Bi-directional semantic similarity for gene ontology to optimize biological and clinical analyses

Sang Jay Bien,[1] Chan Hee Park,[1,2] Hae Jin Shim,[1] Woongcheol Yang,[1,3] Jihun Kim,[1,2] Ju Han Kim[1,2,4]

[1]Seoul National University Biomedical Informatics (SNUBI), Seoul, Korea
[2]Systems Biomedical Informatics Research Center, Seoul, Korea
[3]Department of Computer Science, Princeton University, Princeton, New Jersey, USA
[4]Division of Biomedical Informatics, Seoul National University College of Medicine, Seoul, Korea

Correspondence to
Professor Ju Han Kim, Division of Biomedical Informatics, Seoul National University College of Medicine, 28 Yongon-dong Chongno-gu, Seoul 110799, Korea; juhan@snu.ac.kr

SJB and CHP contributed equally to this work.

Received 22 October 2011
Accepted 16 January 2012

ABSTRACT

Background Semantic similarity analysis facilitates automated semantic explanations of biological and clinical data annotated by biomedical ontologies. Gene ontology (GO) has become one of the most important biomedical ontologies with a set of controlled vocabularies, providing rich semantic annotations for genes and molecular phenotypes for diseases. Current methods for measuring GO semantic similarities are limited to considering only the ancestor terms while neglecting the descendants. One can find many GO term pairs whose ancestors are identical but whose descendants are very different and vice versa. Moreover, the lower parts of GO trees are full of terms with more specific semantics.

Methods This study proposed a method of measuring semantic similarities between GO terms using the entire GO tree structure, including both the upper (ancestral) and the lower (descendant) parts. Comprehensive comparison studies were performed with well-known information content-based and graph structure-based semantic similarity measures with protein sequence similarities, gene expression-profile correlations, protein—protein interactions, and biological pathway analyses.

Conclusion The proposed bidirectional measure of semantic similarity outperformed other graph-based and information content-based methods.

Semantic similarity is a concept whereby a set of documents or terms are assigned a metric based on the likeness of their meaning or the degree of taxonomical proximity. The determination of the semantic similarity between words has been successfully applied in many biomedical areas such as document categorization or clustering,[1 2] information retrieval,[3 4] and genomic data analysis.[5–7] Biomedical semantic similarity has been determined by defining a topological similarity, using statistical means to exploit the amount of co-occurrences between word contexts, or by using ontologies to define the distance between words based on the taxonomical structure.

Methods of determining semantic similarity have recently been very extensively studied for gene ontology (GO), which is becoming one of the most important and rapidly growing biomedical ontologies[8] with the increasing biomedical utility of genomic data with GO annotations. GO is a set of controlled vocabularies, describing biological processes (BP), molecular functions (MF), and cellular components (CC) for the annotation of genes and molecular phenotypes for diseases.[9]

Semantic similarity measures between GO terms can be classified into information content (IC)-based[5 6 10 11] and graph structure-based[7 12] ones. Lord and colleagues[5 6] for the first time applied Resnik's measure of semantic similarity[13 14] to quantify GO term specificities. They evaluated three IC-based measures and concluded that the Resnik's measure showed the best performance.[6] However, Wang *et al*[7] correctly pointed out that IC-based similarity measures tended to vary from species to species because they relied only on the annotation frequency of GO terms to gene products, which were different from species to species. They believed that the specificity of a GO term should be determined by biological meanings, not by their annotation statistics, and proposed a new semantic similarity measure determined only by the GO ontological structures.

However, the measure of semantic similarity of Wang *et al*,[7] given a GO term (or a pair of terms), considers only the ancestral (or upper) terms and neglects the lower (or descendant) ones in a GO graph (see figure 1). The unidirectional nature of the semantic similarity measurement of Wang *et al*[7] has limitations. The lower portions of the GO graphs contain more GO terms that have more specific semantics and semantic relations. GO annotators and curators spend more effort for this detailed portion of the GO graphs. Moreover, many GO term pairs sharing identical ancestors may have very different descendants and vice versa, resulting in severe semantic inconsistencies.

To evaluate semantic similarity measures, Lord *et al*[5 6] and Schlicker *et al*[10] applied protein sequence similarity as the 'gold standard'. Biological pathways and membership of proteins in protein complexes have also been used for evaluation. Guo *et al*[15] proposed a 'positive' dataset including the first-degree neighbors directly connected in the Kyoto encyclopedia of genes and genomes (KEGG)[16] biological pathway graphs and the members of protein complexes. Random pairs of proteins were generated as the 'negative' dataset. The correlations of gene expression profiles from DNA microarray data have also been applied to measure the functional (or semantic) similarity of genes and molecular phenotypes of diseases.[17]

In the present study, we propose a novel method that applies a bidirectional measure of GO semantic similarity, considering the entire GO graph structure including both the upper (ancestral) and the lower (descendant) parts. We first propose a descending semantic similarity measure and demonstrate by means of illustration and comparison studies the necessity of designing a bidirectional measure of carefully combining both ascending and descending semantics. Next, we performed a comprehensive

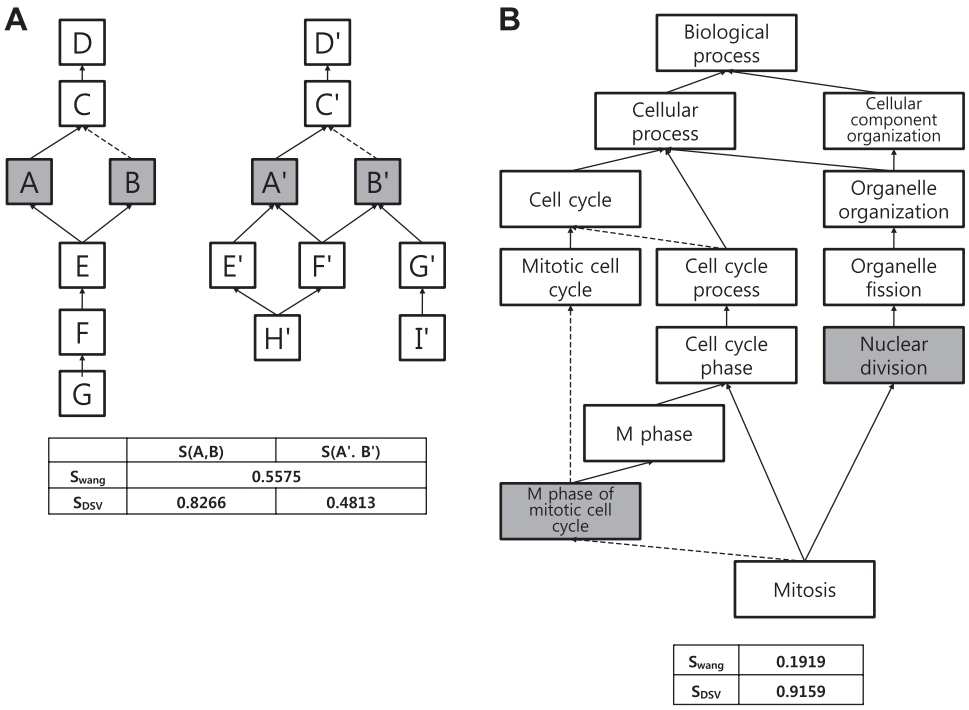

Figure 1 Ascending and descending measures of semantic similarities between gene ontology (GO) terms. (A) Although terms A and B, given the directed acyclic graphs (DAG) structure, must show higher similarity than terms A′ and B′, Wang's semantic similarity considering only the ascending part cannot discern the difference (ie, S_{Wang}=0.5575 for both). In contrast, S_{DSV} clearly discerns the two pairs (ie, 0.8266 and 0.4813). (B) Although the two terms, 'nuclear division' and 'M phase of mitotic cell cycle', must be semantically similar because they share 'mitosis' and its descendants (omitted), ancestor-dependent S_{Wang} is very low (=0.1919). S_{DSV} considering the descendants, however, suggests a high level of semantic similarity (=0.9519). Solid and dotted lines depict 'is_a' and 'part_of' relationships, respectively.

evaluation study comparing established IC and graph structure-based semantic similarity measures using protein sequence similarities, gene expression–profile correlations, protein–protein interactions, and biological pathway membership. Our novel bidirectional measurement of semantic similarity of GO terms outperformed others.

METHODS

Semantic similarity between GO terms

GO consists of three major categories: BP, MF, and CC. BP is a series of events accomplished by one or more ordered assemblies of molecular functions. MF describes activity at the molecular level. CC consists of the location of the cell, from the levels of subcellular structures to macromolecular complexes.

In GO directed acyclic graphs (DAG), a child concept is an instance or a component of the parent concept. As DAG allows multiple inheritances, one concept may have multiple parent concepts with different relations among the five: 'is_a'; 'part_of'; 'regulates'; 'positively regulates'; and 'negatively regulates'. GO obeys a rule called the 'true-path rule'. The more specific the common ancestors of a pair of terms are, the closer the distance between the terms is. On the other hand, as the common ancestors of a pair of terms become general, the distance between the terms becomes farther.

IC-based semantic measures quantify the specificity of a term. The IC of a concept, t_0, is defined as the probability of encountering an instance of the concept t_0 in the corpus,[13 14] and is given by

$$\mathrm{IC} = -\log(p(t_0)) \tag{1}$$

$$p(A) = \frac{freq(t_0)}{freq(root)} \tag{2}$$

$$freq(t_0) = annot(t_0) + \sum_{c \in \text{children}(t_0)} annot(c) \tag{3}$$

where $annot(t_0)$ is the number of occurrences of the term t_0 from the corpus. Resnik's[13 14] semantic similarity measures the similarity of two terms using the IC of the lowest common ancestor of the two terms and thus is defined as

$$S_{\mathrm{Resnik}}(t_A, t_B) = \mathrm{IC}(LCA(t_A, t_B)) \tag{4}$$

Lord *et al*[5 6] for the first time applied the technique from information theory to determine semantic similarity between genes. IC-based measures, however, tend to vary from species to species because they rely on annotation frequency statistics, and different species may have different annotations even for the same genes and molecular phenotypes of diseases. Wang *et al*[7] believed that the specificity of a GO term has to be determined by the GO term's semantics (or biological meanings), not by their annotation frequencies.

Wang *et al*[7] viewed the semantic value of a term, t_0, as the aggregate contribution of semantics from the subgraph, $P(t_0)$, containing t_0 itself and its ancestors all the way up to the root node. For any ancestor term t of term t_0, the ascending S-value of t related to t_0, $AS(t_0, t)$, is defined as

$$\left\{ \begin{array}{c} AS(t_0, t_0) = 1 \\ AS(t_0, t) = \max\{w_e \cdot AS(t_0, t') | t' \in C(t)\} \text{ if } t \neq t_0 \end{array} \right\} \tag{5}$$

where $C(t)$ are the children of term t, and w_e is the semantic contribution factor for the edge that links term t with its child t. Wang *et al*[7] set semantic contribution factors for 'is_a' and 'part_of' relations of GO hierarchy to 0.8 and 0.6, respectively.

Term t_0 has the most specific semantics in $P(t_0)$ and its contribution to its own semantics is defined as 1. Other terms in $P(t_0)$ are more general and thus contribute less to the semantics of t_0. Therefore, the range of w_e is {0,1}. After obtaining the ascending S-values for all terms in $P(t_0)$, the semantic value of term t_0, $SV(t_0)$, is calculated as

$$SV(t_0) = \sum_{t \in P(t_0)} AS(t_0, t) \tag{6}$$

Given $P(t_A)$ and $P(t_B)$ for two GO terms, t_A and t_B, respectively, the semantic similarity between them is calculated as follows:

$$S_{\text{Wang}}(t_A, t_B) = \frac{\sum_{t \in P(t_A) \cap P(t_B)} (AS(t_A, t) + AS(t_B, t))}{SV(t_A) + SV(t_B)} \quad (7)$$

where S_{Wang} refers to Wang *et al*'s[7] measure of semantic similarity.

Each GO term is made for the needs of biologists who describe the real world by biological concepts. As a child term is a special case of the parent, it is assumed that the parent term's semantics are the union of its children's. GO allows for multiple inheritance, and two semantically similar terms are likely to share their child terms, inheriting both concepts of the two terms. Descending semantic similarity can thus also be quantified by the shared child terms. Figure 1A clearly shows that even if the GO term pairs have identical ancestral topologies, their descendant topology may be very different. Therefore, pairs having the same S_{Wang} values can be discerned further using their descendant topologies. Of course, pairs having the same descendant topologies can be discerned further using their ancestral topologies. It is clear that both ascending and descending semantics should be used together in a balanced manner to improve the semantic similarity measures.[7]

We define descending *S*-value (*DS*) and descending semantic value (*DSV*) as follows:

$$\left\{ \begin{array}{c} DS(t) = 1 \text{ if } t \in L \\ DS(t) = \min\{w_e \cdot DS(t') | t' \in C(t)\} \text{ if } t \notin L \end{array} \right\} \quad (8)$$

$$DSV(t_0) = \sum_{t \in C(t_0)} DS(t) \quad (9)$$

$$S_{DSV}(t_A, t_B) = \frac{\sum_{t \in C(t_A) \cap C(t_B)} (2 \cdot DS(t))}{DSV(t_A) + DSV(t_B)} \quad (10)$$

where L are terminal leaf terms. Leaf terms are the most specific ones. Leaves are fixed such that *DS* takes not a relative but an absolute value. Semantic contribution factors are set to 0.8 and 0.6 for 'is_a' and 'part_of' relations, respectively, as in the approach of Wang *et al*.[7] GO recently added three more relationships (ie, regulation, positive regulation, and negative regulation), and we set the semantic contribution factors for them as 0.6 for the purpose of comparison since they are 'part_of' relations in the study of Wang *et al*.[7]

Wang *et al*'s[7] $AS(t_0,t)$ represents the specificity of term t for term t_0 such that t_0 and $AS(t_0,t)$ may differ for each comparison. Computing $DS(t)$ requires more effort than computing $AS(t_0,t)$. Using leaf nodes instead of the term of interest (ie, t_0) in our descending *S*-value, $DS(t)$, has a normalization effect; however, a sub-tree of a term may have multiple leaf nodes. These leaves, called 'source' in graph theory, exert a strong influence on the *DSV* of their parent node. If we choose 'maximum' instead of 'minimum' in equation (8), $DS(t_0,t)$ becomes very unstable due to a shallow sub-tree effect. We chose 'minimum' instead of 'maximum' to prevent this.

Our approach seems to support human intuition. S_{DSV} says that 'M phase of mitotic cell cycle' and 'nuclear division' are semantically similar terms (=0.92, figure 1B). In fact, they are very close to 'mitosis' and their descendants are almost the same. In contrast, S_{Wang} says that they are distant (=0.19) because they share only two ancestors, 'cellular process' and 'biological process', which are very general terms (figure 1B).

We developed a combined measure of bidirectional semantic similarity, S_{BSV}, as follows:

$$S_{BSV}(t_A, t_B) = \frac{\alpha \cdot S_{Wang}(t_A, t_B) + \beta \cdot S_{DSV}(t_A, t_B)}{\alpha + \beta} \quad (11)$$

where α and β are the numbers of total ancestors and total descendants of t_A or t_B, respectively. S_{BSV} complements the limitation of S_{DSV} that considers descending nodes only. More importantly, S_{BSV} tries to include 'depth factor' for comparisons. Due to the recursive dependence on common descendants, S_{DSV} is more likely to impact comparisons involving concepts that are higher in the hierarchy. Notice that S_{Wang} and S_{DSV} are not symmetrical in that S_{Wang} is affected relatively less by the depths of the terms in comparison. Terms eventually converge up in the hierarchy. However, terms having logical reasons to have shared descending semantics may not have common descendants, just because more specific child concepts are not yet created, and reach terminal leaves. S_{BSV} weighs S_{DSV} more when terms are higher in the hierarchy but less when they are lower in the hierarchy. The total number of descendants of two terms, β, complements the drawback of S_{DSV} by reducing the weight of S_{DSV} for terms with a small number of total descendants with a decreased chance of having common descendants regardless of their semantic similarity. β tries to accommodate the similarity measure between higher and lower terms. Due to the property of this 'depth factor', S_{BSV} is different from S_{Wang} even when S_{DSV} equals zero.

Similarity between genes and molecular phenotypes

Gene products are annotated by GO terms. Therefore, semantic similarity between gene products can be regarded as semantic similarity of the GO term sets. Wang *et al*[7] defined set-wise similarity as:

$$S(t, G) = \max_{1 \le i \le k} (S(t, t_i)) \quad (12)$$

$$S(G_1, G_2) = \frac{\sum_{1 \le i \le m} S((t_{1i}, G_2) + \sum_{1 \le j \le n} S(t_{2j}, G_1)}{m + n} \quad (13)$$

where G is a set of GO terms and k is the number of terms in G. G_1 and G_2 consist of m and n terms, respectively.[7] Term-wise similarities can be replaced by S_{Resnik}, S_{Wang}, S_{DSV}, S_{BSV}, etc. The similarities of the most similar pair of terms from each annotation are averaged over to calculate set-wise similarity. We used BP annotations only for the following evaluation steps.

Validation

We performed a comprehensive validation study comparing IC and graph structure-based semantic similarity measures including our newly proposed ones. For the purpose of illustration, we explored the whole GO hierarchy to find the terms showing the biggest discrepancies between the ascending and descending measures. Second, we performed extended replication of the evaluation study of Lord *et al*[5 6] that did not include graph-based measures. We applied protein sequence similarity as the 'gold standard' to compare GO annotation-based semantic similarities calculated by different measures.

Semantic similarity measures are more valuable for investigating functional states such as gene-expression clusters and biological pathway memberships than structures such as protein sequences. To assess the resolution power of the similarity measures, we applied F-statistic comparing for 'between-group' and 'within-group' similarities. For a comprehensive evaluation study, we downloaded three datasets from the gene expression

omnibus:[18] GSE412: treatment-specific changes in gene expression discriminate in-vivo drug response in human leukemia cells; GDS1244: phosgene effect on lungs: time course; and GDS2159: spinal cord injury model: time course. We calculated the correlation coefficients of gene expression profiles for all gene pairs for each dataset. All pairs were sorted according to their correlation coefficients. Figure 2 shows our evaluation scheme. The within-group difference is controlled by s applied equally to the three comparison groups on the horizontal axis of the correlation coefficient. The larger the window size, s, the larger the within-group difference. The difference between the three comparison groups is controlled by the 'between-group' distance, d. We randomly sampled 1000 pairs for each window using 3 s (=0.025, 0.05, 0.1) and 3 d (=0.05, 0.1, 0.2). We repeated the comparison tests for each of the nine s–d pairs for each dataset by sliding the window from the leftmost (R=0) to the rightmost (R=1) levels of the correlation coefficient values shown in figure 2B.

Using receiver operating characteristic curve analysis, we quantitatively evaluated the semantic similarity measures using human protein–protein interaction and biological pathway datasets. The first positive dataset was assembled from the UniProt database from which we gathered all human proteins and their interaction data. After filtering out proteins without interactions, we found 10 348 protein pairs with GO BP annotations. The negative set was created by randomly sampling the same number of protein pairs.

The second dataset comes from BioCarta.[19] We extracted 41 697 protein pairs from the 343 BioCarta pathways with the same number of negative pairs. The third one comes from KEGG[16] with 8839 protein pairs from the selected seven KEGG categories (carbohydrate metabolism, energy metabolism, lipid metabolism, nucleotide metabolism, amino acid metabolism, glycan biosynthesis and metabolism, and metabolism of cofactors and vitamins). Compared with the broader categories of the BioCarta pathways, we used only the metabolism-related categories for KEGG to create a much harder discrimination problem. The negative datasets were created within the comparison categories using the same procedure.

RESULTS

The list of extreme GO term pairs that look very distant by an ascending (or a descending) measure but very close by a descending (or an ascending) measure is exemplified in table 1A (or table 1B). Their ascending and descending similarities were most discrepant among all GO pairs. It is clear that semantic similarity measures depending only on ancestral or only on descendant terms have limitations. All pairs in table 1 are similar in a sense because they are similar in at least one of the measures.

Histamine secretion (GO:0001821) and histamine production involved in acute inflammatory response (GO:0002349), for instance, are very different in terms of ascending semantic similarity (S_{Wang}=0.062) but very similar in terms of descending measure (S_{DSV}=0.872). Bidirectional measures assigned a reasonably high value (S_{DSV}=0.273). Those that have low ascending but high descending semantic similarities in table 1A were those that diverged up in the GO tree and then converged thereafter. Some terms diverged because different biological contexts are required to describe their contextual difference, but then eventually converged because they have the same or at least very close concepts.

As S_{DSV} considers descendant nodes only, the descending semantic similarity of any terminal leaf node pair, even if they are siblings, vanishes. The pairs in table 1B whose ascending similarities are very high (S_{Wang}>0.9) with vanished descending similarities (S_{DSV}=0) were mostly 'sibling' leaf nodes like pointed-end (GO:0010034) and barbed-end (GO:0051016) actin filament capping. As they are siblings deep in the tree, their ascending similarities are very high, but their descending similarities are zeros because they have no children. As a GO tree has so many leaf nodes, approximately two-thirds of all pair-wise S_{DSV} values are zeros. On the contrary, the average S_{Wang} for all pairwise calculations is approximately 1.0. The S_{BSV} values, on the contrary, assign reasonably high but still discernible semantic similarities to both categories.

Protein sequence similarity-based evaluation

Lord and colleagues[5 6] used protein sequence similarities measured by the BLAST algorithm as the 'gold standard' for evaluating IC-based semantic similarity measures. We replicated the same procedure for a fair comparison with an extension to graph-based ones. First, we downloaded SWISS-PROT protein sequences with available GO BP annotations. The number of sequences has approximately doubled to 13 933 compared with the study of Lord *et al*.[5 6] We excluded sequences with no BP annotation, returning 12 376 protein sequences. Next, we performed a BLAST search to find the best matching protein pairs and their bit scores.

Table 2 shows the correlation coefficients between ln(bit score) and the semantic similarities in the comparison. Resnik's measure showed the best performance among the IC-based ones, which is consistent with the findings of the study of Lord *et al*.[5 6] Lord *et al*[5 6] did not have a chance to compare graph-based measures at

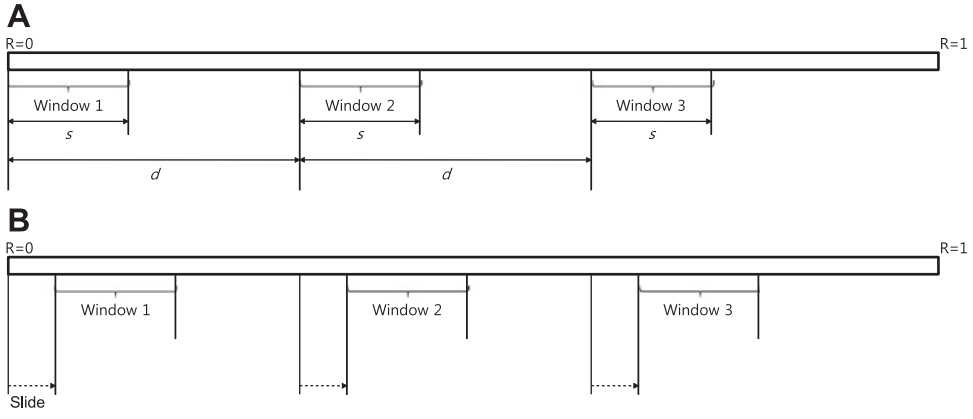

Figure 2 Evaluation schemes for semantic similarity measures. (A) All gene pairs are sorted by expression-profile correlation coefficients, R. Several sliding window sizes (ie, s=0.025, 0.05, 0.1), and distances (ie, d=0.05, 0.1, 0.2) are applied for a vigorous and systematic evaluation. (B) While sliding the windows from R equals 0 to 1, F-values for all comparison are calculated to test the statistical significance of the discriminating powers of different semantic similarities.

Table 1 GO term pairs showing the biggest differences between ascending S_{Wang}, descending S_{DSV}, and bidirectional S_{BSV} measures of semantic similarities

	Term 1	Term 2	S_{Wang}	S_{DSV}	S_{BSV}
(A)	Response to acetate (GO:0010034)	Initiation of acetate catabolic process (GO:0043077)	0.039	0.876	0.164
	Elevation of cytosolic calcium ion concentration (GO:0007204)	Cytosolic calcium ion transport (GO:0060401)	0.028	0.849	0.358
	Neuron projection regeneration (GO:0031102)	Response to axon injury (GO:0048678)	0.088	0.903	0.647
	Histamine secretion (GO:0001821)	Histamine production involved in acute inflammatory response (GO:0002349)	0.062	0.872	0.273
(B)	Pointed-end actin filament capping (GO:0051694)	Barbed-end actin filament capping (GO:0051016)	0.960	0.000	0.936
	Suppression by virus of host extracellular antiviral response (GO:0019053)	Suppression by virus of host intracellular antiviral response (GO:0019052)	0.956	0.000	0.852
	Replication fork protection (GO:0048478)	Replication fork arrest (GO:0043111)	0.951	0.000	0.858
	Pointed-end actin filament uncapping (GO:0051696)	Barbed-end actin filament uncapping (GO:0051638)	0.949	0.000	0.919

GO, gene ontology.

that time. Table 2 demonstrates that graph-based measures including S_{Wang}, S_{DSV} and S_{BSV} outperform the IC-based measures of Resnik,[13 14] Lin[20] and Jiang and Conrath.[21] Our combined measure, S_{BSV}, showed the highest correlation coefficient but the differences are too small to achieve statistical significance among the graph-based ones. We concluded that our descending and bidirectional measures are at least as good as the classic ascending measures in terms of protein sequence similarity prediction.

Gene expression-profile similarity-based evaluation

Figure 3 shows the results of the evaluation study based on gene expression-profile similarity. The bidirectional measure, S_{BSV} (black lines), seems to take advantage of both ascending and descending measures in that S_{BSV} follows S_{Wang} when it performs well and S_{DSV} when it performs well (figure 3A–C). Although S_{Resnik} is a well-known and highly performing semantic similarity measure, it had poorer F-values (in the vertical axis) than that of most graph-based measures in our evaluation study. Consistent with the study of Wang *et al*,[7] ontology structure-based S_{Wang} had a better resolution power than IC-based S_{Resnik}.

Although S_{Resnik} showed low performance in general, S_{Resnik} got better when s and d were very big, as shown in the upper right corners (s=0.1, d=0.2) in figure 3A,B). In contrast to the right upper corners representing easier discrimination problems, the left lower corners (s=0.025, d=0.05) represent harder ones. The descending measure, S_{DSV}, showed very high performance in the left lower corners, representing better discerning power for gene expression profiles with higher similarities. S_{DSV} outperformed S_{Wang} and the others except for the large s and large d regions. It seems that the bidirectional S_{BSV} measure compensates for the unidirectional descending S_{DSV} and ascending S_{Wang} measures for their areas of weaknesses.

Biological knowledge-based evaluation

Figure 4 shows that semantic similarity measures can be used to predict protein–protein interactions and biological pathway memberships with reasonably high performance. As the KEGG metabolic pathway constitutes a harder problem than BioCarta, the receiver operating characteristic curves for KEGG (figure 4C) showed poorer performances than those for BioCarta (B) for all four of the measures. Once again, we see that the descending S_{DSV} measure outperformed the ascending S_{Wang} measure for those harder discrimination tasks and the bidirectional S_{BSV} used the advantages of both. All graph-based measures outperformed the IC-based S_{Resnik}. Other IC-based methods were omitted from the graph due to lower performances than that of S_{Resnik}.

Figure 4A shows that the descending S_{DSV} measure may have a point where the discriminating power is saturated. The UniProt database is well annotated and rich in very specific GO terms. The S_{DSV} of more than one-third of the protein pairs (n=3515) was thus zero. It is a nice demonstration of the limitation of S_{DSV}. The descending S_{DSV} measure is poor in distinguishing the semantic distances of leaf-to-leaf pairs or of pairs near the terminal leaves having few descendants. Nevertheless, S_{DSV} shows very good performance before the saturation point, and our bidirectional measure, S_{BSV}, shows the best performance among these, using the advantages of both the ascending S_{Wang} and descending S_{DSV} measures. Please note that S_{BSV} is not equal to S_{Wang} even when S_{DSV} equals zero due to the 'depth factor' (see the Methods section). The measure of Resnik[13 14] shows relatively high performance for this protein–protein interaction dataset (figure 4A) compared with the others (figure 4B,C). It seems that the high specificity annotations of the UniProt database complements S_{Resnik}'s low resolution problem, returning a high level of discriminatory performance.

Impact of introducing bidirectional semantic similarity measure

Because approximately a third of GO BP terms are leaves, it is important to have an approximate idea of what proportion of the comparisons will yield a different value with the new measure. S_{BSV} is not merely a weighted summation of S_{Wang} and S_{DSV} but applies a 'depth factor' such that S_{BSV} does not become S_{Wang} even if S_{DSV} equals zero (see the Methods section). Moreover, entities such as genes, gene clusters, and biological pathways are annotated with more than one GO term. Figure 5 shows the proportions of changes, $(S_{Wang}-S_{BSV})/S_{Wang}$, introduced by applying bidirectional measures, S_{BSV}, between gene pairs. The dotted line depicts the frequency distribution of the proportion of semantic similarity changes of 9088 among the 12 376 best matched protein pairs by BLAST for the protein sequence similarity evaluation-based study. Only 229 pairs showed no change (or $(S_{Wang}-S_{BSV})/S_{Wang}$=0). We removed the 3288 (=12 376–9088) pairs having perfectly identical GO annotations because their semantic similarities cannot be changed from 1.0 by any measure. The average of the proportions of changes was 0.290.

Table 2 Correlation coefficients between protein sequence similarity measured by BLAST bit score and various semantic similarities

	IC-based			Graph-based		
	S_{Resnik}	S_{Lin}	$S_{JiangConrath}$	S_{Wang}	S_{DSV}	S_{BSV}
Correlation coefficient	0.220	0.170	0.192	0.353	0.356	0.357

IC, information content.

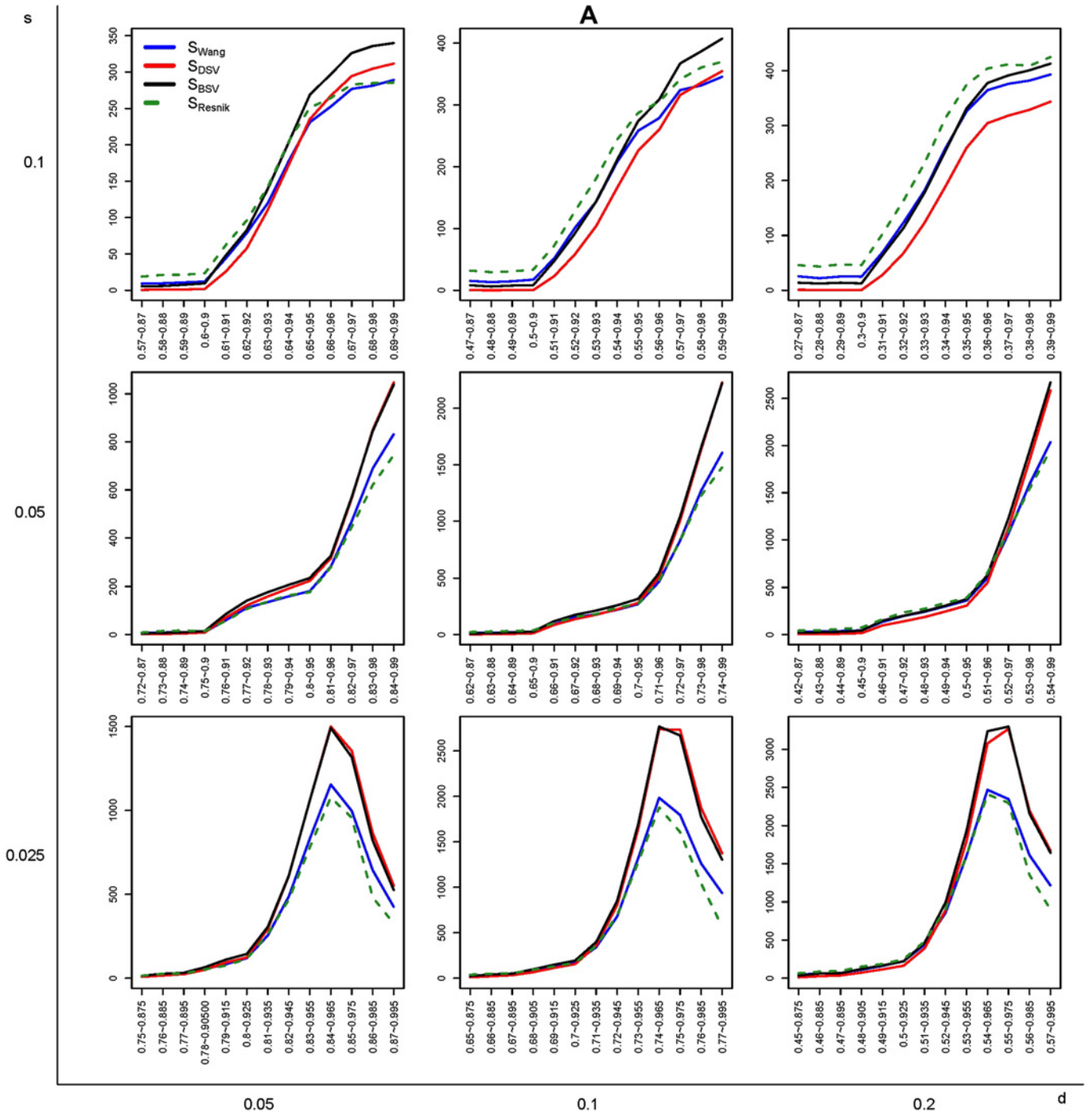

Figure 3 Evaluation of semantic measures using microarray gene expression-profile similarities for (A) GSE412, (B) GDS1244, and (C) GDS2159 datasets downloaded from the gene expression omnibus. Correlation coefficients between gene expression profiles were calculated for all gene pairs. F-test was applied for testing the discriminant power of the semantic measures by varying window sizes (s=0.025; 0.5; 0.1) and window distances (d=0.05; 0.1; 0.2) across different levels of correlation coefficients by sliding the windows (see figure 2 for the evaluation scheme). Inner horizontal and vertical axes represent correlation coefficient and F-statistic, respectively. Outer horizontal and vertical axes represent window distance *d* and window size *s*, respectively.

We downloaded 7804 human genes having at least one GO annotation from the GO annotation database and randomly sampled 9000 pairs. We discarded gene pairs having perfectly identical GO annotations during the sampling procedure because their semantic similarities can be changed by no measure. The solid line depicts the frequency distribution of the proportion of changes. Only two pairs showed no change. The average of the proportions of changes was 0.623. Best-matched protein pairs showed smaller changes (dotted line) than randomly sampled gene pairs from the GO annotation database (solid line) because sequence-matched proteins are semantically more similar. Some pairs, however, showed big change even though they are the best-matched pairs by BLAST.

Introducing a descending measure for computing semantic similarity can be justified by the existence of multiple inheritances. We found that 11763 (68.3%) among 17217 GO BP

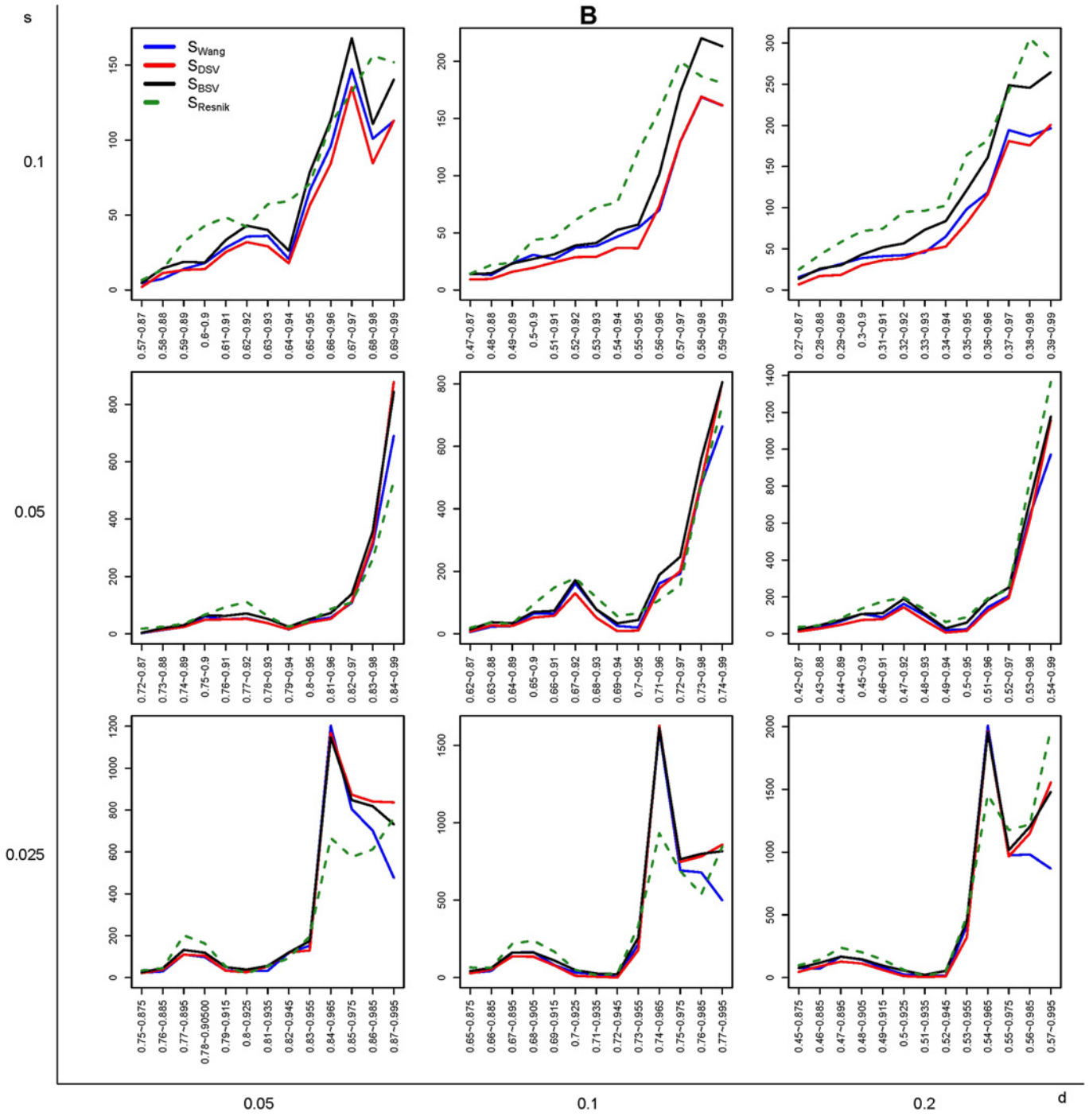

Figure 3 Continued

terms have more than one parent. It was 50.3% (=14 646/29 139) in all GO terms. The average number of parents of a term was approximately 2.03. In medical subject heading (MeSH) hierarchy, we found that 36 817 (73.9%) among 49 836 terms have more than one parents. MeSH showed more multiple inheritances, with 2.94 parents per a term.

DISCUSSION

S_{DSV} and S_{BSV} extend and improve the ascending semantic similarity of Wang *et al*.[7] We applied our method for measuring the semantic similarity of genes and molecular phenotypes of diseases using ontological relations of GO terms. We performed comprehensive evaluation studies and theoretical analysis. While the scope of the present study has been limited to GO term similarities, the improved measure of semantic similarity can be applied as is to other biomedical ontologies such as ICD, MeSH, SNOMED-CT, etc.[22 23] As Jensen and Bork[8] pointed out, GO has become the dominating biomedical ontology over a period of just 5 years at least in terms of how often they are mentioned in PubMed abstracts. Almost all biomedical ontologies are either simple tree structures that represent hierarchical classifications or DAG. The difference is that the latter allows a term to be related to multiple broader terms, whereas the former does not. Moreover, some of the GO terms are middle phenotypes between the cellular and molecular function levels and the disease levels, explaining pathophysiology.

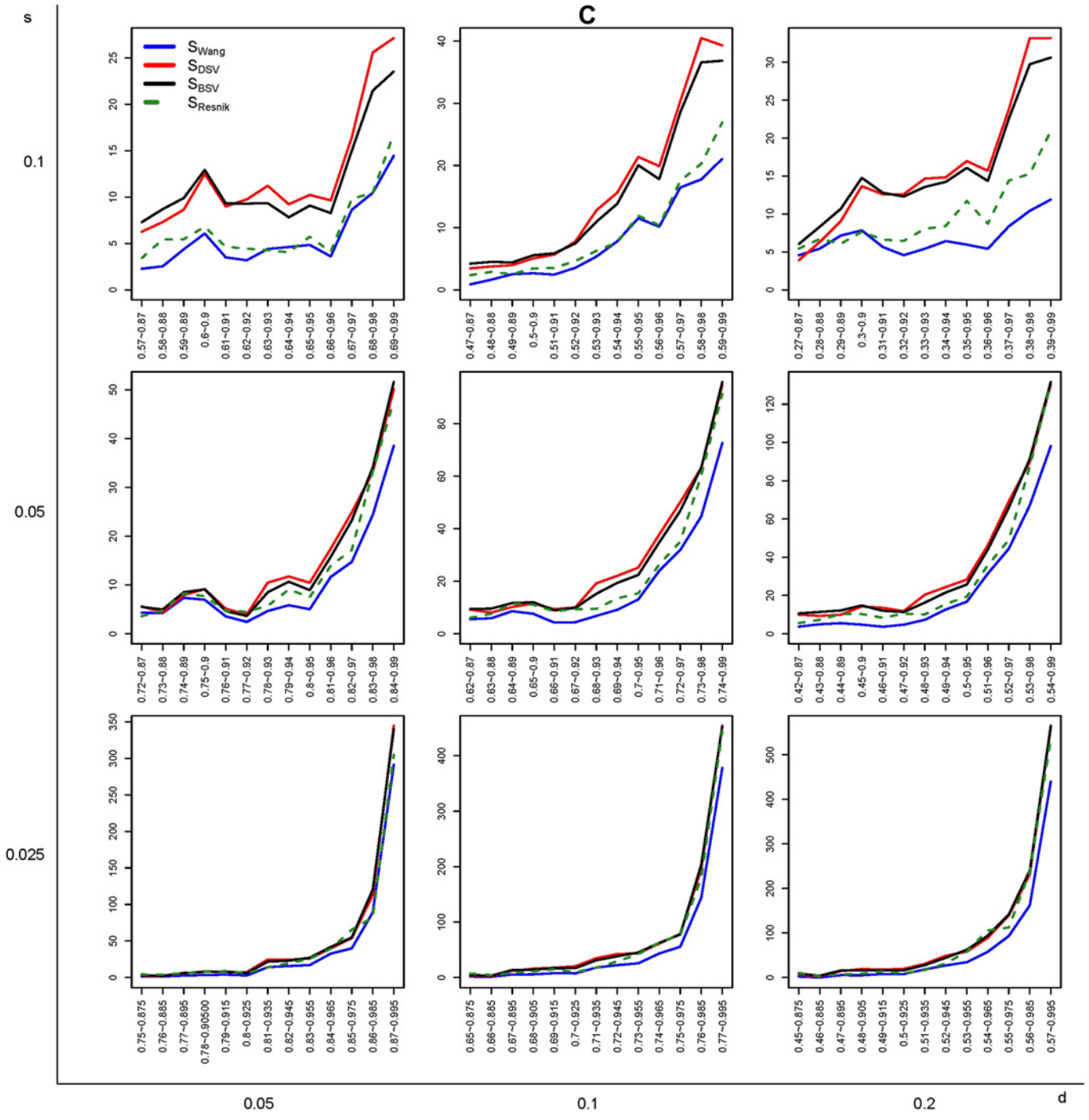

Figure 3 Continued

Although S_{Wang} is a well-known highly performing measure, it suffers from limitations. First, when two terms in comparison are near the root, they have few common terms and the similarity measure becomes unstable. This is a symmetrical problem to the 'leaf-to-leaf pair' problem of descending S_{DSV} measure that is well demonstrated in table 1B and figure 4A. Cell communication (GO:0007154) and cell death (GO:0008219), for example, are children of cellular process (GO:0009987), which is a child of the GO BP root term such that their ascending semantic similarity is relatively high (ie, S_{Wang}=0.507). Cell death has one more path to the root via its parent death (GO:0016265), which is a child of the GO BP root term. Although they are very high in the hierarchy they have very few shared descendants such that S_{DSV}=0.003 and S_{BSV}=0.004.

Second, when two terms are descendants of distant parents but soon converge, the ascending S_{Wang} measure regards them as distant pairs by neglecting their many shared descendants (see table 1A). This 'diverge-then-converge' pairs are inevitable given the DAG structure and there are many such pairs in the GO DAG structure. As described in the Results section, 68.3% of GO BP, 50.3% of all GO, and 73.9% of MeSH terms have more than one parent.

The descending *S*-value is designed in a very different way than the ascending *S*-value. Wang *et al*[7] defined the ascending *S*-value (or AS) of a term, t, as a contribution of t to t_0. Term t_0

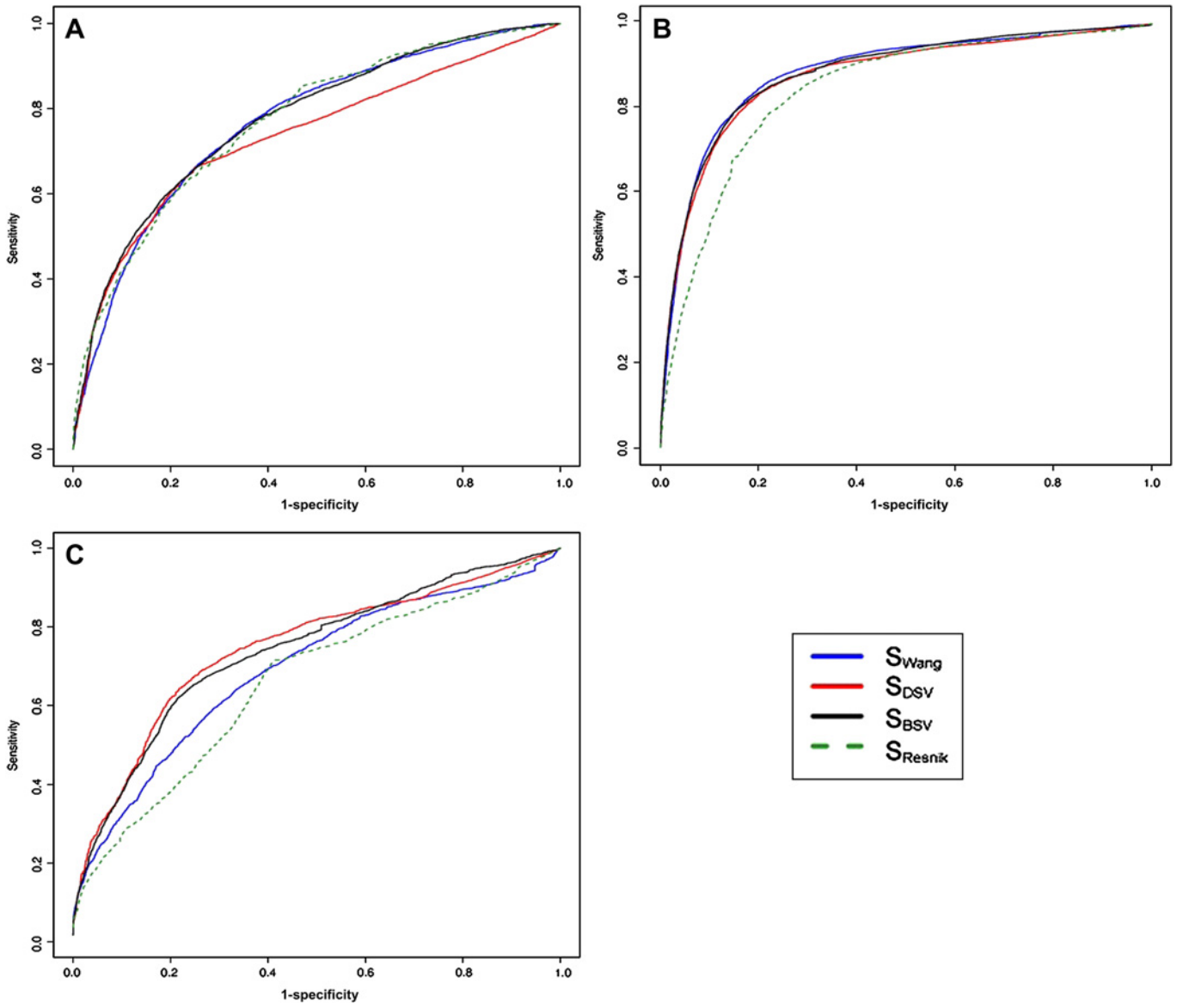

Figure 4 Receiver operating characteristic curve analyses to evaluate semantic similarity measures. The positive datasets were extracted from (A) UniProt protein–protein interaction data, (B) Biocarta, and (C) KEGG biological pathways. Negative sets were created by random sampling from the corresponding datasets. Other information content-based measures are omitted because of their poorer performances compared to Resnik's measure.

has the most specific semantics in $P(t_0)$ and its contribution to its own semantics is defined as 1. The ascending *S*-value of a term, *t*, thus varies according to t_0 such that computing the ascending semantic value requires two variables, $AS(t_0, t)$. While the ascending *S*-value of a term, *t*, is a relative value, the descending *S*-value (DS) of a term, *t*, is an absolute one because terminal leaves always have the most specific semantics of 1 in $C(t_0)$. Computing DS requires only one variable, $DS(t)$, and the minimum value will be chosen among the many paths. Both semantic values can be obtained by summating ascending and descending *S*-values of all members of the subgraphs, $P(t_0)$ and $C(t_0)$, respectively. One can pre-compute $DS(t)$ for all GO terms because DS has absolute value.

The descending measure, S_{DSV}, is designed to impact comparisons involving concepts that have large numbers of descendants. S_{BSV} complements the limitation of S_{DSV} that considers descending nodes only. Comparisons may involve higher and lower terms together. The total number of descendants of two terms, β, or the 'depth factor' works as a reasonable compensator. Wang *et al*[7] demonstrated that graph-based measure shows better resolution power for harder problems. The present study demonstrated that S_{DSV} improves the resolving power by utilizing more specific terms down the hierarchy and S_{BSV} complements its drawback. Figure 4B,C demonstrates that S_{DSV} improves performance for harder problems. Figure 3 demonstrates that S_{DSV} shows very high performance in the left lower corners (s=0.025, d=0.05), representing better discerning power for gene expression profiles with higher similarities. Moreover, S_{BSV} is not merely a weighted summation of S_{Wang} and S_{DSV}. While S_{DSV} applies the commonality (or conjunction) of descendants, its weight, β, applies the union of descendants. S_{BSV} does not become S_{Wang} even when S_{DSV} equals zero.

GO hierarchy has a large proportion of terminal leaves on which S_{DSV} has only limited impact. However, more useful real-world tasks involve genes, pathways, disease,[24] and medical concepts,[22] having rich GO annotations rather than terms

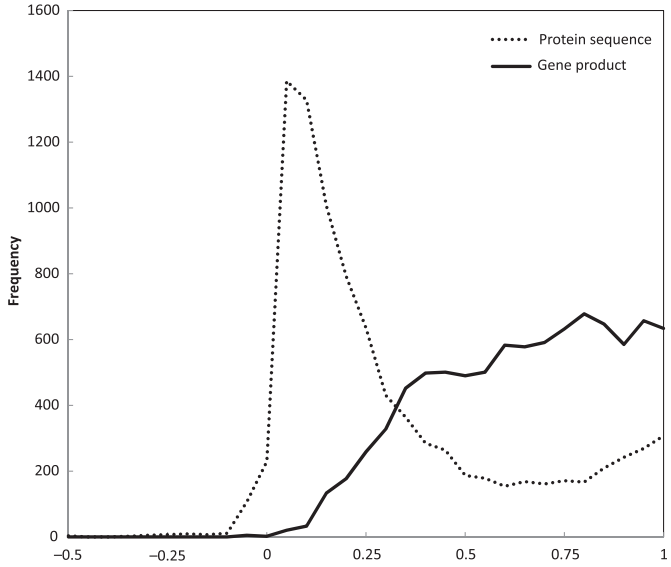

Figure 5 Distribution of the differential proportions between ascending and bidirectional semantic similarity measures for gene pairs. The proportions of changes between two measures are computed as $(S_{Wang}-S_{BSV})/S_{Wang}$. Almost all semantic similarities of gene (or protein) pairs are affected by introducing bidirectional semantic similarity measure. Highly selected similar protein pairs (dotted line) by the best BLAST sequence match showed relatively smaller degree of change (0.29 in average) than randomly selected gene pairs (solid line, 0.63 in average).

themselves. Figure 5 demonstrates the magnitude of the impact of introducing descending similarity measure to gene pair semantic comparison studies.

Intentional definition (or coactive definition) works from more general to more specific, which is informally called a 'top-down' approach. Extensional definition (or denotative definition) works the other way (ie, 'bottom-up'), moving from specific observations to broader generalizations. Current methods of determining semantic similarities are limited in that they are applying 'top-down' approaches only. We propose a novel method that applies bidirectional measures of semantic similarity, considering the entire DAG structure including both the upper (ancestral) and the lower (descendant) parts.

Contributors SJB and CHP conducted the study and wrote the manuscript. WY raised the problem of semantic measures in the beginning and HJS modeled the problem in a schematic way. JHK evoked the initial conception of the study and supervised the study. The first two authors contributed equally to this work.

Funding This research was supported by the basic science research programme through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0028631). CHP's educational grant was supported in part by the Korea Communications Commission, Korea, under the R&D programme supervised by the Korea Communications Agency (KCA-2011-11911-01108).

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Cilibrasi R,** Vitányi PM. The google similarity distance. *IEEE Trans Knowl Data Eng* 2006;**19**:370—83.
2. **Aseervatham S,** Bennani Y. Semi-structured document categorization with a semantic kernel. *Pattern Recognit* 2009;**42**:2067—76.
3. **Lee J,** Kim M, Lee Y. Information retrieval based on conceptual distance in is-a hierarchies. *J Doc* 1993;**49**:188—207.
4. **Ratprasartporn N,** Po J, Cakmak A, *et al*. Context-based literature digital collection search. *Int J Very Large Data Bases* 2009;**18**:277—301.
5. **Lord PW,** Stevens RD, Brass A, *et al*. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 2003;**19**:1275—83.
6. **Lord PW,** Stevens RD, Brass A, *et al*. Semantic similarity measures as tools for exploring the gene ontology. *Pac Symp Biocomput* 2003:601—12.
7. **Wang JZ,** Du Z, Payattakool R, *et al*. A new method to measure the semantic similarity of GO terms. *Bioinformatics* 2007;**23**:1274—81.
8. **Jensen LJ,** Bork P. Ontologies in quantitative biology: a basis for comparison, integration, and discovery. *PLoS Biol* 2010;**8**:e1000374.
9. **Harris MA,** Clark J, Ireland A, *et al*. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004;**32**:D258—61.
10. **Schlicker A,** Domingues FS, Rahnenführer J, *et al*. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* 2006;**7**:302.
11. **Tao Y,** Sam L, Li J, *et al*. Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics* 2007;**23**:i529—38.
12. **Wu X,** Zhu L, Guo J, *et al*. Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Res* 2006;**34**:2137—50.
13. **Resnik P.** Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*; 20—25 August 1995, Montreal, Quebec, Canada, 1995:448—53.
14. **Resnik P.** Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J Artif Intell Res* 1999;**11**:95—130.
15. **Guo X,** Liu R, Shriver CD, *et al*. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics* 2006;**22**:967—73.
16. **Kanehisa M,** Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27—30.
17. **Sevilla JL,** Segura V, Podhorski A, *et al*. Correlation between gene expression and GO semantic similarity. *IEEE/ACM Trans Comput Biol Bioinform* 2005;**2**:330—8.
18. **Barrett T,** Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol* 2006;**411**:352—69.
19. *BioCarta*. http://www.biocarta.com/ (accessed 1 Apr 2010).
20. **Lin D.** An information-theoretic definition of similarity. In: Shavlik JW, ed. *Proceedings of the Fifteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 296—304. doi:10.1.1.55.1832
21. **Jiang JJ,** Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of the International Conference on Research in Computational Linguistics ROCLING X*. Taiwan, 1998.
22. **Pedersen T,** Pakhomov SV, Patwardhan S, *et al*. Measures of semantic similarity and relatedness in the biomedical domain. *J Biomed Inform* 2007;**40**:288—99.
23. **Pesquita C,** Faria D, Falcão AO, *et al*. Semantic similarity in biomedical ontologies. *PLoS Comput Biol* 2009;**5**:e1000443.
24. **Mathur S,** Dinakarpandian D. Finding disease similarity based on implicit semantic similarity. *J Biomed Inform*. Published Online First: 7 December 2011. http://dx.doi.org/10.1016/j.jbi.2011.11.017

Bi-directional semantic similarity for gene ontology to optimize biological and clinical analyses

Sang Jay Bien, Chan Hee Park, Hae Jin Shim, et al.

J Am Med Inform Assoc published online February 28, 2012
doi: 10.1136/amiajnl-2011-000659

Notes