Computer Science and Information Systems 12(4):1235–1253 DOI: 10.2298/CSIS141130053Z

An Integrated Information-Based Similarity Measurement of Gene Ontology Terms

Shu-Bo Zhang[1] and Jian-Huang Lai[2]

[1] Department of Computer Science, Guangzhou Maritime Institute
510700 Guangzhou, P.R. China
845996912@qq.com
[2] School of Information Science and Technology,Sun Yat-sen University
510275 Guangzhou, P.R. China
stsljh@mail.sysu.edu.cn

Abstract. Measuring the semantic similarity between pairs of terms in Gene Ontology (GO) can help to compare genes that can not be compared by other computational methods. In this study, we proposed an integrated information-based similarity measurement (IISM) to calculate the semantic similarity between two GO terms by taking into account multiple common ancestors that they share, and aggregating the semantic information and depth information of the non-redundant common ancestors. Our method searches for non-redundant common ancestors in an effective way. Validation experiments were conducted on both gene expression dataset and pathway dataset, and the experimental results suggest the superiority of our method against some existing methods.

Keywords: Gene Ontology, GO terms, semantic similarity, biological pathways, gene expression profile.

1. Introduction

Measuring the similarity between pairs of genes or gene products is a fundamental and important research issue in molecular biology, as it can help to infer the biological function of genes, which has a vast variety of applications in the fields such as gene function prediction [1], gene expression data analysis [2], gene clustering [3], disease gene prioritization [4], analysis of protein interactions [5], and so on. As laboratory methods are costly, laborious and time-consuming, computational approaches to this issue are attracting more and more attention from the community of bioinformatics and biomedicine. With the advent of high-throughput techniques, a growing number of genes and gene products have been studied and annotated. Gene Ontology (GO) offers a consistent description of gene function from heterogeneous annotation data, and provides us with a promising way to compare genes or gene products that could not be compared by other computational approaches. The GO is composed of two components: GO graph and gene annotation, the GO graph is structured as a rooted Directed Acyclic Graph (rDAG) [6] with controlled vocabulary of terms as its nodes and the relationship between terms as its edges, while the gene annotation relates genes with terms in the GO graph. By studying the relationships between GO terms that annotate the genes involved, we can deduce the semantic similarity of genes at functional level.

The similarity between two genes or gene products can be quantified by computing the number of GO terms that annotate them simultaneously [7, 8], however, such approaches only take advantage of the annotation information, and are restricted to the genes or genes products with the same annotations [9]. Another class of approaches widely used is to consider both annotation and structure information of the GO graph, where the similarity between two genes is quantified by computing the semantic similarity between two terms or two sets of terms in the GO graph [10]. The semantic similarity measure between GO terms is close related to the structure of the GO graph, as the similarity score can be quantified by deriving both similarity and dissimilarity from their common ancestors and distance of path connecting them, respectively. We shall give a brief discussion of some most representative methods in the next section, the reader can refer to Pesquita et al. [10] and Gan et al. [11] for details about different semantic similarity measures and a comprehensive survey of literature. To date, many measures have been introduced to quantify the semantic similarity of GO terms, and some of them have been proved to be useful in relevant fields [12, 13]. However, due to the complexity of the GO structure, measuring the semantic similarity between two terms is still challenge problem.

In this study, we proposed an integrated information-based similarity measurement (IISM) to measure the semantic similarity of GO terms. This method is based on the following observations: (1) A term in the GO graph inherits semantics from its ancestors; (2) some common ancestors of two GO terms provide redundant semantics for quantifying their similarity; (3) the similarity between terms near the root of GO graph is smaller than that of terms further from the root. The first two observations come from the semantics inheritance property between terms in the GO graph, and the third observation dovetails with the human intuition of biological knowledge that a GO term at lower levels has more specific biological meaning. Our method quantified the similarity measure between pairs of GO terms by combining the node property and the depth of nodes in the Go graph. We first detected the non-redundant ancestor of two terms and their non-redundant common ancestors. After computing the semantic value of the non-redundant ancestors of both terms and those of their non-redundant common ancestors, the similarity value based on the node property was quantified by the ratio of semantics two terms share. The similarity measure based on the depth of nodes was quantified based on the probabilities they occurred in the corpus, and then the average of these two measures was defined as the similarity values of the term pair. The effectiveness of our similarity measurement was evaluated through the study of similarity among genes in the pathway of Saccharomyces genome database (SGD) database, another experiment was conducted on a gene expression dataset, and the results suggest the superior performance of our method.

The rest of this paper is organized as follows: after reviewing some related works in section 2, we describe our measure for quantifying semantic similarity over GO terms in section 3. Section 4 demonstrates some experimental results to evaluate the effectiveness of the proposed measure. Finally, this paper is finished with a conclusion in Section 5.

2. Related Work

The methods to compute the semantic similarity between two GO terms can be generally classified into three categories: edge-based, node-based and hybrid methods, while the semantic similarity between two sets of GO terms can be computed in pairs or in groups.

The edge-based approaches measure the similarity between two terms based on the number of edges and their type in the GO graph. A commonly adopted strategy is to derive the similarity measure from the distances between two nodes in the graph, where the distance is quantified by the length of the shortest path, or average length of the all paths connecting two nodes, and the distance measure can be easily converted into a similarity value. Rada et al. [14] were the earlier researchers that used this measure on a biomedical ontology MSH (i.e. Medical Subject Headings), and Al-Mubaid et al. [15] used it on GO terms for the first time. Alternatively, the similarity measure of this category can be directly quantified by the number of edges in the path from the root node to the most informative common ancestor of two nodes [16]. However, Pesquita et al. [10] suspected the effectiveness of the edge-based measures as they are established on the basis of two assumptions that are seldom true in biological ontologies, that is to say, the nodes and edges in the graph are evenly distributed, and the edges at the same level have the same semantic distance from the root. Even though several attempts have been made to lessen these disadvantages by weighting the edges at different depth, considering the link type or the density of nodes [17, 18], the problem caused by these assumptions are still not solved effectively.

Node-based methods used the nodes and their properties as data source to deduce the semantic similarity between two terms. The nodes used in this categories of method includes those relate to the term pair investigated and their ancestors or descendants in the GO graph [10]. The most commonly used property of a node is the information content (IC), which indicates how informative and specific a term is, and is defined as the negative logarithm of the probability of a term occurs in a corpus [19] or in the GO graph [20, 21]. Resnik [22] proposed a similarity measure based on the shared information, which is deduced from the information content of their most informative common ancestor (MICA). Since the similarity value of Resnik's measurement may be greater than one, Lin [23] and Jiang and Conrath [24] proposed their improved versions of measure that normalized the similarity value to (0,1). Nevertheless, these measurements defined the similarity based on Resnik's measurement that only consider the information content of a single common ancestor, namely, the MICA that inherited by both terms. This is proper in the case that the GO is a tree, but it will become problematic in the Directed Acyclic Graph (DAG) structure of GO, as a node may have more than one parent nodes and thus some biological information inherited from some ancestors will be neglected. Some measures were proposed to address this problem by considering the effect of multiple ancestors [25, 26] or multiple descendants [27, 28] of two terms investigated, In [25], Couto et al. employed the concept of disjunctive common ancestors (DCAs) and proposed a graph-based similarity measure (GraSM), where the disjunctive common ancestors is determined in a recursive way, and they redefined the shared information content as the average information content of all their disjunctive common ancestors. They later updated GraSM with dubbed Disjunctive Shared Information (DiShIn) [26] to calculate the shared information content between two terms by counting the number of distinct paths from common ancestors to the terms, even so, the computational complexity of GraSM and DiShIn are rather high, as they need to search for the paths between two terms. In real-time scenario, they need to perform a preliminary calculation and stored the results for later application.

As mentioned above, both the edge-based and the node-based measures have their limitations. Some works take into account both nodes and edges in the graph into account. Wang et al. [18] presented a similarity measurement by combining the structure of GO graph and the semantic information of the GO terms, where they quantified the semantics of a term by S-value, which integrated the contribution of all term in a GO subgraph including all the ancestors and the term itself. Their similarity measure between two terms was defined as the percentage of S-value they share. Several works [29-31] have demonstrated the advantage of this measurement. However, this approach suffers from some shortcomings [32], namely that the semantic contribution value of a edge is empirically determined, and that the dynamically calculation for the semantic values of GO terms is rather time consuming. Recently, Wu et al. [9] proposed a hybrid measure, where they used the node information to improve the edge-based measure they introduced previously, and shown the superiority in determining the protein-protein interaction. Bandyopadhyay and Mallick [33] developed a new hybrid method to address the issue of shallow annotation in the GO structure. Song et al. [32] introduced an aggregate information content approach where they defined the semantics of a term as the aggregate contribution of semantic weight of all its ancestors and the term itself and the similarity between two terms was defined as the ratio of semantics they shared.

3. The Proposed Measure

3.1. Non-redundant Common Ancestors of Two GO Terms

The GO graph is structured as a rooted Directed Acyclic Graph (rDAG) [6] with controlled vocabulary of terms as its nodes and the relationship between terms as its edges. The terms in the graph describe genes and gene products with three aspects of biological meanings: molecular function (MF), biological process (BP) and cellular component (CC). The edges in the graph link different terms to each other by certain relationships, such as “part-of”, “is-a”, “regularized”, and so on. In the DAG structure of GO graph, all terms in the graph are organized in a hierarchical way, a term closer to the root of the graph has more general biological meaning, while those far from the root have more specific biological meanings, that is to say, a term at lower level inherits the semantics from its parents, which are more specific in biological meanings. The biological meanings of nodes in the graph are more and more specific from ancestral terms to descendant terms, and the semantics of GO terms also become more and more specific.

Due to the inheritance relationship between terms at different levels of the GO graph, the semantics conveys from upper terms to their descendants step by step through the edges linking them, and the child nodes aggregate all semantics from its direct parents and form a more specific semantics, which will be transmitted downwards. Based on this observation, one may think that the semantics of a node's ancestral terms are redundant to all its descendants. This is the case when there is a unique direct parent of those child nodes. However, as a node in the GO graph may have multiple parents, and a parental node may have more than one child nodes, the semantics of an ancestral node may be specified through a path and flows into some descendants, but not flows into other child nodes. For the similarity measure between pairs of GO terms, this kind of ancestors is informative and is referred to as non-redundant common ancestors. How to identify such common ancestors is a focus of this study.

In order to distinguish between redundant and non-redundant common ancestors of two GO terms investigated, we examine the direct child nodes of each common ancestor. If they are all inherited by the both terms, we can see that from these direct child nodes upwards, the two terms have the same common ancestors, and the semantics of the common ancestors are inherited by both terms through the paths passing these direct child nodes. Thus, the common ancestors upwards from the direct child nodes are redundant as it provides no more information about the relationship between the term pair. On the contrary, if there is a direct child node inherited by either term involved, the semantics of the common ancestor may flow into this child node and downward to the term exclusively. This suggests that the common ancestor is informative to the semantic relationship between two terms, and the common ancestor of this kind is distinguished as non-redundant common ancestor, they should be taken into consideration when we define the similarity measurement between two terms.

The non-redundant common ancestors shared by two terms can be identified in a direct way. First of all, we detect the parental node set for each term involved, and then compute the common ancestor set from the two parental node sets. Consequently, for each element in the common ancestor set, we check its direct child nodes, if there is a child node that are not included in the common ancestor set, we consider it to be a non-redundant common ancestor. This algorithm has a computational complexity of $O(n)$, which is much cheaper than other methods [18, 26] based on multiple common ancestors.

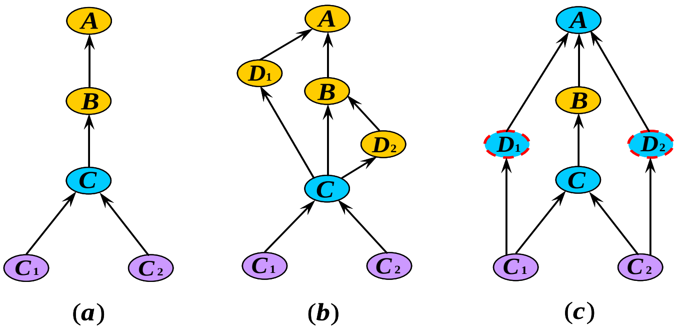

Fig. 1. Illustrative examples of the inheritance relationship and the non-redundant common ancestors of two terms in the GO graph. Ancestral nodes in yellow color are redundant ancestors and those in blue color are non-redundant ancestors of C_1 and C_2. D_1 and D_2 in (c) are the non-redundant ancestor of C_1 and C_2, respectively

Figure 1 gives an illustrative example to show the inheritance relationship among nodes in the GO graph and the non-redundant ancestors of two terms, as well as their non-redundant common ancestors. The graph in Figure 1(a) is a tree, in which A and B the redundant ancestors of C_1 and C_2 as the semantics of A is inherited by B and then uniquely inherited by C, which is directly inherited by C_1 and C_2, and thus C is their unique non-redundant common ancestor. Similarly, C in Figure 1(b) is also a unique non-redundant common ancestor of C_1 and C_2 as all semantics of their ancestors are combined in C and directly inherited by the two terms. Figure 1(c) gives another example of multiple inheritance. Unlike D_1 and D_2 in Figure 1(b), these two nodes are uniquely inherited by C_1 and C_2, and they serves as non-redundant ancestor of this two nodes respectively, C is the direct parent of the term pair and serves as a non-redundant common ancestor. The semantics of B is a redundant common ancestor as it contributes all its semantics to node C. As for A in this graph, it is a non-redundant common ancestor of C_1 and C_2, as the semantics of this node flows into the two leaves nodes through different paths, even though it is not directly inherited by the two terms.

3.2. The Semantic Similarity between GO Terms

As discussed in the previous section, many existing measures have their limitations. In order to address these problems, we proposed a hybrid scheme to quantify the similarity between pairs of terms in the GO graph, where both structural information and information content of nodes are taken into consideration. Moreover, we took advantage of more than one common ancestor of two terms. The similarity value composes of two parts: one is deduced from the semantic values of both non-redundant common ancestors and the non-redundant ancestors of the term pair involved, the other is deduced from the depth of their non-redundant common ancestors, which is based on the probability that these non-redundant common ancestor occur in the corpus.

Suppose we have two terms c_1 and c_2 ,and $RCA(c_1, c_2)$ is their redundant common ancestor set, then the commonality of the term pair can be characterized by their non-redundant common ancestor set, which can be calculated as,

$$NRCA(c_1, c_2) = CA(c_1, c_2) - RCA(c_1, c_2) \tag{1}$$

where $CA(c_1, c_2)$ is the common ancestor set of c_1 and c_2. Consequently, the semantic value that they share can be defined as the summation of information content that all their non-redundant ancestors contain, and computed as follows,

$$SSV(c_1, c_2) = \sum_{t \in NRCA(c_1, c_2)} IC(t) \tag{2}$$

The non-redundant ancestor set of term c_1 can then be defined as,

$$NRA(c_1) = Parent(c_1) - RCA(c_1, c_2) \tag{3}$$

where $Parent(c_1)$ is the ancestor set of c_1. Then the semantics of this term can be characterized by the elements in $NRA(c_1)$ and the term itself. In this study, we define the semantic value of a GO term by adding the information content of this term and those of its non-redundant ancestors,

$$SV(c_1) = \sum_{t \in NRS(c_1)} IC(t) \tag{4}$$

where $NRS(c_1) = NRA(c_1) \text{U}\{c_1\}$. Note that $SV(c_1)$ provides the information to describe what a term c_1 is, and $SSV(c_1, c_2)$ provides all information needed to state the commonality of c_1 and c_2 in the GO graph, we can quantified their similarity measure, based on the information they share, in a normalized form by the ratio between the amount of $SSV(c_1, c_2)$ and the semantic values of c_1 and c_2 as follows,

$$sim_{sv}(c_1, c_2) = \frac{2 \times SSV(c_1, c_2)}{SV(c_1) + SV(c_2)} \tag{5}$$

According to the definition in formulas (1) and (3), we have $NRCA(c_1, c_2) \subseteq NRA(c_1)$ and $NRCA(c_1, c_2) \subseteq NRA(c_2)$, which will result in $SSV(c_1, c_2) \leq SV(c_1)$ and $SSV(c_1, c_2) \leq SV(c_2)$ and then $sim_{sv}(c_1, c_2) \leq 1$.

In the case where c_1 and c_2 are identical, we have $SV(c_1) = SV(c_2)$ and $SSV(c_1, c_2) = SV(c_2)$, and the similarity value of this kind will be equal to 1. This means that the specificities of terms in the GO graph are ignored and it will cause the problem of shallow annotation, which has been highlighted in previous work. In order to address this limitation, we take advantage of the information related with the depth of non-redundant common ancestors in the GO graph.

In the hierarchical structure of GO graph, if a gene is annotated with a term, it is also annotated with the ancestors of this term. That is to say, once a term occurs in a corpus, all its ancestors will also occur in the same corpus implicitly, which will result in smaller probability of occurrence for the terms at lower level. As we have discussed in previously that GO term pairs at lower level are more specific and have larger similarity values, we can relate the probability that the non-redundant common ancestors occur with the similarity measurement of two terms, and define this kind of similarity value as follows,

$$sim_{depth}(c_1, c_2) = \frac{1}{N} \sum_{t \in NRCA(c_1, c_2)} (1 - p(t)) \tag{6}$$

where N is the number of non-redundant common ancestors of c_1 and c_2, $p(t)$ is the probability that a term t occurs in the corpus. It is easy to see that if the non-redundant common ancestors are far away from the root, the similarity value of this kind will become larger.

Finally, we define the overall semantic similarity value between c_1 and c_2 by integrating the above two similarity measures as below:

$$sim_{IISM}(c_1, c_2) = (sim_{depth}(c_1, c_2) + sim_{sv}(c_1, c_2)) / 2 \tag{7}$$

To show how the non-redundant ancestors are identified and the semantic similarity measurement IISM is calculated, we take a snippet of GO graph for example. Figure 2 is a sub-graph of biological process aspect of GO that consists of 7 terms and the corresponding relationships among them. In this context, *cellular protein localization* (denoted by n_5) and *establishment of protein localization* (denoted by n_6) are two terms investigated, and the set of their common ancestors is $\{n_0, n_1, n_2, n_4\}$, n_0 and n_1 are redundant common ancestors as all their semantics flow into n_2 and then into n_5 and n_6 through different paths, so the non-redundant common ancestors set of n_5 and n_6 is $NRCA(n_5, n_6) = \{n_2, n_4\}$. After removing the redundant common ancestors away from the corresponding ancestor set of n_5 and n_6, we get their non-redundant ancestor set, $NRA(n_5) = \{n_2, n_3, n_4\}$ and $NRA(n_6) = \{n_2, n_4\}$, respectively.

Table 1. The information content value, semantic value and probability of the terms presented in Figure 2

Id	GO id	IC value	S value	P value
n_0	GO:0008150	0	0	1
n_1	GO:0051179	0.7508	0.7508	0.1775
n_2	GO:0033036	1.2365	1.9873	0.0580
n_3	GO:0070727	1.5644	3.5517	0.0273
n_4	GO:0008104	1.3416	3.3289	0.0455
n_5	GO:0034613	1.5819	10.4498	0.0262
n_6	GO:0045184	1.4176	6.7338	0.0382

IC value denotes the information content of each node and S value is the semantic value of this term and its non-redundant ancestors, P value denotes the probability that each node occurs in the occurs in a corpus

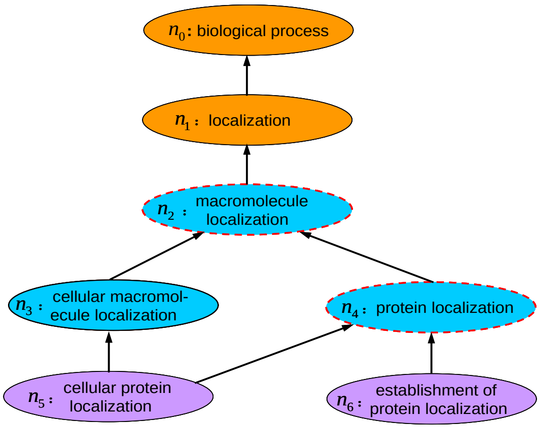
n_0: biological process
n_1 : localization
n_2 : macromolecule localization
n_3 : cellular macromol-ecule localization
n_4 : protein localization
n_5 : cellular protein localization
n_6 : establishment of protein localization

Fig. 2. A fragment of GO graph that contains cellular protein localization and establishment of protein localization. The nodes in blue color with red dashed circles (n_2 and n_4) are the non-redundant common ancestors of n_5 and n_6 , while those in orange color are their redundant common ancestors

3.3. The Implementation of IISM

The algorithm for the implementation of IISM is described below. It starts by selecting the parents and the common ancestors of the both terms (line 1) in GO and by initializing the set of redundant common ancestors as an empty set (line 2). The algorithm selects each common ancestor (line 3). For each selected ancestor, the algorithm calculates its direct descendants and checks if they are all inherited by both terms (line 4-5), if the common ancestor is redundant, it is added to the set of redundant common ancestors (line 6). Then the non-redundant common ancestor set and the non-redundant ancestor sets of the both terms are calculated (line 9-11), and the common semantics as well as the semantics of each term is computed (line 12 -22), the similarity measure based on the semantic values is quantified as the ratio of common semantics to the total semantics of two terms (lines 22). After the similarity based on the depth of the non-redundant common ancestors was computed (line 23 - 27), the algorithm calculates the similarity as the average of the two kinds of similarity values (lines 28). The algorithm for the implementation of IISM is shown in Figure 3.

Algorithm 1. *IISM* (c_1, c_2)

```
1:  CA = GetCommonAnc($c_1$, $c_2$), $P_1$=GetParent($c_1$), $P_2$=GetParent($c_2$)
2:  $RCA$ = {}
3:  for all $a$ in $CA$ do
4:      DirectChildSet = GetDirectDescendant(a)
5:      if DirectChildSet $\subset$ CA then
6:          $RCA = RCA \cup \{a\}$
7:      end if
8:  end for
9:  NRCA=CA-RCA,
10: $NRA_1$=$P_1$ - RCA
11: $NRA_2$=$P_2$ - RCA
12: SSV = 0, $SV_1$ = 0,$SV_2$ = 0
13: for all $ca$ in $NRCA$ do
14:     SSV +=IC(ca)
15: end for
16: for all $c_1$ in $NRA_1$ do
17:     $SV_1$ +=IC($c_1$)
18: end for
19: for all $c_2$ in $NRA_2$ do
20:     $SV_2$ +=IC($c_2$)
21: end for
22: $sim_{sv}$= 2×SSV/($SV_1$+$SV_2$)
23: $P = 0$
24: for all $ca$ in $NRCA$ do
25:     $P$ += $p(ca)$
26: end for
27: $sim_{depth}$= P/N
28: return sim= ($sim_{sv}$+ $sim_{depth}$)/2
```

Fig. 3. The algorithm for the implementation of our approach

4. Validation of the Proposed Approach

In this section, the performance of our similarity measure over terms is validated through the effectiveness that they quantified the functional similarity between pairs of genes or gene products. As a gene may be annotated with more than one term, the gene similarity is usually established on two terms sets. The similarity value between two term sets can be computed by three schemes: maximum (MAX) [18], average (AVE) [34] and best match average (BMA) [35] rule. In this study, the BMA rule was adopted to estimate the semantic similarity value between two genes.

There are some strategies to validate the effectiveness of similarity measure for genes or gene products: comparing the semantic similarity with the gene expression similarity, comparing it with sequence similarity, comparing it with interaction relationship of proteins, or comparing it with the gene relationships in biological pathways. In this study, we used gene expression dataset and pathway dataset as bench mark to validate the proposed measure, the semantic similarity was compared with expression similarity and gene functional similarity in pathway. Our experimental results were evaluated against those produced by some existing measures.

4.1. Dataset

Three datasets were used in this study: the GO dataset (released in April 2013), a gene expression dataset and a pathway dataset were used. The Gene Ontology data and gene annotation dataset were downloaded from the Gene Ontology database [36], which contains 25370 BP, 3295 CC, and 10445 MF terms. The gene annotation dataset contains 91133 annotations of 6381 genes for the yeast genome, and 358244 annotations of 43245 genes. The Spellman dataset [37] comprises of 6178 gene expression profiles of genes in yeast cell cycle.

The pathway dataset was downloaded from the website of SGD (http://pathway.yeastgenome.org/), which contains classification and annotation information of genes in each pathway. There are 187 biological pathways in the SGD database (as of September 23, 2013). Most of these pathways contain more than three genes manually annotated by both Enzyme Commission (EC) numbers and molecular function GO terms. For instance, there are eleven genes, BAT2, BAT1, PDC6, PDC5, PDC1, SFA1, ADH5, ADH4, ADH3, ADH2 and ADH1 in the amino acid degradation pathway valine degradation. According to SGD, the eleven genes in this pathway are manually divided into three classes as illustrated in Figure 4. The genes in each group participate in the same biological process and are annotated by an EC number. Table 2 lists the EC number of all genes in this pathway. Moreover, the genes in the same group are annotated by mostly the same GO terms. In this study, the entities lacking EC number or gene name, as well as pathway with less three genes, were removed from SGD. The final dataset contains 109 pathways with at least three genes annotated by EC numbers and GO terms.

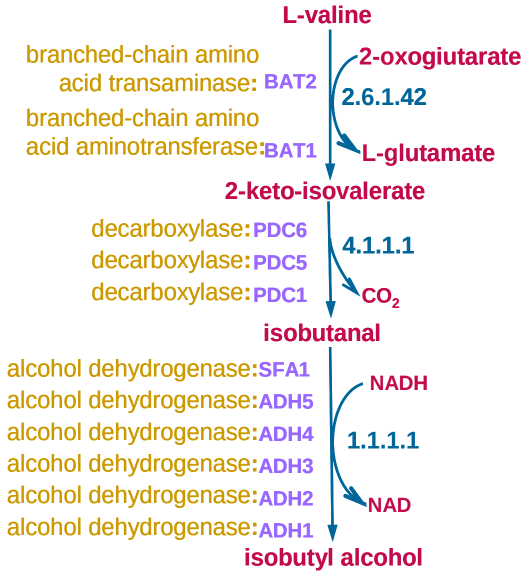
L-valine
branched-chain amino acid transaminase: BAT2
branched-chain amino acid aminotransferase:BAT1
2-oxogiutarate
2.6.1.42
L-glutamate
2-keto-isovalerate
decarboxylase:PDC6
decarboxylase:PDC5
decarboxylase:PDC1
4.1.1.1
CO_2
isobutanal
alcohol dehydrogenase:SFA1
alcohol dehydrogenase:ADH5
alcohol dehydrogenase:ADH4
alcohol dehydrogenase:ADH3
alcohol dehydrogenase:ADH2
alcohol dehydrogenase:ADH1
NADH
1.1.1.1
NAD
isobutyl alcohol

Fig. 4. Functions of genes in a S.cerevisiae pathway valine degradation

4.2. Experimental Results

Results on pathway dataset. In the SGD dataset, genes in each pathway are manually classified according to their molecular function, and annotated with molecular function terms of GO and the corresponding EC numbers. Two genes annotated with the same EC number means that they participate in the same reaction and tend to perform similar biological function. This kind of priori knowledge provides similarity information between genes at functional level, i.e., the genes with the same EC number is more similar than those with different EC numbers. We took this priori functional information as ground truth to validate our approach by comparing it to the clustering results derived from our approach. The genes in a pathway are clustered according to the semantic similarity measurement proposed in this paper and then compared with the ground truth. If the clustering result is consistent with the artificial classification result based on the biological reactions, it suggests that the similarity measure is effective in charactering the functional similarity between genes.

Because of the space crunch, we only take the pathway 'valine degradation' as an example to demonstrate the performance of our approach and that of Wang's method [18]. The gene names and corresponding EC numbers in this pathway are listed in Table 2. We see that there are 11 genes involved in 3 reactions in valine degradation pathway, SFA1, ADH1, ADH2, ADH3, ADH4 and ADH5 are in a reaction annotated with EC number '1.1.1.1', BAT1 and BAT2 are in another reaction with EC number '2.6.1.42', while PDC1, PDC5 and PDC6 participate in the reaction numbered with '4.1.1.1'.

Table 2. Functions of genes in valine degradation pathway

Class id	EC number	Gene name	Class id	EC number	Gene name
C1	1.1.1.1	ADH3	C2	4.1.1.1	PDC1
	1.1.1.1	ADH4		4.1.1.1	PDC5
	1.1.1.1	SFA1		4.1.1.1	PDC6
	1.1.1.1	ADH5	C3	2.6.1.42	BAT1
	1.1.1.1	ADH1		2.6.1.42	BAT2
	1.1.1.1	ADH2			

Table 3. Similarity values among genes in valine degradation pathway obtained by different methods

	ADH3	ADH4	ADH5	ADH2	ADH1	SFA1	PDC1	PDC5	PDC6	BAT1	BAT2
ADH3	1.000	0.896	0.851	0.897	0.837	0.915	0.577	0.530	0.647	0.625	0.648
ADH4	0.917	1.000	0.820	0.812	0.824	0.838	0.581	0.557	0.667	0.648	0.669
ADH5	0.886	0.867	1.000	0.850	0.788	0.867	0.551	0.556	0.677	0.686	0.673
ADH2	0.823	0.852	0.884	1.000	0.892	0.911	0.527	0.573	0.725	0.665	0.722
ADH1	0.879	0.866	0.842	0.924	1.000	0.881	0.513	0.550	0.699	0.652	0.682
SFA1	0.934	0.872	0.804	0.929	0.913	1.000	0.513	0.541	0.666	0.700	0.662
PDC1	0.660	0.673	0.640	0.463	0.615	0.607	1.000	0.963	0.628	0.595	0.628
PDC5	0.629	0.658	0.646	0.657	0.646	0.632	0.975	1.000	0.670	0.621	0.666
PDC6	0.736	0.749	0.754	0.794	0.774	0.746	0.703	0.731	1.000	0.912	0.994
BAT1	0.717	0.736	0.764	0.747	0.736	0.568	0.661	0.683	0.928	1.000	0.915
BAT2	0.735	0.750	0.750	0.790	0.763	0.740	0.701	0.727	0.997	0.468	1.000

The values in the upper triangular matrix are obtained by our method, while those in the lower triangular matrix are obtained by Wang's method

The semantic similarity values among genes in this pathway were computed by our measure and Wang's, and the genes were clustered based on the semantic similarity values. Table 3 lists the similarity values obtained by our method (the upper triangular matrix) and those obtained by Wang's measure (the lower triangular matrix). It shows that the two measures generally product higher similarity values within the same cluster. They both mostly produce semantic similarity values lager than 0.8 for the two reaction annotated with "1.1.1.1"and "4.1.1.1", respectively. However, Wang's method produced a rather small similarity value of 0.468 between BAT1 and BAT2, which is only about half of our similarity value. Figure 5 demonstrates the comparison results of average semantic similarity values of inner and between groups. We can see that the inner similarity scores are generally larger than the similarity between groups except one (produced by Wang's method). In comparison, the inner group similarities of C1-C1 and C2-C2 produced by Wang's method are somewhat larger than those produced by IISM, while those for the third group C3-C3 is much smaller than IISM. As for similarity values between groups, IISM consistently has larger scores, which suggests that our method has stronger ability to distinguish between groups than those of Wang's.

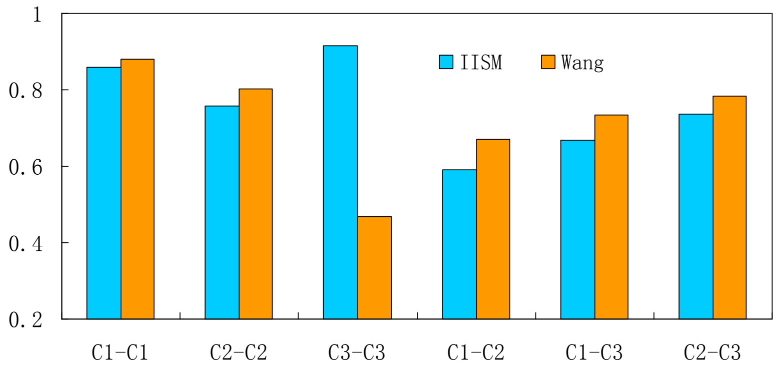

Fig. 5. Comparison results of inner and between groups similarity values for IISM and Wang's method. Larger inner similarity score and smaller inter similarity values implies stronger ability to distinguish different classes

The clustering results based on the semantic similarity values obtained by our method and those obtained by Wang's method are shown in Figure 6 and Figure 7, respectively. Figure 6 shows that the 11 genes in the valine degradation pathway are clustered into three clusters based on our measure. SFA1, ADH1, ADH2, ADH3, ADH4 and ADH5 form the first class (denoted by class 1), while PDC1, PDC5 and PDC6 are clustered into another group (denoted by class 2) and BAT1 and BAT2 form the third cluster (denoted by class 3). The clustering results suggest that our semantic similarity measure can effectively capture the functional relationships among genes in the pathway.

The clustering results based on Wang's measure are demonstrated in Figure 7. We see that the eleven genes are not clear grouped in to three classes, as the second class (PDC1, PDC5 and PDC6) and the third class (BAT1 and BAT2) do not seem to form clear groups. Specifically, PDC6 seems more close to the first group of genes, while PDC1 and PDC5 seem more close to the third class. In addition, the genes in the third class are clustered into proper group either. Thus, the clustering result is not consistent with our prior knowledge about the valine degradation pathway. It implies that the similarity scores obtained by Wang's approach can't effectively characterize the functional relationship between genes in this pathway.

Here we give a biological interpretation of the clustering result of our method in this pathway to show how it is consistent with the priori functional knowledge of these genes. According to the annotation information in the SGD database, the eleven genes in this pathway are functionally related with the valine degradation activity, which is comprised of the following steps: 1) deamination of the amino acid to the corresponding alpha-keto acid; 2) decarboxylation of the resulting alpha-keto acid to the respective aldehyde; and, 3) reduction of the aldehyde to form the corresponding long chain or complex alcohol. Specifically, the genes in class 1 (SFA1, ADH1, ADH2, ADH3, ADH4 and ADH5) tend to be involved in alcohol dehydrogenase activity (GO:0004022) and participate in amino acid catabolic process to alcohol via Ehrlich pathway (GO:0000947) and NADH oxidation (GO:0006116) in cytoplasm (GO:0005737). By contrast, PDC1, PDC5 and PDC6 are inclined to perform pyruvate decarboxy-lase activity (GO:0004737) in nucleus (GO:0005634) or cytoplasm (GO:0005737), and be involved in catabolic and metabolic process, as they are mostly annotated with glycolytic fermentation to ethanol (GO:0019655), aromatic amino acid family catabolic process (GO:0000949), tryptophan catabolic process (GO:0006569), L-phenylalanine catabolic process (GO:0006559) and pyruvate metabolic process (GO:0006090). As for the third group of genes, BAT1 and BAT2 tend to be involved in branched-chain-amino-acid transaminase activity (GO:0004084) and participate in the branched-chain amino acid biosynthetic process (GO:0009082) and branched-chain amino acid catabolic process (GO:0009083). In terms of the relationship among the three groups, class 2 and class 3 are more similar to each other since the genes in these two groups are all annotated with ‘catalytic activity’ in the SGD database. This result suggests that our clustering result is consistent with the human perspective of gene functions in this pathway. Additionally, the genes in class 1 also participate in some other degadation pathway, such as tryptophan degradation, leucine degradation, phenylalanine degradation and isoleucine degradation, and those in class 2 (PDC1, PDC5 and PDC6) also be involved in the pathways of glucose fermentation, phenylalanine degradation, tryptophan degradation, isoleucine degradation and acetoin biosynthesis II, while the two genes in the third group tend to be active in the biosynthesis pathways like isoleucine biosynthesis, leucine biosynthesis and valine biosynthesis. Once again, we see the superior of our method, as our semantic similarity scores can clearly distinguish the genes between class 2 and class 3, which are similar in biological meanings.

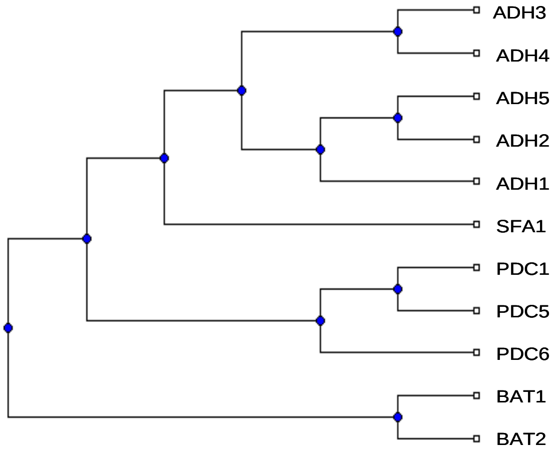

Fig. 6. Clustering results of genes in valine degradation pathway based on similarity values obtained by our method

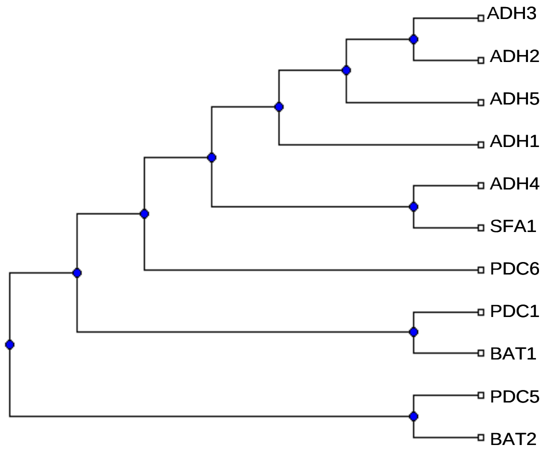

Fig. 7. Clustering results of genes in valine degradation pathway based on similarity values obtained by Wang's method

Results on gene expression dataset. In the experiment conducted on the gene expression dataset, the performance of our method is evaluated by comparing the semantic similarity values with the gene expression similarity. The hypothesis for this is that genes express at similar level tend to perform similar biological function or involved in the same biological process. In this study, the relationship between the expression similarity and semantic similarity was characterized by the Pearson's correlation coefficient. The expression similarity relationship between genes is quantified by the absolute correlation coefficients between the expression profiles, and the semantic similarity values of genes were deduced from the biological process (BP) ontology. Like several previous studies [32, 38, 39], the gene pairs are divided into equal intervals based on the sorting expression correlation coefficients between pairs of genes. The average expression correlation coefficients of genes within each interval characterize the mean statistical property of expression correlations. The corresponding average semantic similarity values based on GO terms in each interval were computed and then the Pearson's correlation coefficients between the averages of the two kinds of similarities in each interval are calculated. The larger coefficient suggests that the stronger association between the semantic similarity and the expression similarity, and the more effective semantic similarity measure between GO terms.

To evaluate the effectiveness of our metric against other methods on this gene expression profile dataset, we split the gene pairs into 4-13 bins like [32], the Pearson's correlation coefficients based on different methods were computed and listed in Table 4. We can see that the scores of correlation coefficient generally decrease as the number of groups grows. This means that the data more close to the 'raw' data will produce smaller coefficients, which will make it difficult to capture the relationship between the expression similarity and semantic similarity, and subsequently evaluate the effectiveness of the similarity measure. As for different measures, the similarity scores produced by multiple common ancestors based methods, including Wang's, Song's and IISM, generally correlate better with expression similarity than those derived by single common ancestor based methods (including Resnik's, Li's and Jiang's methods).

Among the methods based on multiple common ancestors, our method (IISM) almost produces larger correlation coefficients than those produced by Song’s, which is followed by Wang’s approach. In particular, our measure correlations better with expression profile than all other approaches in the case that the number of bins is larger than 8, which implies that our method based on integrated information performs better even if the resolution is high. This may due to the fact that our approach not only combines both semantic information and depth information of the GO terms, but also takes into account information from multiple common ancestors of two terms.

Table 4. Correlation Coefficients between Gene Expression Similarities and Semantic Similarities Derived from Different Approaches

Classes	Resnik [22]	Lin [23]	Jiang [24]	Wang [18]	Song [32]	IISM
4	0.614	0.789	0.930	0.929	0.966	0.943
5	0.561	0.717	0.889	0.802	0.850	0.864
6	0.413	0.569	0.700	0.745	0.774	0.785
7	0.519	0.622	0.761	0.725	0.733	0.758
8	0.496	0.597	0.675	0.706	0.714	0.743
9	0.417	0.659	0.664	0.745	0.778	0.791
10	0.403	0.620	0.730	0.733	0.772	0.793
11	0.419	0.665	0.691	0.725	0.761	0.776
12	0.246	0.485	0.722	0.715	0.782	0.797
13	0.321	0.525	0.715	0.709	0.791	0.832

5. Conclusions

In this study, we proposed a novel approach for the semantic similarity measurement over GO terms by taking into account multiple common ancestors, and aggregating both semantic information and depth information of GO terms. The semantic value of a GO term was established on the concept of non-redundant ancestors and the common semantics two terms share was derived from their non-redundant common ancestors. The information associated with the probability that a non-redundant common ancestor occurs in the corpus was integrated into our IISM measurement to address the problem of ‘shallow annotation’. The validation experiments were conducted on both pathway dataset and gene expression dataset, and results on the both datasets show the superiority of our approach. Moreover, the computational complexity of our IISM approach is $O(n)$, which is more effective than other methods based on multiple common ancestors, this makes it suitable for large-scale study. IISM is an alternative to other methods based on multiple ancestors.

In addition, there are two issues should be noticed in this work: (1) the IC value of a GO term used in this paper is derived from the probability that it appears in a specific corpus, which will be affected by the annotation of specific species, and the IC value may vary across the annotations of different species. In addition, the IC value may change with the increase of our biological knowledge in the future; (2) our method only took into account the common ancestors in the upper part the GO graph. As a matter of fact, the descendants two terms share can also provide information to characterize their commonality. It will be help to consider the common descendants of two terms for better semantic similarity measure.

Acknowledgment. This work was supported in part by the National Natural Sciences Foundation of China under grants No. 60675016 and 60633030, fully by Educational Commission and Communications Department of Guangdong Province, China grants 2012-02-45, Natural Sciences Foundation of Guangzhou Maritime Institute grants K31012B09, and by Foundation for innovation project in Higher Education of Guangdong, China (A510602, B510620). An earlier version of this paper was presented at the International Conference on ICSAI2014.

References

1. Nariai, N., Kolaczyk, E. D., Kasif, S.: Probabilistic protein function prediction from heterogeneous genome-wide data. PLoS One, Vol. 2, No. 3, e337. (2007)
2. Alexa, A., Rahnenführer, J., Lengauer, T.: Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics, Vol. 22, No. 13, 1600-1607. (2006)
3. Yang, D., Li, Y., Xiao, H., Liu, Q., Zhang, M., Zhu, J., Ma, W., Yao, C., Wang, J., Wang, D., Guo, Z., Yang, B.: Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories. Bioinformatics, Vol. 24, No. 2, 265-271. (2008)
4. Mathur, S., Dinakarpandian, D.: Finding disease similarity based on implicit semantic similarity. Journal of biomedical informatics, Vol. 45, No, 2, 363-371. (2012)
5. Wang, H., Zheng, H., Browne, F.: Integration of Gene Ontology-based similarities for supporting analysis of protein–protein interaction networks. Pattern Recognition Letters, Vol. 31, No, 14, 2073-2082. (2010)
6. Benabderrahmane, S., Smail-Tabbone, M., Poch, O., Napoli, A., Devignes, M.D.: IntelliGO: a new vector-based semantic similarity measure including annotation origin. BMC bioinformatics, Vol. 11, No. 1, 588. (2010)
7. Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D., Jacq, B.: GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome Biology, Vol. 5, No. 12, R101. (2004)
8. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Wanker, E.E.: A human protein-protein interaction network: a resource for annotating the proteome. Cell, Vol. 122, No. 2, 957-968. (2005)
9. Wu, X., Pang, E., Lin, K., Pei, Z.M.: Improving the Measurement of Semantic Similarity between Gene Ontology Terms and Gene Products: Insights from an Edge-and IC-Based Hybrid Method. PloS one, Vol. 8, No. 5, e66745. (2013)
10. Pesquita, C., Faria, D., Falcao, A.O., Lord, P., Couto F.M.: Semantic similarity in biomedical ontologies. PLoS computational biology, Vol. 5, No. 7, e1000443. (2009)
11. Gan, M., Dou, X., Jiang, R.: From Ontology to Semantic Similarity: Calculation of Ontology-Based Semantic Similarity. The Scientific World Journal, Vol. 793091, 1-11. (2013)
12. Consortium, G. O.: Gene ontology consortium: going forward. Nucleic acids research, Vol. 43, No. D1, D1049-D1056. (2015)
13. Peng, J., Uygun, S., Kim, T., Wang, Y., Rheeet, S.Y., Chen, J.: Measuring semantic similarities by combining gene ontology annotations and gene co-function networks. BMC bioinformatics, Vol. 16, No. 1, 44. (2015)

14. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, Vol. 19, No. 1, 17-30. (1989)
15. Nagar, A., Al-Mubaid, H.: A new path length measure based on GO for gene similarity with evaluation using sgd pathways. In Proceedings of the 21st International Symposium on Computer-Based Medical Systems, 2008 CBMS'08. Jyväskylä, Finland, 590-595. (2008)
16. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, 133-138. (1994)
17. Richardson, R., Smeaton, A.F., Murphy, J.: Using WordNet as a knowledge base for measuring semantic similarity between words. In Technical Report Working Paper CA-1294. School of Computer Applications, Dublin City University. (1994)
18. Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., Chen, C.F.: A new method to measure the semantic similarity of GO terms. Bioinformatics, Vol. 23, No. 10, 1274-1281. (2007)
19. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In Proceeding of the 14th International Joint Conference on Artificial Intelligence, Montreal, San Fransisco, 448-453. (1995)
20. Othman, R.M., Deris, S., Illias, R.M.: A genetic similarity algorithm for searching the Gene Ontology terms and annotating anonymous protein sequences. Journal of biomedical informatics, Vol. 41, No. 1, 65-81. (2008)
21. Seco, N., Veale, T., Hayes, J.: An intrinsic information content metric for semantic similarity in WordNet. In Proceedings of the 16th European Conference on Artificial Intelligence, Valencia, Spain, 1089-1090. (2004)
22. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research, Vol. 11, 95-130. (1999)
23. Lin, D.: An information-theoretic definition of similarity. In Proceedings of the 15th international conference on Machine Learning, Madison, Wisconsin, USA, 296-304. (1998)
24. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of International Conference Research on Computational Linguistics, Taiwan, 19-33. (1997)
25. Couto, F.M., Silva, M.J., Coutinho, P.M.: Measuring semantic similarity between Gene Ontology terms. Data & knowledge engineering, Vol. 61, No. 1, 137-152. (2007)
26. Couto, F.M., Silva, M.J.: Disjunctive shared information between ontology concepts: application to Gene Ontology. Journal of Biomedical Semantics, Vol. 2, 1-5. (2011)
27. Bien, S.J., Park, C.H., Shim, H.J., Yang, W., Kim, J., Kim, J.H.: Bi-directional semantic similarity for gene ontology to optimize biological and clinical analyses. Journal of the American Medical Informatics Association, Vol. 19, No. 5, 765-774. (2012)
28. Yang, H., Nepusz, T., Paccanaro, A.: Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. Bioinformatics, Vol. 28, No. 10, 1383-1389. (2012)
29. Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Hayashizaki, Y.: An atlas of combinatorial transcriptional regulation in mouse and man. Cell, Vol. 140, No. 5, 744-752. (2010)
30. Washington, N.L., Haendel, M.A., Mungall, C.J., Ashburner, M., Westerfield, M., Lewis, S.E.: Linking human diseases to animal models using ontology-based phenotype annotation. PLoS biology, Vol. 7, No. 11, e1000247. (2009)
31. Xu, T., Du, L., Zhou, Y.: Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. BMC bioinformatics, Vol. 9, No 1, 472. (2008)
32. Song, X., Li, L., Srimani, P.K., Philip, S.Y., Wang, J.Z.: Measure the Semantic Similarity of GO Terms Using Aggregate Information Content. IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 11, No. 3, 468-476. (2014)

33. Bandyopadhyay, S., Mallick, K.: A New Path Based Hybrid Measure for Gene Ontology Similarity. IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 11, No. 1, 116-127. (2014)
34. Azuaje, F., Wang, H., Bodenreider, O.: Ontology-driven similarity approaches to supporting gene functional assessment. In Proceedings of the SIG meeting on Bio-ontologies, Detroit, Michigan, 9-10. (2005)
35. Teng, Z., Guo, M., Liu, X., Dai, Q., Wang, C., Xuan, P.: Measuring gene functional similarity based on group-wise comparison of GO terms. Bioinformatics, Vol. 29, No. 11, 1424-1432. (2013)
36. Ashburner, M., Ball, C.A., Blake, J.A., et al.: Gene Ontology: tool for the unification of biology. Nature genetics, Vol. 25, No. 1, 25-29. (2000)
37. Spellman, P.T., Zhang, M.Q., et al.: Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Molecular biology of the cell, Vol. 9, No. 12, 3273-3297. (1998)
38. Li, B., Wang, J.Z., Feltus, F.A., Zhou, J., Luo, F.: Effectively integrating information content and structural relationship to improve the go-based similarity measure between proteins. Arxiv preprint arXiv:1001.0958. 1-54. (2010)
39. Sevilla, J.L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J.M., Martinez-Cruz, L.A., Corrales, F.J., Rubio, A.: Correlation between gene expression and GO semantic similarity. IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 2, No. 4, 330-338. (2005)

Shu-Bo Zhang received the MSc and PhD degree in applied mathematics from Sun Yat-sen University, China, in 2005 and 2009, respectively. He joined Guangzhou Maritime Institute in 2009, where he is currently serves as an associate professor with the Department of Computer Science and Maritime Institute of Information Technology. His current research interests include pattern recognition, bioinformatics, ontologies. He is a senior member of the China Computer Federation.

Jian-Huang Lai received the MSc degree in applied mathematics and the PhD degree in mathematics from Sun Yat-sen University, China, in 1989 and 1999, respectively. He is currently a professor with the Department of Automation of School of Information Science and Technology. His current research interests include the areas of digital image processing, pattern recognition, multimedia communication, wavelet, and its applications. He serves as a standing member of the Image and Graphics Association of China and also serves as a standing director of the Image and Graphics Association of Guangdong. He is a senior member of the IEEE.

Received: November 30, 2014; Accepted: April 15, 2015.