 respectively. Inspired by simUI, simGIC [22] take the IC value of terms into consideration. For genes G_1 and G_2, simGIC is calculated by:

$$simGIC(G_1, G_2) = \frac{\sum_{t_i \in S_{G_1} \cap S_{G_2}} IC(t_i)}{\sum_{t_j \in S_{G_1} \cup S_{G_2}} IC(t_j)} \quad (5)$$

While simUI doesn't consider the specificity of terms in the GO graph and simGIC only take the IC of terms into consideration, which may lead to some semantics loss.

In addition to the methods mentioned above, there are some approaches [19], [20] based on vector space and we called the basic vector space model VSM in this paper. In VSM, the one-hot coding is adopted to assign values to vectors and the dimension of a vector is equal to the total number of terms in GO. Each dimension represents by a binary digit, denoting the presence or absence of a term in the set annotating the gene. The similarity of two genes calculated by VSM is defined as:

$$Sim_{VSM}(G_1, G_2) = \frac{v_1 \cdot v_2}{|v_1|\,|v_2|} \quad (6)$$

where v_1 and v_2 correspond to the term set annotating gene G_1 and G_2 respectively. Mathematically speaking, the similarity between two genes measured by VSM method is the cosine similarity of two vectors. In human perspectives, VSM ignores the relationship of terms and the specificity of terms in the GO graph.

In summary, a method for measuring gene functional similarity may be an effective method when it takes the specificity of nodes and edges into full consideration in the GO graph. Therefore, a novel method based on Semantic-Weighted Edge called SWE is proposed to measure gene functional similarity precisely. Firstly, we calculate the IC of a term considering the max depth of itself and its ancestors as well as the topology of its descendant in the GO graph. Secondly, we compute the IC of the intersection and the union set of term set annotating two genes utilizing the inherited relationship between terms respectively. Finally, the gene function similarity is the ratio between the IC of the intersection and the union set.

II. Methods

A. Calculate the IC of a term

Generally speaking, the lower a term's level, the more special the term, so it's evident that terms at lower levels have larger IC than terms at higher levels. In addition, Terms with more ancestors will be more specific than terms with less ones. An effective method consider not only the specificity of the term itself, but also inheritance relationships in the GO graph. In this premise, an effective method is given by:

$$IC(t) = log(depth(t)) * \left(log\left(\sum_{t_i \in Ance(t)} depth(t_i)\right) + 1\right) * \left(1 - \frac{log(|Desc(t)|)}{log(N+1)}\right) \quad (7)$$

where $depth(t)$ represents the max depth of term t, $Ance(t)$ denotes the ancestor terms of term t, $Desc(t)$ represents the descendant terms of term t and N denotes the total terms in the GO graph.

B. Calculate the IC of a term set based on semantic-weighted edge

Firstly, in terms of the inheritance relationship existing in the GO graph, the weight of an edge based on semantic-weighted edge is expressed as:

$$\omega_{ij} = \frac{\sum_{t_m \in Desc(t_j)} IC(t_m)}{\sum_{t_n \in Desc(t_i)} IC(t_n)} \quad (8)$$

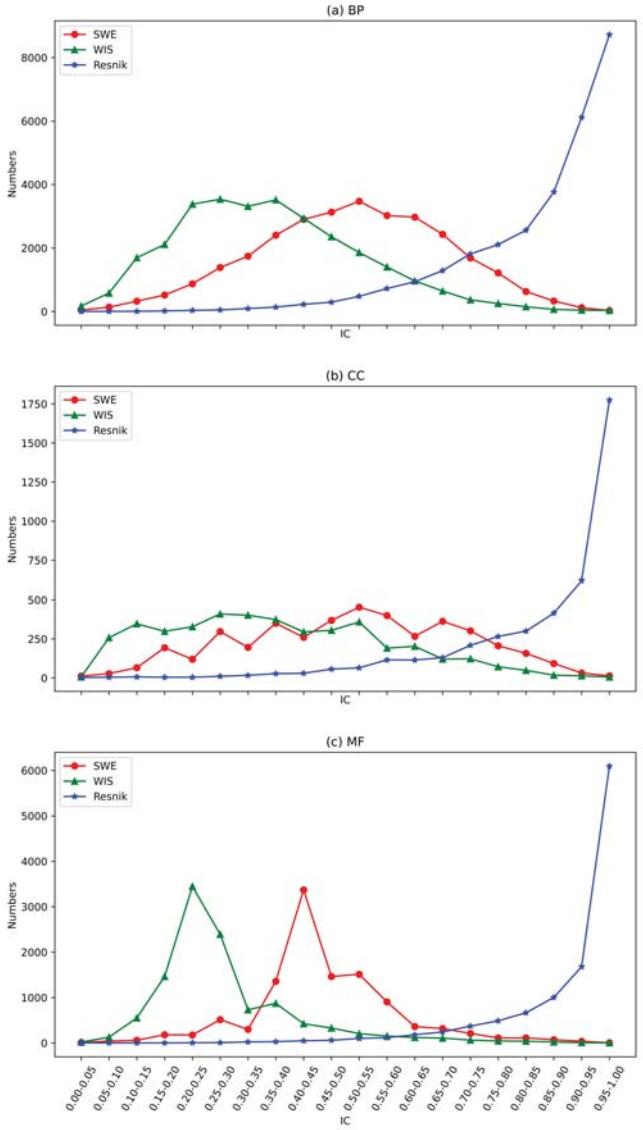

Fig. 1. The distributions of term quantity based on (a):BP ontology, (b):CC ontology and (c):MF ontology for three methods. The abscissa represents the range of IC value and the ordinate denotes the quantity of terms

where t_i is a parent term of t_j and $Desc(t)$ is the descendant term set of term t. In order to fully consider the specificity of nodes and edges, we use the IC of the term set as the parameters of weight instead of simply use the numbers of the term set.

Subsequently, according to the true path rule, a gene annotated with some terms also is annotated with their ancestor terms. In a term set existing inheritance relationship between terms, the IC of a term t_j is consisted of two parts: $IC_{inherited}(t_j)$ and $IC_{own}(t_j)$. The first one is inherited semantics, which is calculated by:

$$IC_{inherited}(t_j) = \sum_{t_i \in Parent(t_j)} \omega_{ij} * IC(t_i) \tag{9}$$

where $Parent(t_j)$ is the parent term set of term t_j and ω_{ij} is computed by (8). The second one belong to itself, which is defined as:

$$IC_{own}(t_j) = IC(t_j) - IC_{inherited}(t_j) \tag{10}$$

Finally, the IC of a term set is the sum of the own IC of all terms in the set, which is given by:

$$IC(T) = \sum_{t \in T} IC_{own}(t) \tag{11}$$

Apparently, it can avoid to calculate the overlap semantics between two terms that using SWE to calculate the IC of a term set.

C. Calculate the functional similarity between genes

Given two genes G_1 and G_2 annotated with term sets T_{G_1} and T_{G_2} respectively, the functional similarity between G_1 and G_2 is defined as:

$$simSWE(G_1, G_2) = \frac{IC(T_{G_1} \cap T_{G_2})}{IC(T_{G_1} \cup T_{G_2})} \tag{12}$$

where $\cap$ and $\cup$ represent the computation of intersection and union respectively.

III. Validations and Results

A. The database of GO and GO annotations

Gene Ontology file is downloaded from the Gene ontology database(http://geneontology.org/docs/download-ontology/, dated December 2019) containing 44,674 ontology terms subdivided into 29,380 biological process terms, 4,181 cellular component terms and 11,113 molecular function terms.

Gene Ontology annotation files for proteins are downloaded from the Gene Ontology database(http://geneontology.org/docs/download-go-annotations/) for Homo sapiens(dated December 2019) and Saccharomyces cerevisiae(dated December 2019).

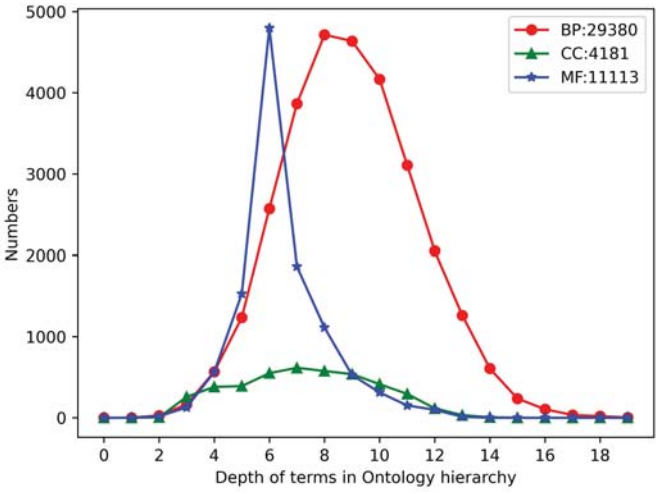

Fig. 2. The distributions of term quantity for BP, CC and MF terms based on the depth of GO graph

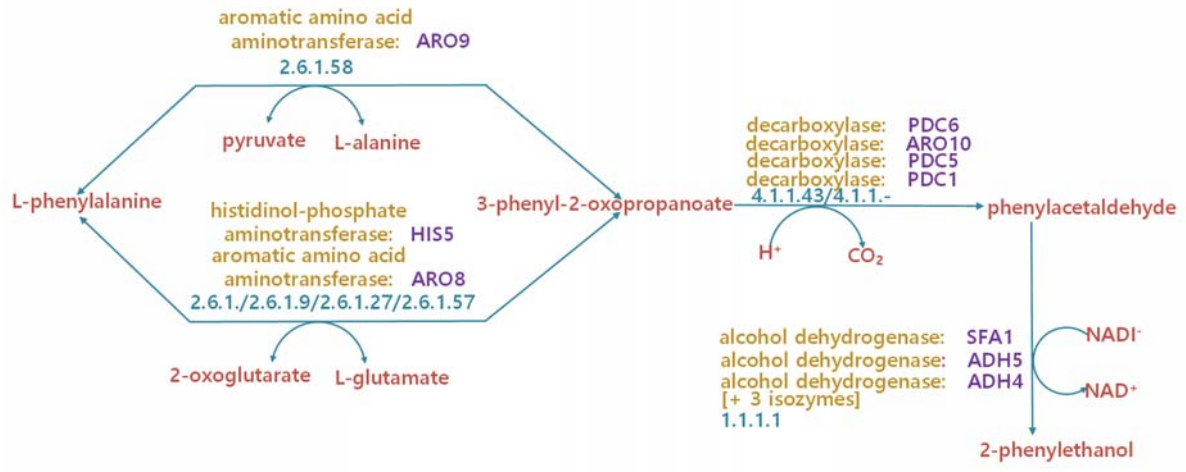

Fig. 3. Functions of genes in Saccharomyces cerevisiae Pathway: phenylalanine degradation

B. Biological pathway

Classifying the genes according to the molecular function is an important validation for a gene functional similarity measure [23]. In the current study, we utilize the yeastpathways data of Saccharomyces genome database(SGD, https://pathway.yeastgenome.org/) to study functional classification of genes.

C. Protein-Protein interaction dataset

For PPI experiments, we collect Homo sapiens datasets from WIS [15] and Saccharomyces cerevisiae datasets from Zhang [18]. After updating, we construct a new dataset which are core set of DIP database. Negative datasets with the same number of PPIs for yeast and human are independently generated by randomly choosing annotated gene pairs for BP, CC and MF ontology, which are absent from a combined dataset of all possible PPIs.

D. Gene expression dataset for Saccharomyces cerevisiae

The gene expression dataset was downloaded from Jain and Davis [24]. The dataset contains 11,966 S. cerevisiae gene pairs after updated, including gene pairs of 3,888, 4,211 and 3,867 based on BP, CC and MF ontology respectively.

E. The distribution of term IC in different ranges based on different method and of terms based on the range of depth in the GO

An effective computational model of IC is a key factor to measuring the gene functional similarity accurately. The distribution of term IC in different ranges based on BP, CC and MF ontologies are presented in Fig. 1.

In human perspectives, the IC of term which have high level is less than the term located on the lower level because the lower a term's level, the more special the term. The distribution of terms based on the range of depth in three ontologies are given in Fig. 2. According to the Fig. 2, it's shows that more than 89 percent of terms are at the middle levels of GO graph no matter what ontology they belong to. Therefore, the number of terms whose IC value are in the medium range should be in the majority.

For Resnik's model, there are more than 85 percent of term IC is more than 0.9 and this fact is show that Resnik's model not consider the specificity of different terms in the Ontology. WIS model has a huge improvement over Resnik's model. The distribution of IC does not change dramatically at any point like Resnik's model. However, a lot of terms are focused on the range of lower IC value presented by WIS and this curve is no consistent with the distribution of terms. On the contrary, the distribution of term IC given by SWE is consistent with the distribution of terms and human perspectives because we take the specificity of terms full into consideration based on the GO graph.

F. Functional classification of genes in a biological pathway

Using Saccharomyces genome database as the reference for our measurement of gene functional similarity is an effective way to compare the results of our method and other method. The SGD database contains more than 80 biological pathways. Most of these pathways contain at least three genes annotated by both GO molecular function terms and EC numbers. For example, there are 10 genes and 8 EC numbers in the 'phenylalanine degradation' pathway depicted in Fig. 3. We calculate the functional similarity among these genes based on MF ontology by method Resnik, method Wang, VSM and SWE. The result is presented in Fig. 4.

For Resnik's model, as is shown in Fig. 4(a), only a pair of genes has a similarity value greater than 0.5, although some of them have the same EC number in the biological pathway. For example, gene 'PDC1' and 'PDC5' have the same EC number in the biological pathway, but their similarity value only has 0.43. Therefore, the result obtained from Resnik's model inconsistent with human perception to a great extent. As is shown in Fig. 4(d), the result obtained by SWE can distinguish different genes precisely based on the EC number

genes have. But not only that, it also shows that for those genes whose EC number are entirely different, the similarity value between them is quite low. In summary, SWE can reflect the closeness of gene's biological meanings in human perspectives.

G. Protein-Protein interaction of S. cerevisiae and H. sapiens

The result of protein-protein interaction experiment is a desirable evaluation criterion for gene functional similarity. In our evaluation, we conduct out experiment using the PPI dataset aforementioned and the results of experiment were depicted by receiver operating characteristic (ROC) curves, with area under the curve(AUC) as the main accuracy criterion.

Functional similarity values between genes in S. cerevisiae and H. sapiens are calculated by SWE and other six methods which are simGIC, Resnik, WIS, simUI, VSM, Wang. AUC values for each method in terms of BP, CC and MF ontology on S. cerevisiae and H. sapiens PPI datasets are given in Table I. For S. cerevisiae, SWE run first on four experiments which based on BP_IEA+, CC_IEA+, BP_IEA- and MF_IEA- and was only inferior to method simGIC on MF_IEA+ and CC_IEA-. Furthermore, the results of method Wang and Resnik on PPI for S. cerevisiae datasets are inferior to other methods on almost all experiments. Therefore, this result demonstrates groupwise methods perform better than pairwise methods on S. cerevisiae datasets.For H. sapiens, it can be seen that SWE win the first place with an overwhelming advantage on BP_IEA+, BP_IEA- and MF_IEA-. simGIC obtain the best results on other three experiments and SWE run on simGIC's heels. For example, the AUC of simGIC

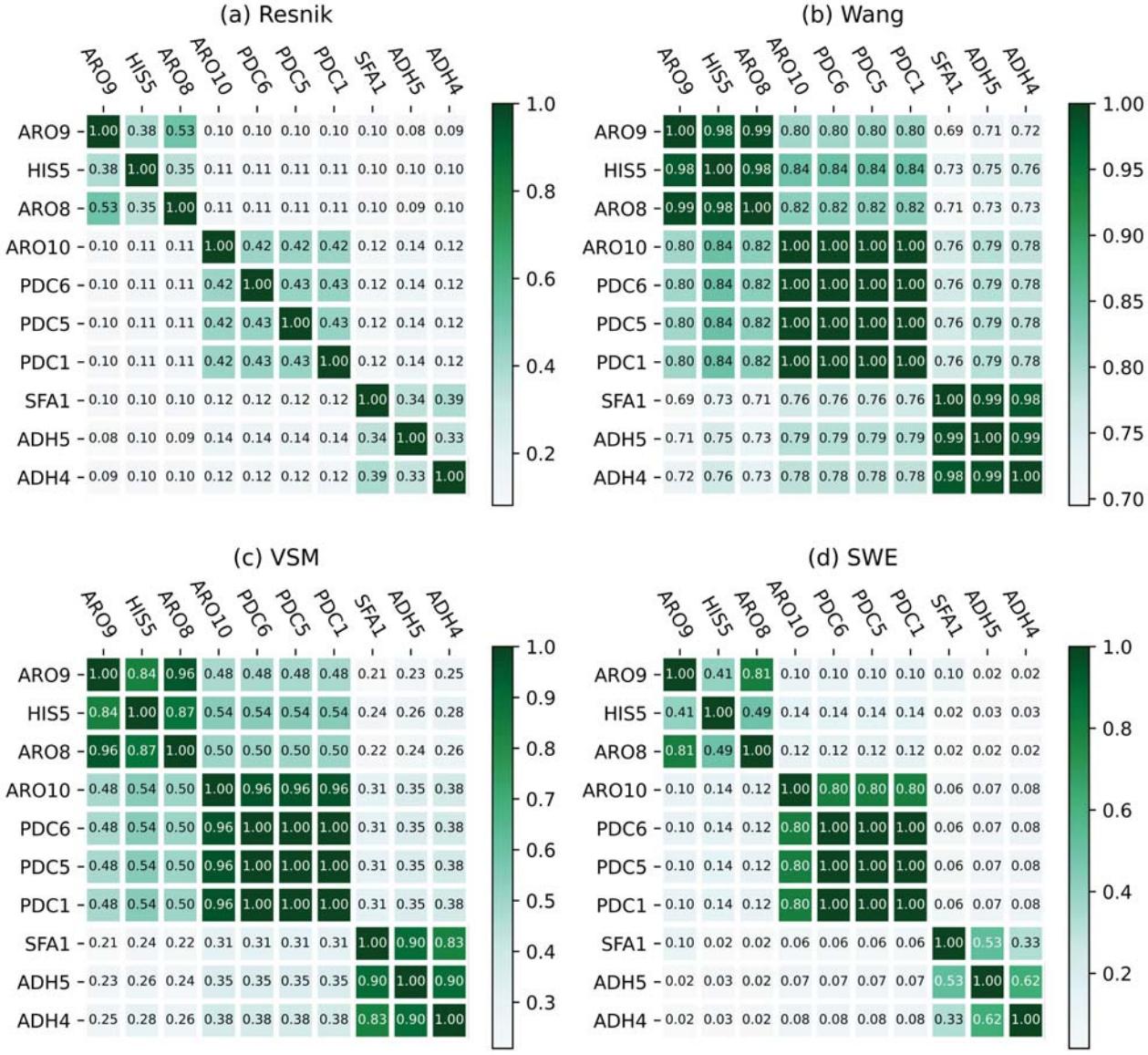

Fig. 4. Results of pathway experiments based on MF ontology using Method (a)Resnik, (b) Wang, (c)VSM and (d) SWE

TABLE I
AUC OF SEVEN FUNCTIONAL SIMILARITY MEASURES FOR BP, CC AND MF ONTOLOGY WITH IEA+ AND IEA- IN PPI DATASETS

Datasets	Methods	IEA+			IEA-		
		BP	CC	MF	BP	CC	MF
S. cerevisiae	SWE	**0.8234**	**0.8317**	0.7441	**0.8724**	0.8343	**0.7460**
	simGIC	0.8198	0.8223	**0.7497**	0.8647	**0.8392**	0.7023
	Resnik	0.7888	0.8211	0.6987	0.7949	0.8043	0.6182
	WIS	0.8184	0.8249	0.7371	0.8643	0.8122	0.7259
	simUI	0.8095	0.8213	0.7253	0.8447	0.8004	0.7098
	VSM	0.8115	0.8246	0.7294	0.8477	0.8033	0.7088
	Wang	0.7932	0.8028	0.7110	0.8262	0.7948	0.6905
H. sapiens	SWE	**0.8624**	0.7504	0.7228	**0.7940**	0.6839	**0.6907**
	simGIC	0.8381	**0.7614**	**0.7597**	0.7839	**0.6867**	0.6730
	Resnik	0.6696	0.6714	0.7033	0.7264	0.6638	0.6662
	WIS	0.8049	0.6734	0.6637	0.7718	0.6604	0.6835
	simUI	0.7921	0.6484	0.6208	0.7734	0.6586	0.6836
	VSM	0.7896	0.6564	0.6297	0.7825	0.6675	0.6732
	Wang	0.7334	0.6260	0.5824	0.7404	0.6466	0.6474

The best results are in bold

TABLE II
PEARSON'S CORRELATION COEFFICIENT OF SEVEN FUNCTIONAL MEASURES FOR BP, CC AND MF ONTOLOGY WITH GENE EXPRESSION DATASET (IEA+ AND IEA-)

Methods	IEA+			IEA-		
	BP	CC	MF	BP	CC	MF
SWE	0.4048	0.4197	0.2411	0.4403	0.5412	**0.1998**
simGIC	**0.4053**	0.4212	**0.2546**	**0.4418**	**0.5540**	0.1972
Resnik	0.3135	**0.4405**	0.2219	0.3818	0.5286	0.1439
WIS	0.3993	0.4125	0.2457	0.4318	0.5162	0.1980
simUI	0.3799	0.4003	0.2241	0.4252	0.5151	0.1889
VSM	0.3416	0.3621	0.1941	0.4024	0.4999	0.1806
Wang	0.2160	0.2292	0.0695	0.3141	0.4046	0.0563

The best results are in bold

only 0.0028 more than SWE on CC_IEA- experiment. In addition, the performance of groupwise methods is better than pairwise methods on most experiments.

H. Pearson's correlation coefficient analysis based on gene expression data

In order to analyze the correlation with gene expression data, we updated the datasets mentioned in III-D, and then randomly selected 3500 gene pairs for BP, CC and MF ontology respectively as the experimental data. We calculated the similarity value for every gene pair adopted seven methods which are SWE, simGIC, Resnik, WIS, simUI, VSM and Wang with IEA+ and IEA- on BP, CC and MF ontology first. Next, Pearson's correlation between gene functional similarity values and gene expression values can be computed and is listed in Table II. In general, higher this value, better is the measure.

As is shown in Table II, for different ontologies, the correlations which belong to CC ontology have the highest values on either $IEA-$ or $IEA+$ experiments, followed by BP and MF ontology. For different methods, simGIC ranks first on BP_IEA+, MF_IEA+, BP_IEA- and CC_IEA- experiments, while SWE has highest correlation on MF_IEA-. Method Resnik is superior to other methods on CC_IEA+, but its performance on BP_IEA+, MF_IEA+, BP_IEA-, MF_IEA- experiments are less than satisfactory. Although SWE runs first on MF_IEA- only and slightly worse than simGIC, their differences are almost negligible. In this experiment, the overall performance of groupwise methods such as SWE and simGIC are better than pairwise methods including method Wang and method Resnik, though method Resnik has highest correlation than other methods on CC_IEA+ experiment. In conclusion, the performance of SWE has reached a satisfactory level in gene expression experiment.

DISCUSSION

In this paper, an effective model is designed for measuring the IC of a term, which takes the overall specificities of the term based on GO graph. These specificities contain depth, ancestors, descendants and total number of terms. According to Fig. 2, GO is a tree contain much nodes at middle levels and less nodes at top and bottom levels. Therefore, in human perspectives, the number of terms with IC in medial range should be in the majority. The results listed in the Fig. 1 demonstrate that the performance of SWE is more consistent with human perspectives than the other two models.

What's more, SWE introduce the concept of semantic-weighted edge inspired by WIS. In order to avoid the defect exisiting in WIS, the edge is semantically weighted in SWE.

The results of SWE demonstrate it is effective to estimate gene functional similarity.

However, SWE has a shortage that is the definition of the IC of terms. How to evaluate the effectiveness of the model is a challenge problem. In this paper, we conducted an experiment and show the results in Fig. 1. But this is may be far from enough. In the current research, we evaluate and compare the results on three mainstream experiments and ignore some other metrics. We defer that as future work and some applications such as gene functional similarity network can have a widely use.

Conclusion

In the current study, we have proposed a novel functional similarity method between genes combining the gene ontology structure with the gene ontology annotation, namely SWE. SWE belongs to groupwise method and treats the GO terms annotating different genes as a term set. In addition, the specificity of every GO term and each edge in the ontology are took fully consideration.

As is shown in this article, we have resolved the difficulty of IC's definition first, and then the specificity of each term and edge were fully took into account. Finally, it can be found that the quantity distribution of terms in terms of GO depth and IC range has the same meaning, that is, the number of terms in the middle layers in the go structure is dominant and their IC are mostly between 0.3 and 0.7.

In order to estimate the reliability and effectiveness of SWE, three mainstream experiments, which are functional classification of genes in biological pathway, protein-protein interaction experiment and pearson's correlation coefficient analysis based on gene expression data, are applied into the evaluation. In addition to SWE, we also apply some classic approaches into each experiment as controls. The experimental results demonstrate that SWE is a more reliable and effective approach to measure gene functional similarity than other tested methods.

Acknowledgment

This work was supported by the Natural Science Foundation of China (Grant No. 61772475,61801432, 62003308).

References

[1] "Gene ontology consortium: The gene ontology (GO) database and informatics resource," *Nucleic Acids Res.*, vol. 32, no. Database-Issue, pp. 258–261, 2004.

[2] T. G. O. Consortium, "Expansion of the gene ontology knowledgebase and resources," *Nucleic Acids Res.*, vol. 45, no. Database-Issue, pp. D331–D338, 2017.

[3] T. G. O. Consortium, "The gene ontology resource: 20 years and still going strong," *Nucleic Acids Res.*, vol. 47, no. Database-Issue, pp. D330–D338, 2019.

[4] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler, "The gene ontology annotation (GOA) database: sharing knowledge in uniprot with gene ontology," *Nucleic Acids Res.*, vol. 32, no. Database-Issue, pp. 262–266, 2004.

[5] "Gene ontology annotations and resources," *Nucleic Acids Res.*, vol. 41, no. Database-Issue, pp. 530–535, 2013.

[6] R. P. Huntley, T. Sawford, P. Mutowo-Meullenet, A. Shypitsyna, C. Bonilla, M. J. Martin, and C. O'Donovan, "The GOA database: Gene ontology annotation updates for 2015," *Nucleic Acids Res.*, vol. 43, no. Database-Issue, pp. 1057–1063, 2015.

[7] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *CoRR*, vol. abs/1105.5444, 2011.

[8] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of the 10th Research on Computational Linguistics International Conference, ROCLING 1997, Taipei, Taiwan, August 1997* (K. Chen, C. Huang, and R. Sproat, eds.), pp. 19–33, The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), 1997.

[9] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconsin, USA, July 24-27, 1998* (J. W. Shavlik, ed.), pp. 296–304, Morgan Kaufmann, 1998.

[10] A. Islam and D. Z. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Trans. Knowl. Discov. Data*, vol. 2, no. 2, pp. 10:1–10:25, 2008.

[11] O. Ferret, "Testing semantic similarity measures for extracting synonyms from a corpus," in *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta* (N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, eds.), European Language Resources Association, 2010.

[12] M. Ahmed, C. Dixit, R. E. Mercer, A. Khan, M. R. Samee, and F. Urra, "Multilingual corpus creation for multilingual semantic similarity task," in *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020* (N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, eds.), pp. 4190–4196, European Language Resources Association, 2020.

[13] D. Sánchez and M. Batet, "A semantic similarity method based on information content exploiting multiple ontologies," *Expert Syst. Appl.*, vol. 40, no. 4, pp. 1393–1399, 2013.

[14] Z. Teng, M. Guo, X. Liu, Q. Dai, C. Wang, and P. Xuan, "Measuring gene functional similarity based on group-wise comparison of GO terms," *Bioinform.*, vol. 29, no. 11, pp. 1424–1432, 2013.

[15] Z. Tian, C. Wang, M. Guo, X. Liu, and Z. Teng, "An improved method for functional similarity analysis of genes based on gene ontology," *BMC Syst. Biol.*, vol. 10, no. 4, pp. 465–484, 2016.

[16] P. Dutta, S. Basu, and M. Kundu, "Assessment of semantic similarity between proteins using information content and topological properties of the gene ontology graph," *IEEE ACM Trans. Comput. Biol. Bioinform.*, vol. 15, no. 3, pp. 839–849, 2018.

[17] M. Milano, G. Agapito, P. H. Guzzi, and M. Cannataro, "An experimental study of information content measurement of gene ontology terms," *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 3, pp. 427–439, 2018.

[18] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinform.*, vol. 23, no. 10, pp. 1274–1281, 2007.

[19] J. Zhang, K. Jia, J. Jia, and Y. Qian, "An improved approach to infer protein-protein interaction based on a hierarchical vector space model," *BMC Bioinform.*, vol. 19, no. 1, pp. 161:1–161:14, 2018.

[20] S. Benabderrahmane, M. Smaïl-Tabbone, O. Poch, A. Napoli, and M. Devignes, "Intelligo: a new vector-based semantic similarity measure including annotation origin," *BMC Bioinform.*, vol. 11, p. 588, 2010.

[21] S. Falcon and R. Gentleman, "Using gostats to test gene lists for GO term association," *Bioinform.*, vol. 23, no. 2, pp. 257–258, 2007.

[22] C. Pesquita, D. Faria, H. P. Bastos, A. E. N. Ferreira, A. O. Falcão, and F. M. Couto, "Metrics for GO based protein semantic similarity: a systematic evaluation," *BMC Bioinform.*, vol. 9, no. S-5, 2008.

[23] S. Bandyopadhyay and K. Mallick, "A new path based hybrid measurefor gene ontology similarity," *IEEE ACM Trans. Comput. Biol. Bioinform.*, vol. 11, no. 1, pp. 116–127, 2014.

[24] S. Jain and G. D. Bader, "An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology," *BMC Bioinform.*, vol. 11, p. 562, 2010.