JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 21, Number 12, 2014

Pp. 964–974
DOI: 10.1089/cmb.2014.0143

A Weighted Multipath Measurement Based on Gene Ontology for Estimating Gene Products Similarity

LIZHEN LIU,[1] XUEMIN DAI,[1] HANSHI WANG,[1] WEI SONG,[1] and JINGLI LU[2]

ABSTRACT

Many different methods have been proposed for calculating the semantic similarity of term pairs based on gene ontology (GO). Most existing methods are based on information content (IC), and the methods based on IC are used more commonly than those based on the structure of GO. However, most IC-based methods not only fail to handle identical annotations but also show a strong bias toward well-annotated proteins. We propose a new method called weighted multipath measurement (WMM) for estimating the semantic similarity of gene products based on the structure of the GO. We not only considered the contribution of every path between two GO terms but also took the depth of the lowest common ancestors into account. We assigned different weights for different kinds of edges in GO graph. The similarity values calculated by WMM can be reused because they are only relative to the characteristics of GO terms. Experimental results showed that the similarity values obtained by WMM have a higher accuracy. We compared the performance of WMM with that of other methods using GO data and gene annotation datasets for yeast and humans downloaded from the GO database. We found that WMM is more suited for prediction of gene function than most existing IC-based methods and that it can distinguish proteins with identical annotations (two proteins are annotated with the same terms) from each other.

Key words: depth of LCAs, different weights, every path, gene ontology, semantic similarity.

1. INTRODUCTION

GENE ONTOLOGY (GO) (Ashburner et al., 2000) is a standard vocabulary of functional terms that is used for coherent annotation of gene products (Xu et al., 2008). GO comprises three orthogonal ontologies: biological processes (BP), molecular function (MF), and cellular components (CC) (Ashburner et al., 2000). These ontologies are represented as three directed acyclic graphs (DAGs) in which the nodes correspond to the terms describing a certain biological semantic category and the edges represent relationships between terms (Ashburner et al., 2000). The most common relationships are ''is-a,'' which indicates that the child is a subclass of the parent, and ''part-of,'' which means that the child is a component of the parent. In a DAG, a term inherits the semantics of its ancestors and distributes them to its descendants, and so the lower term contains more information (McHale, 1998). GO terms have been widely used to annotate genes and gene products with

[1]Information and Engineering College, Capital Normal University, Beijing, China.
[2]Agresearch Ltd., Hamilton, New Zealand.

functional terms (Wu et al., 2006a) in the Gene Ontology Annotation project (Barrell et al., 2009). GO data provide a novel way to measure the functional relationship between gene products regarding their MF and biological role, which enables biologists to benefit from studying gene correlation (Azuaje et al., 2006). In the past few decades, a variety of methods have been proposed to quantify the semantic similarity of GO terms.

These measures have been used in a broad range of applications, such as protein function (Fontana et al., 2009), cellular localization prediction (Lei and Dai, 2006), protein–protein interaction prediction (Wu et al., 2006b; Xu et al., 2008), automatic annotation validation (Couto et al., 2006), pathway modeling (Guo et al., 2006), and the evaluation of similarity between gene products with respect to expression profiles (Sevilla et al., 2005). However, all existing similarity measurement methods have drawbacks. Methods of Resnik (1995), Lin (1998), and Jiang and Conrath (1997) use information content (IC) to represent the specificity of GO terms. These methods hinder their ability to determine the functional similarity of genes. Resnik's method (Resnik, 2011) ignores the information contained in the structure of the ontology. A serious drawback of Lin's method (Lin, 1998) and Jiang's method (Jiang and Conrath, 1997) is that shallow annotation (two gene products are well annotated near the root of the ontology) makes the semantic similarity always be close to 1, which leads to a misleading result. Considering the drawbacks of these two methods, Wang et al. (2007) developed an IC-independent method, in which each edge is assigned a weight that was named the semantic contribution factor (ω_e), according to the type of relationship. They represent a GO term A as $\mathrm{DAG}_A = (A, T_A, E_A)$, a subgraph of GO, where T_A is the set of all ancestors of A and itself, and E_A is the set of corresponding links. The contribution of any term t to the semantics of a term A is defined as the S-value of the term t related to term A, $S_A(t)$, which can be calculated by

$$\begin{cases} S_A(A) = 1 \\ S_A(A) = \max\{\omega_e \cdot S_A(t') | t' \in \text{children of}(t)\} & \text{if } t \neq A \end{cases} \tag{1}$$

where ω_e is the semantic contribution factor for edge $e \in E_A$ linking term t with its child term t'. The semantic value of the term A, $S_A(A)$, is the aggregate semantic contribution of all terms in the DAG_A.

$$SV(A) = \sum_{t \in T_A} S_A(t) \tag{2}$$

Given that $\mathrm{DAG}_A = (A, T_A, E_A)$ and $\mathrm{DAG}_B = (B, T_B, E_B)$ for the two terms A and B, respectively, the semantic similarity between them, $S_{\mathrm{GO}}(A, B)$, is defined as

$$S_{\mathrm{GO}}(A, B) = \frac{\sum_{t \in T_A \cap T_B} [S_A(t) + S_B(t)]}{SV(A) + SV(B)} \tag{3}$$

Nevertheless, each ancestor term may have multiple direct child terms, and this property indicates that there may be multiple paths from a given term to its ancestor. Wang's method (Wang et al., 2007) ignored the affection brought by different paths. Therefore, it is not sensitive to GO updating.

Besides their individual drawbacks, a common problem in above methods is that different researchers may get different semantic similarity values for the same two GO terms if they use different gene annotation data. In some special cases, researchers need to have a fixed semantics when the terms are used to annotate genes. Hence, it is desirable to determine the semantic similarity of GO terms only based on their structure and annotation specification of gene ontologies (Wang et al., 2007). However, most ontology-structure-based methods (Gentleman, 2005; Pesquita et al., 2007) determine the semantic similarity either based on their distances to the closest common ancestor terms or based on the number of their common ancestor terms. There are some other methods that rely on distance measures (Couto et al., 2006; Wu et al., 2006b), for example, counting the number of edges on the shortest path between the involved terms in GO to calculate the similarity of GO terms and ignoring the affection brought by other paths, and thus the results obtained by these methods have a low accuracy.

2. METHODS

To address the weaknesses of the most existing methods, we propose a weighted multipath measurement (WMM) based on the DAG structure of GO to measure the semantic similarity of GO terms. In addition to

taking every path into account, we also assign a weight to each edge according to the type of relationship like Wang's method (Wang et al., 2007) instead of quantifying each edge as 1.

2.1. Semantic similarity of GO terms based on the structure of GO graph

Figure 1 shows a subgraph that is extracted from GO for a given term, for example, GO: 0048471. As shown in Figure 1, term t_3 is the father of t_5 and the brother of t_4. Interpreted by human common sense, the similarity value between t_3 and t_5 should be higher than that of t_3 and t_4. Some related studies also demonstrated that the path of two terms affected the semantic similarity of them, and the longer the path is,

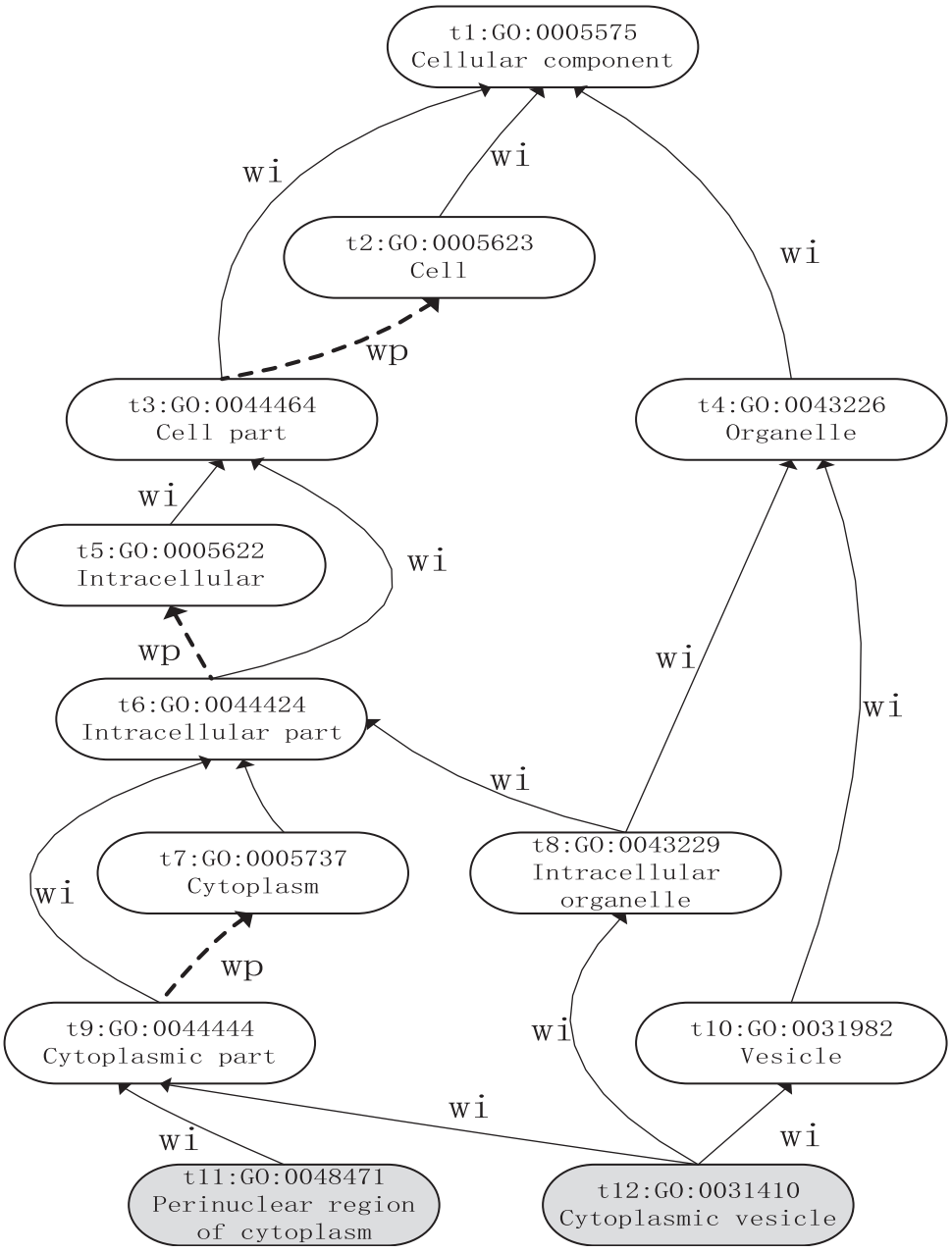

FIG. 1. A subgraph generated from GO of two seed terms, GO:0048471 and GO:0031410. Gene ontology is represented as a directed acyclic graph in which the nodes correspond to the terms and the edges represent relationships between terms. The solid arrows represent the ''is-a'' relationship and the dotted arrows show the ''part-of'' relationship. GO, gene ontology.

smaller the similarity value is. For example, sim (t_3, t_4) < sim (t_3, t_5) (the similarity value of term t_3 and term t_4 is smaller than that of term t_3 and term t_5), sim (t_7, t_8) < sim (t_6, t_8), and sim (t_{11}, t_{12}) < sim (t_9, t_{12}); the results obtained by Wang's method (Wang et al., 2007) and Jiang's method (Jiang and Conrath, 1997) both confirmed it (see Fig. 2).

Given two GO terms t_a and t_b in a DAG, the path from t_a to t_b is defined as

$$\text{path}(t_a, t_b) = \{\prec t_1, t_2, \ldots, t_n \succ | (t_a = t_1) \wedge (t_a = t_n) \wedge (\forall i : (1 \le i \le n) \wedge (t_i \in parents(t_{i+1})))\} \quad (4)$$

where function parent (t) represents the set of parents of t. If term t_a is an ancestor of term t_b, then there is at least one path from t_a to t_b, and so the set of ancestors of term t can be defined as

$$\text{ancestors} = \{t_1, t_2, \ldots, t_n, t | (\forall i : (1 \le i \le n) \wedge \text{paths}(t_i, t) \neq \oslash)\} \quad (5)$$

The common ancestors of t_a and t_b are defined as

$$\text{CAs}(t_a, t_b) = \text{ancestors}(t_a) \cap \text{ancestors}(t_b) \quad (6)$$

The set of lowest common ancestors (LCAs) is described as

$$\text{LCAs}(t_a, t_b) = \{t | (\text{node}(\text{paths}(t, t_a)) \cap \text{node}(\text{paths}(t, t_b)) \cap \text{CAs}(t_a, t_b)) = t\} \quad (7)$$

Considering that one GO term may have multiple parent terms with different semantic relations in a DAG, there may be multiple LCAs between two terms and more than one path from one term to a certain LCA, and so given two terms t_a and t_b, the similarity between them based on path distance (PD) can be defined as

$$\text{sim}_{\text{PD}(t_a, t_b)} = \frac{\alpha}{\overline{\text{dis}(\text{paths}(t, t_a))} + \overline{\text{dis}(\text{paths}(t, t_b))} + \alpha}, t \in \text{LCAs}(t_a, t_b) \quad (8)$$

where $\overline{\text{dis}(\text{paths}(t, t_x))}$ means the average PD from t to t_x, and LCAs (t_a,t_b) denotes the LCAs of t_a and t_b. α is a parameter ranging from 0 to 1.

Because the specificity of a GO term is usually determined by its location in the GO graph and a GO term's semantics (biological meanings) are inherited from all its ancestor terms, two terms sharing the same parent that are near the root of the ontology should have a larger semantic difference than two terms having the same parent that are far away from the root of the ontology, which means that the semantic similarity value of two terms whose closest common ancestor is near the root of the ontology is expected to be less

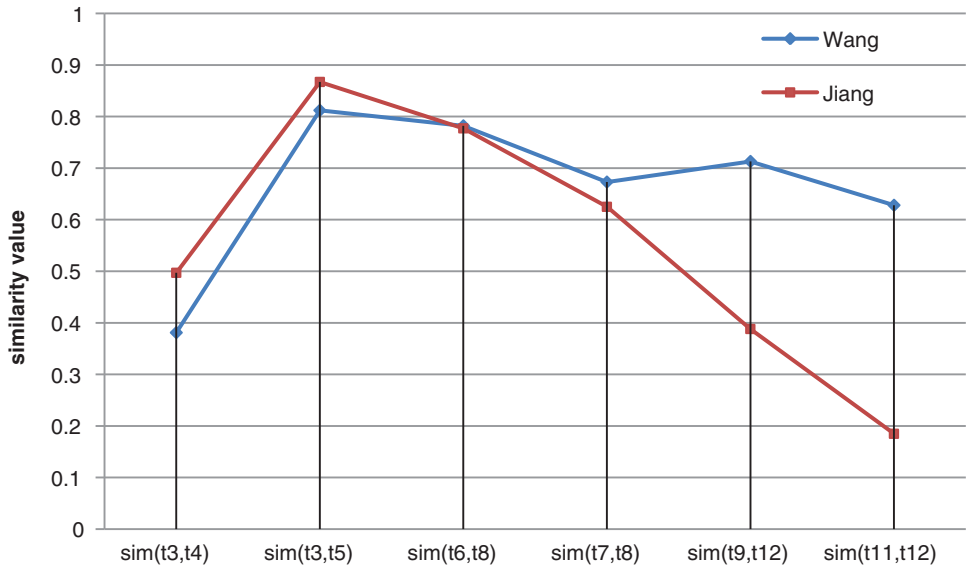

FIG. 2. Similarity values of certain term pairs obtained by Wang's and Jiang's methods. Term pairs are extracted from Figure 1. The results obtained by Wang's and Jiang's methods demonstrate that the path of two terms affects the semantic similarity of term pairs, and the longer the path is, the smaller the similarity value is.

than that of two terms whose LCA is far away from the root. So the semantic difference of two GO terms cannot be accurately represented by their distances to their closest common ancestor terms, and the distance from the LCAs to the root of the ontology is an important factor that affects the semantic similarity. For instance, sim (t_7, t_8) > sim (t_3, t_4); this is because the LCA of t_7 and t_8 is t_6, and the LCA of t_3 and t_4 is t_1, and t_1 is the root of the ontology. The results we derived are consistent with the results obtained by Jiang's method (Jiang and Conrath, 1997) and Wang's method (Wang et al., 2007). Figure 2 gives an intuitive comparison. Based on this, a concept called common path distance (CPD) is defined, which represents the average distance from the root to its LCAs of t_a and t_b. Thus, the similarity value is given by

$$\mathrm{sim}_{\mathrm{CPD}(t_a, t_b)} = \exp\left(\frac{-\beta}{\overline{\mathrm{dis}(\mathrm{paths}(\mathrm{root}, t))} + \beta}\right), t \in \mathrm{LCAs}(t_a, t_b) \tag{9}$$

where $\overline{\mathrm{dis}(\mathrm{paths}(\mathrm{root}, t))}$ means that the average distance from the root to t, that is, β, is a parameter ranging from 0 to 1. Existing methods always make the semantic similarity between any term and itself be 1, and it ignores differences of location on the ontology hierarchy. Equation 9 takes the depth of term into account, which is useful to calculate the semantic similarity of two proteins annotated by identical annotations (see the Comparing identical annotations section for details).

In fact, the common ancestor of two GO terms may have different contributions to the semantics of these specific child terms because the distance from terms to their common ancestor and the semantic relations (edges in the GO graph) may be different. In order to get a more accurate result, different weights are assigned to different edges according to the type of relationship. In this study, we assign w_i and w_p for '''is-a'' and ''part-of'' relations, respectively. The values of w_i and w_p are calculated according to Wang's method (Wang et al., 2007).

In order to get a more reasonable and accurate result, $\mathrm{sim}_{\mathrm{PD}}$ and $\mathrm{sim}_{\mathrm{CPD}}$ are combined by a parameter λ. Finally, the semantic similarity of two GO terms is defined as

$$\mathrm{sim}(t_a, t_b) = \lambda \mathrm{sim}_{\mathrm{PD}(t_a, t_b)} + (1-\lambda)\mathrm{sim}_{\mathrm{CPD}(t_a, t_b)} \tag{10}$$

By replacing $\mathrm{sim}_{\mathrm{PD}(t_a, t_b)}$ and $\mathrm{sim}_{\mathrm{CPD}(t_a, t_b)}$ with Equation 8 and Equation 9, respectively, we have

$$\mathrm{sim}(t_a, t_b) = \lambda . \frac{\alpha}{\overline{\mathrm{dis}(\mathrm{paths}(t, t_a))} + \overline{\mathrm{dis}(\mathrm{paths}(t, t_b))} + \alpha} + (1-\lambda) \cdot \exp\left(\frac{-\beta}{\overline{\mathrm{dis}(\mathrm{paths}(\mathrm{root}, t))} + \beta}\right), t \in \mathrm{LCAs}(t_a, t_b) \tag{11}$$

where α is a parameter that regulates the contribution rate of PD, and β regulates the contribution rate of the depth, and λ regulates the contribution rate of distance and depth to the similarity value. The values of them range from 0 to 1.

2.2. *Function similarity of gene products*

It is meaningless to compare only the semantic similarity of GO terms. All methods for calculating the similarity of GO terms are proposed to finally compare the function of genes or gene products.

There are two strategies quantifying the relationship between two gene products. One is pairwise strategy, which includes maximum (MAX), the average, and best-match average (BMA). The other strategy is named as groupwise, such as simUI (Gentleman, 2005), simGIC (Pesquita et al., 2007), and SORA (Teng et al., 2013). Different strategies are best suited in different contexts (Pesquita et al., 2009a; Guzzi et al., 2012), and no measure is clearly preferred over the others for biological problems. As noted by Pesquita et al. (2008), the maximum and average approaches have limitations from a biological point of view, and the BMA (Pesquita et al., 2007) performs better than the MAX (Pesquita et al., 2007), because MAX strategy considers the best match among all term pairs of two gene products, and it could be potentially affected by incorrect annotations or the noise along with the IEA annotations (Guzzi et al., 2012). In this article, the BMA is applied.

Let A and B be two gene products of interest, and T_A and T_B are the sets of all the GO terms assigned to proteins A and B, respectively. So the relationship strength between A and B is defined through pairwise

rule, that is, BMA. The BMA strategy finds all the best semantic similarity values for each term in T_A and T_B, and Equation 13 demonstrates it. The functional similarity of proteins A and B is calculated by

$$\mathbf{FPC}^{\mathrm{GO}}_{\mathrm{BMA}}(A,B)=\frac{\sum_{t_i\in T_A}\mathrm{FPC}(t_i,T_A)+\sum_{t_j\in T_B}\mathrm{FPC}(t_j,T_B)}{|T_A|+|T_B|} \tag{12}$$

where

$$\mathrm{FPC}(t_x,M)=\max_{m_y\in M}[\mathrm{FPC}(t_x,m_y)] \tag{13}$$

3. RESULTS AND DISCUSSION

Most of the existing semantic similarity measurements assess their performance in terms of correlations with sequence similarity (Mistry and Pavlidis, 2008), protein family similarity (Couto et al., 2005; Schlicker et al., 2006), and human ratings (Rong et al., 2006), etc. In this article, we first evaluate the performance of WMM by comparing the calculated semantic similarities with human ratings, and then we compare WMM with other existing methods using the Collaborative Evaluation of GO-based Semantic Similarity Measures (CESSM), an online tool for evaluating GO-based semantic similarity measures using Pearson's correlation with sequence, Pfam domain, and EC classification (ECC) similarities (Pesquita et al., 2009b); third, we compare the resolutions of WMM with other methods; finally, we discuss about the contribution of the WMM method in calculating the similarity of proteins annotated by identical annotations.

In this article, GO data (released in April 2012) and gene annotation datasets (released in April 2012) for yeast and human downloaded from the GO database (Ashburner et al., 2000) are used. The GO contains 22,506 BP, 2980 CC, and 9341 MF terms.

3.1. Comparison of WMM with human ratings

Ten biologists grade 25 pairs of GO terms with high, intermediate, and low similarities from 0 (no similarity) to 10 (synonymy) individually. After repeated testing, the maximum value of Pearson's correlation coefficient (PCC) is obtained when $w_i = 0.8$, $w_p = 0.6$, $\alpha = 0.8$, $\beta = 0.1$, and $\lambda = 0.6$. Table 1 shows the PCCs between similarity values obtained by seven measures and that of human ratings. A higher PCC represents better performance, which means that the method with a higher PCC has a higher capability to achieve semantic similarity closer to human performance. As shown in Table 1, the results obtained by the WMM method most closely match the human perception. Although WMM does not show significant improvement compared with Combine's (Guzzi et al., 2012) method, it outperforms two intrinsic methods, SimUI method (Gentleman, 2005) and Wang's method (Wang et al., 2007).

3.2. Evaluation of WMM by CESSM

CESSM is an online tool made available by the XLDB research team at the University of Lisbon. In total, 13,430 protein pairs involving 1039 distinct proteins and Uniprot GO annotations can be downloaded

Table 1. Pearson's Correlation Coefficients vs. Human Ratings

Method	*PCC*
WMM	0.8849
Combine's (Guzzi et al., 2012)	0.8638
Lin's (Lin, 1998)	0.8496
SimUI (Gentleman, 2005)	0.8397
Wang's (Wang et al., 2007)	0.8257
Resnik's (Resnik, 1995)	0.8241
ZZL's (Zhong et al., 2002)	0.7144

The PCCs between the semantic similarities obtained by the seven measures and that of human ratings. PCCs, Pearson's correlation coefficients; WMM, weighted multipath measurement.

Table 2. The Performances of Different Methods Evaluated by Collaborative Evaluation of GO-based Semantic Similarity Measures

GO	*Standard*	*WMM*	*GI*	*UI*	*RB*	*LB*	*JB*
MF	ECC	0.6506	0.6220	0.6366	0.6027	0.6417	0.5613
	Pfam	0.6386	0.6380	0.6181	0.5715	0.5639	0.4909
	SeqSim	0.6826	0.7127	0.5925	0.6683	0.6063	0.5459
BP	ECC	0.4606	0.3981	0.4023	0.4444	0.4352	0.3707
	Pfam	0.4586	0.4547	0.4505	0.4588	0.3727	0.3319
	SeqSim	0.8225	0.7733	0.7304	0.7397	0.6369	0.5864
CC	ECC	0.4006	0.3612	0.3575	0.3777	0.3683	0.2599
	Pfam	0.5376	0.4974	0.5214	0.4931	0.4851	0.2599
	SeqSim	0.8025	0.7500	0.6721	0.7113	0.6398	0.5014

Performance is measured by the PCC between the semantic similarity given by each method and the functional similarity estimated from EC classification (ECC), Pfam annotation (Pfam), and sequence similarity (SeqSim), respectively. Molecular function (MF), biological processes (BP), and cellular components (CC) are the three ontologies of gene ontology (GO).

from CESSM online. CESSM provides three standards of evaluation: ECC similarity (ECC), Pfam similarity (Pfam), and sequence similarity (SeqSim). CESSM is used in our research to compare the WMM method with existing methods, such as Resnik's (RB) method (Resnik, 1995), Lin's (LB) method (Lin, 1998), and Jiang's (JB) method (Jiang and Conrath, 1997), coupled with simGIC (GI) (Pesquita et al., 2007) and simUI (UI) (Gentleman, 2005) (i.e., GI, UI, RB, LB, and JB) in three ontologies (MF, BP, and CC). First, the performance is evaluated by measuring the PCC between the semantic similarities given by each method and the functional similarity estimated from ECC, Pfam annotation (Pfam), and sequence similarity (SeqSim), respectively, under different three ontologies, MF, BP and CC. Table 2 shows the performance of each method evaluated by correlation received from CESSM. Figures 3–5 give a more intuitive comparison. As shown in Table 2 and Figure 3, both WMM and simGIC outperformed the others in sequence similarity (with a correlation of approximately 0.8 in BP). By analyzing Figure 4, we can find that WMM, simGIC, and simUI show a higher correlation in Pfam similarity than the others (with a correlation of approximately 0.6 in MF). In Figure 5, the three aforementioned methods and Resnik show a similar correlation for ECC (0.6 in MF).

Pesquita et al. (2008) recommended a measurement called resolution instead of the correlation coefficient to evaluate how well the semantic similarity matches the sequence similarity because the relationship

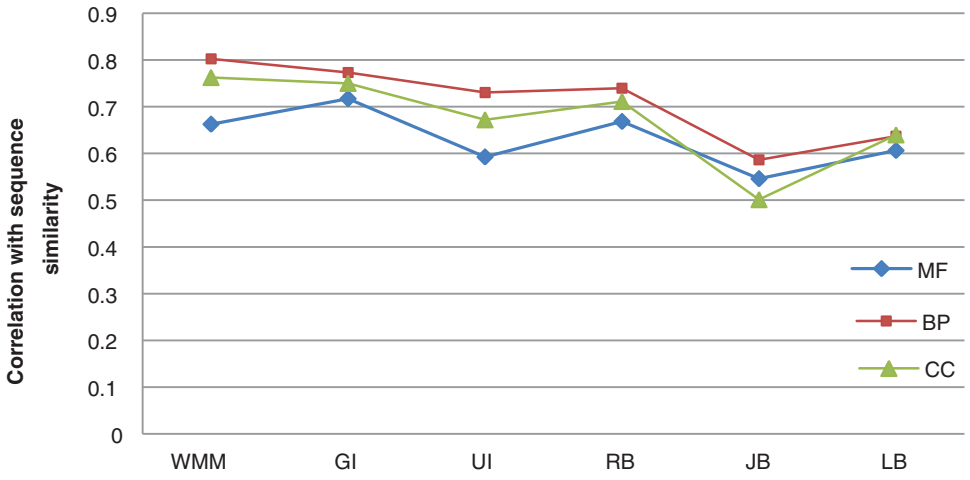

FIG. 3. Correlation between semantic similarity calculated by WMM and sequence similarity displayed by CESSM. Both WMM and simGIC outperform the others in sequence similarity (with a correlation of approximately 0.8 in BP). The evaluation is carried out for UniProt protein pairs from the CESSM database in the MF, BP, and CC ontologies. BP, biological processes; CC, cellular components; CESSM, Collaborative Evaluation of GO-based Semantic Similarity Measures; MF, molecular function; WMM, weighted multipath measurement.

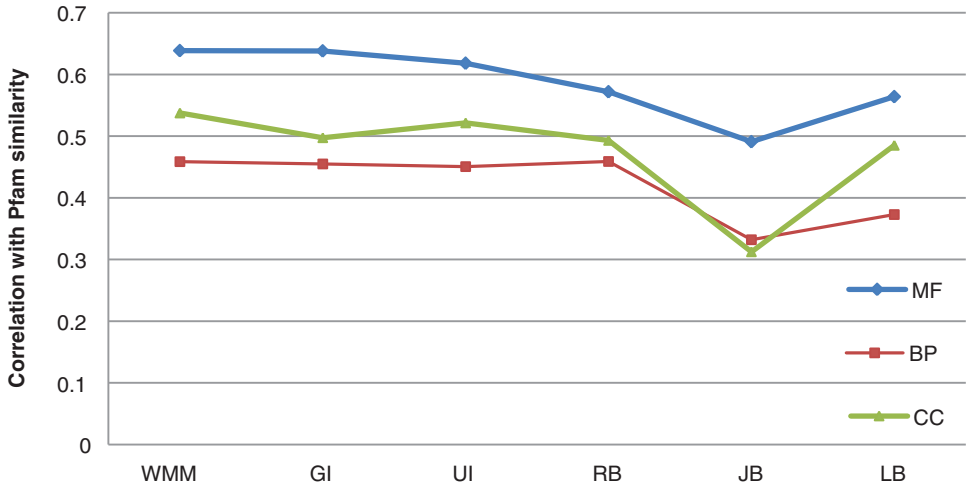

FIG. 4. Correlation between semantic similarity calculated by WMM and Pfam similarity displayed by CESSM. WMM, simGIC, and simUI show a higher correlation in Pfam (protein family) similarity than the others (with a correlation of approximately 0.6 in MF). The evaluation is carried out for UniProt protein pairs from the CESSM database in the MF, BP, and CC ontologies.

between semantic similarity and sequence similarity is not linear. Resolution is the relative intensity whereby variations in the sequence similarity scale are translated into the semantic similarity scale. The method with a higher resolution has a higher capability to distinguish protein functions between different levels. Figure 6 shows the resolutions when sequence similarity is compared with the semantic similarity measured by WMM and some other existing methods. For a more intuitive comparison, see Figure 7. It shows that WMM performs comparably to the other five methods.

3.3. Comparing identical annotations

Identical annotation occurs when two proteins are annotated with the same terms. It is a pity that most existing semantic similarity measurements assume that the similarity of any pair of proteins annotated by the same GO terms (known as identical annotation) will always be 1, which does not match the human perception that the similarity between proteins annotated with more specific terms should be greater than those annotated with more general ones. For instance, in the simplest case in Figure 1, in a given protein pair, P_1 and P_2 both annotated with the single term GO: 0005623, and in another protein pair, P_3 and P_4

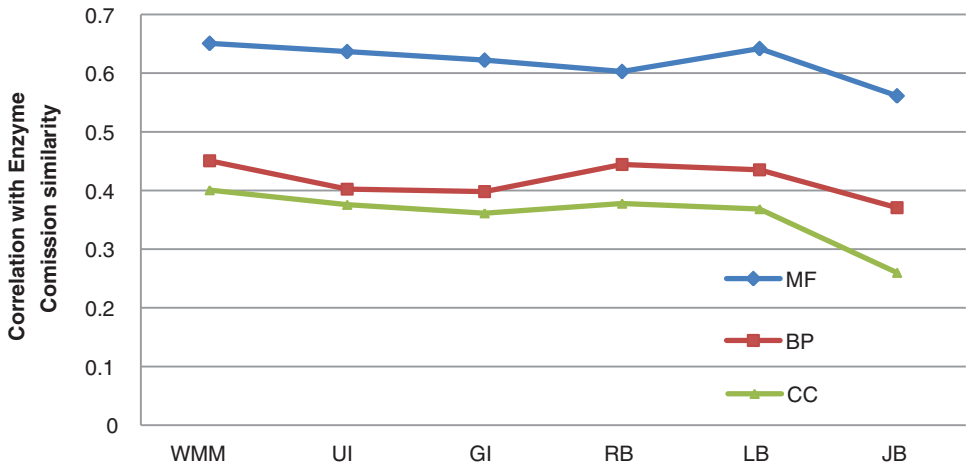

FIG. 5. Correlation between semantic similarity calculated by WMM and ECC similarity displayed by CESSM. WMM, simGIC, simUI, and Resnik show a similar correlation (0.6 in MF) for ECC.The evaluation is carried out for UniProt protein pairs from the CESSM database in the MF, BP, and CC ontologies. ECC, enzyme comission classification.

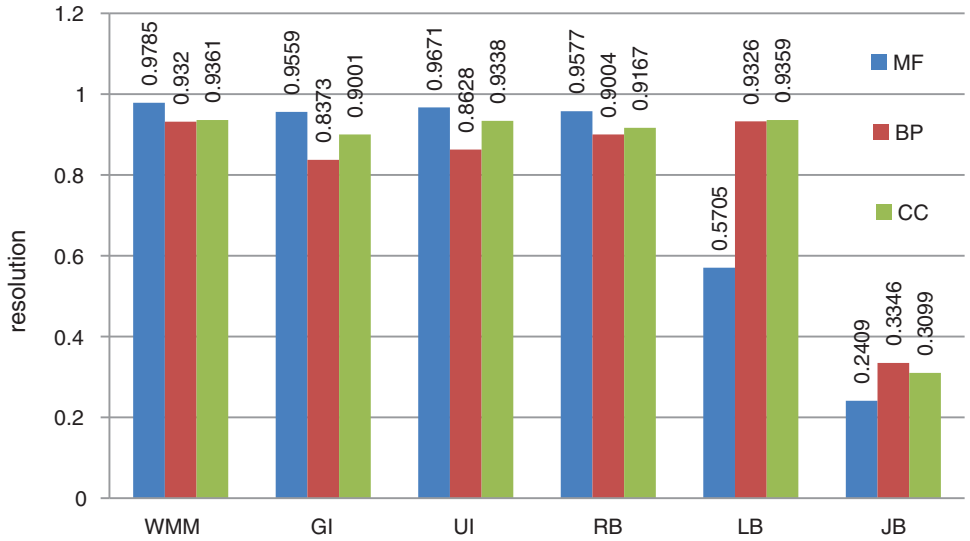

FIG. 6. The resolutions of different methods. Resolution is defined as the relative intensity whereby variations in the sequence similarity scale are translated into the semantic similarity scale.

both annotated with the single term GO: 0044424, and both of them have similarities of 1 under all of the existing measurements, but it is not true if rated by a human expert. As the specificity of terms increases downward through the DAG, it is reasonable to assume that there will be more similarity in the latter case than in the former case, and the similarity value between a term and itself can be calculated by Equation 9. As a result, our method shows that the similarity between P_1 and P_2 is 0.5583 and that the similarity between P_3 and P_4 is 0.7154.

As shown in Table 2 and Figure 7, WMM performs better than all the other five methods. Although it does not show significant improvements, it does show that it can compare proteins with identical annotations, providing a more authentic and unbiased result. These results confirm that WMM can be used as an alternative method to evaluate the functional similarity between proteins.

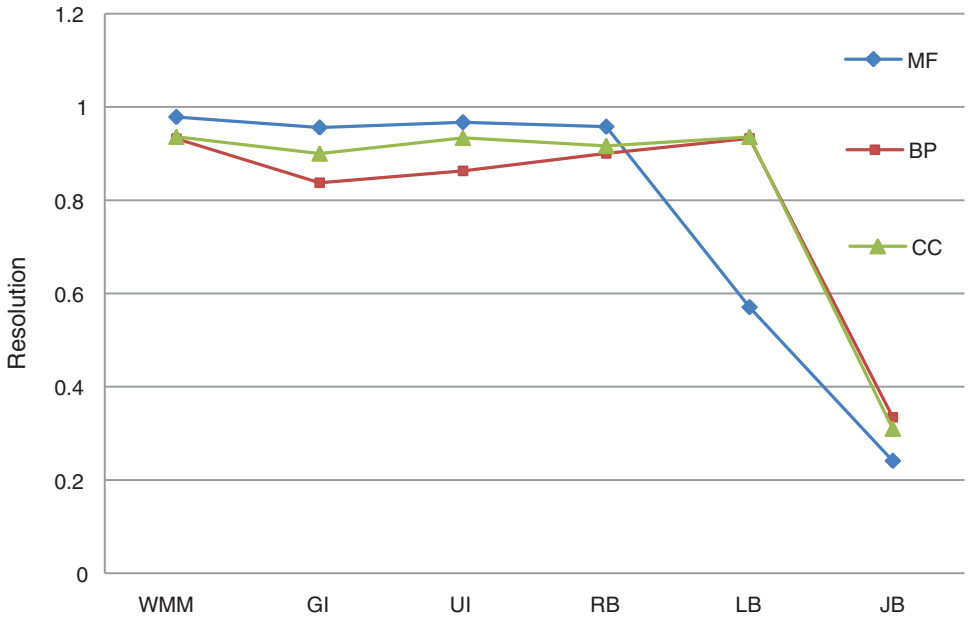

FIG. 7. The comparison of resolutions obtained by CESSM. This figure is drawn from the data in Figure 6 and shows that WMM outperforms the other five methods.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation of China under Grant No. 61303105; the Humanity & Social Science general project of Ministry of Education under Grant No. 14YJAZH046; the Beijing Educational Committee Science and Technology Development Planned under Grant No. KM201410028017; Academic Degree Graduate Courses group projects; and the Beijing Key Disciplines of Computer Application Technology.

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

Ashburner, M., Ball, C.A., Blake, J.A., et al. 2000. Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29.

Azuaje, F., Al-Shahrour, F., and Dopazo, J. 2006. Ontology-driven approaches to analyzing data in functional genomics. *Methods Mol. Biol.* 316, 67–86.

Barrell, D., Dimmer, E., Huntley, R.P., et al. 2009. The GOA database in 2009—an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.* 37, D396–D403.

Couto, F.M., Silva, M.J., and Coutinho, P.M. 2005. Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. *In* Proceedings of the 14th ACM International Conference on Information and Knowledge Management, ACM.

Couto, F.M., Silva, M.J., Lee, V., et al. 2006. GOAnnotator: linking protein GO annotations to evidence text. *J. Biomed. Discov. Collab.* 1, 19.

Fontana, P., Cestaro, A., Velasco, R., et al. 2009. Rapid annotation of anonymous sequences from genome projects using semantic similarities and a weighting scheme in gene ontology. *PLoS One* 4, e4619.

Gentleman, R. 2005. Visualizing and distances using GO. Available at: www.bioconductor.org/docs/vignettes.html

Guo, X., Liu, R., Shriver, C.D., et al. 2006. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics* 22, 967–973.

Guzzi, P.H., Mina, M., Guerra, C., and Cannataro, M. 2012. Semantic similarity analysis of protein data: assessment with biological features and issues. *Brief. Bioinform.* 13, 569–585.

Jiang, J.J., and Conrath, D.W. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008.

Lei, Z., and Dai, Y. 2006. Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. *BMC Bioinform.* 7, 491.

Lin, D. 1998. An information-theoretic definition of similarity. ICML.

McHale, M. 1998. A comparison of WordNet and Roget's taxonomy for measuring semantic similarity. *In* Proceedings of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems.

Mistry, M., and Pavlidis, P. 2008. Gene Ontology term overlap as a measure of gene functional similarity. *BMC Bioinform.* 9, 327.

Pesquita, C., Faria, D., Bastos, H., et al. 2007. Evaluating go-based semantic similarity measures. *In* Proceedings of 10th Annual Bio-Ontologies Meeting.

Pesquita, C., Faria, D., Bastos, H., et al. 2008. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinform.* 9, S4.

Pesquita, C., Faria, D., Falcao, A.O., et al. 2009a. Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.* 5, e1000443.

Pesquita, C., Pessoa, D., Faria, D., and Couto, F. 2009b. CESSM: collaborative evaluation of semantic similarity measures, JB2009: *Challenges Bioinform*. 157.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007.

Resnik, P. 2011. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.* 11, 95–130.

Rong, L., Shunliang, C., Yuanyuan, L., et al. 2006. A measure of semantic similarity between gene ontology terms based on semantic pathway covering. *Prog. Nat. Sci.* 16, 721–726.

Schlicker, A., Domingues, F.S., Rahnenführer, J., and Lengauer, T. 2006. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinform.* 7, 302.

Sevilla, J.L., Segura, V., Podhorski, A., et al. 2005. Correlation between gene expression and GO semantic similarity. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2, 330–338.

Teng, Z., Guo, M., Liu, X., et al. 2013. Measuring gene functional similarity based on group-wise comparison of GO terms. *Bioinformatics* 29, 1424–1432.

Wang, J.Z., Du, Z., Payattakool, R., et al. 2007. A new method to measure the semantic similarity of GO terms. *Bioinformatics* 23, 1274–1281.

Wu, C.H., Apweiler, R., Bairoch, A., et al. 2006a. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* 34, D187–D191.

Wu, X., Zhu, L., Guo, J. et al. 2006b. Prediction of yeast protein–protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Res.* 34, 2137–2150.

Xu, T., Du, L., and Zhou, Y. 2008. Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. *BMC Bioinform.* 9, 472.

Zhong, J., Zhu, H., Li, J., and Yu, Y. 2002. Conceptual graph matching for semantic search. *In* ICCS '02 Proceedings of the 10th International Conference on Conceptual Structures: Integration and Interfaces. Springer, London, pp. 92–106.

Address correspondence to:
Hanshi Wang
Information and Engineering College
Capital Normal University
No. 105. West 3rd Ring North Road
Haidian District
Beijing 100048
China

E-mail: wanghanshi1014@126.com