BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 10 2012, pages 1383–1389
doi:10.1093/bioinformatics/bts129

Databases and ontologies

Advance Access publication April 19, 2012

Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty

Haixuan Yang, Tamás Nepusz and Alberto Paccanaro*

Department of Computer Science and Centre for Systems and Synthetic Biology, Royal Holloway, University of London, Egham, TW20 0EX, UK

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Several measures have been recently proposed for quantifying the functional similarity between gene products according to well-structured controlled vocabularies where biological terms are organized in a tree or in a directed acyclic graph (DAG) structure. However, existing semantic similarity measures ignore two important facts. First, when calculating the similarity between two terms, they disregard the descendants of these terms. While this makes no difference when the ontology is a tree, we shall show that it has important consequences when the ontology is a DAG—this is the case, for example, with the Gene Ontology (GO). Second, existing similarity measures do not model the inherent uncertainty which comes from the fact that our current knowledge of the gene annotation and of the ontology structure is incomplete. Here, we propose a novel approach based on downward random walks that can be used to improve any of the existing similarity measures to exhibit these two properties. The approach is computationally efficient—random walks do not need to be simulated as we provide formulas to calculate their stationary distributions.

Results: To show that our approach can potentially improve any semantic similarity measure, we test it on six different semantic similarity measures: three commonly used measures by Resnik (1999), Lin (1998), and Jiang and Conrath (1997); and three recently proposed measures: simUI, simGIC by Pesquita *et al.* (2008); GraSM by Couto *et al.* (2007); and Couto and Silva (2011). We applied these improved measures to the GO annotations of the yeast *Saccharomyces cerevisiae*, and tested how they correlate with sequence similarity, mRNA co-expression and protein–protein interaction data. Our results consistently show that the use of downward random walks leads to more reliable similarity measures.

Availability: We have developed a suite of tools that implement existing semantic similarity measures and our improved measures based on random walks. The tools are implemented in Matlab and are freely available from: http://www.paccanarolab.org/papers/GOsim/

Contact: alberto@cs.rhul.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 28, 2011; revised on February 29, 2012; accepted on March 11, 2012

1 INTRODUCTION

The introduction of ontologies for gene functional annotation allows us to compare genes by quantifying the similarity of the terms with which they are annotated. These comparisons are important as they contribute to the inference of functional relationships between gene products by providing a perspective that complements both experimental information and sequence-based approaches.

Table 1. Statistics of GO terms and yeast annotation

	Overlap in GO		Uncertainty in yeast	
	Multiple parents	Single parent	Terms	Non-leaf
BP	13 517	6349	3322	898
CC	1765	1005	754	203
MF	1424	7475	1857	366

BP, biological process; CC, cellular component; MF, molecular function. First column: number of GO terms with more than one parent; second column: number of GO terms with only one parent; third column: total number of GO terms to which yeast gene are annotated; fourth column: number of GO terms to which some yeast genes are annotated while not being annotated to any of their children. Only non-empty nodes were considered. Annotations with evidence codes IEA, NR, ND and IC were excluded.

Standard ontologies usually have a structure that can be modelled by a rooted and oriented tree [e.g. MIPS (Mewes *et al*., 2006), GenProtEC (Riley and Space, 1996)], or more generally by a directed acyclic graph (DAG), like the Gene Ontology (GO; Ashburner *et al*., 2000) which has become a standard and is the focus of this article. In general, given two terms, comparisons are not straightforward due to the complex structure of the ontologies. Lord *et al*. (2003), Sevilla *et al*. (2005) and Schlicker *et al*. (2006) discuss in detail the issues related to such comparisons, here we only mention a few of them. For instance, comparisons should take into account the position of the terms in the ontology structure as terms in higher levels (i.e. closer to a root term) are less specific and likely to be less informative. At the same time, the depth of the term may not be an exact indicator of its specificity as some edges (relations) in the ontology may cover a larger conceptual distance, whereas others may cover shorter distances. Calculating the depth of a term becomes even more problematic in DAG structures, as a term may have multiple parent terms and thus multiple paths of different lengths leading to the root term (see column 1 in Table 1). Finally, ontologies and annotations are constantly refined and one should not neglect the possibility of new annotations or new terms being added later to the ontologies.

2 MOTIVATION FOR THIS WORK

Several semantic similarity measures have been proposed that have proved to be useful tools in a variety of biological problems

*To whom correspondence should be addressed.

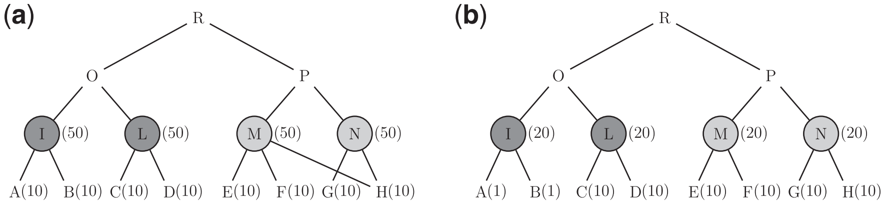

Fig. 1. Illustration of two properties that should hold for a semantic similarity measure. Nodes are represented by a letter, the number of genes associated to some of the nodes is shown in parentheses next to the nodes. In both (**a**) and (**b**), the semantic similarity between M and N should be greater than the similarity between I and L because: in (a) M and N have a higher proportion of genes in common than I and L; in (b) M has less uncertainty than I.

(Guo *et al.*, 2006; Jain and Bader, 2010; Othman *et al.*, 2008). However there are two important aspects that, to the best of our knowledge, present state-of-the-art methods fail to take into account. First, these measures compare two terms by examining only the part of the hierarchy that is above them and do not consider the hierarchy below the terms being examined—note that here and in the rest of the article we are representing the Gene Ontology as an uprooted DAG, where the most general terms are located at the top and their (more specific) children are positioned below them. Second, existing measures do not model the uncertainty in the ontology structure and the annotation. Possibly new functional terms are currently missing from the existing ontology which may affect the semantic similarity of the terms being compared. At the same time, genes currently annotated to non-leaf terms will eventually be moved to leaf nodes once enough experimental evidence will become available—column 4 of Table 1 quantifies this uncertainty in annotation for yeast.

We shall now illustrate, by looking at the two toy examples in Figure 1, the importance of each of these two aspects which constitute the motivation for our work. Note that here and in the rest of the article we shall only consider GO terms to which at least one gene product is annotated.

The ontology structure beneath the terms under consideration. Figure 1a shows a small example where we coloured two pairs of terms (I,L) and (M,N) for which the ontology structure is identical above them, but different below. In fact the children of term I and term L are completely disjoint, whereas terms M and N have a common child. Without loss of generality, we can assume that the sets of genes annotated to the leaves A,B,C,D,E,F,G,H are disjoint and that the set of genes assigned to I,L,M,N but not to their children are also disjoint. Due to the ontology structure, gene products annotated to H are also annotated to both M and N. Therefore, nodes M and N are more similar than nodes I and L because, given the structure of the hierarchy, they share 10 gene products. Ideally, the semantic similarity between M and N should be higher than that between I and L.

In other words, given the structure of the ontology, the similarity between M and N is affected by the number of genes annotated to H: the higher this number, the higher the similarity between them. Therefore a well-defined measure of semantic similarity should take into account the number of genes in H, which is located in the part of the ontology below the nodes under consideration. That is, the semantic similarity between two terms depends not only on their common ancestors, but also on their common descendants.

The uncertainty in the ontology structure and current annotation. We shall explain the role of uncertainty in semantic similarity using the example in Figure 1b. As before, without loss of generality we can assume that the sets of genes annotated to the leaves A,B,C,D,E,F,G,H are disjoint and that the set of genes assigned to I,L,M,N but not to their children are also disjoint.

Here the ontology structure above and below the pairs of terms (I,L) and (M,N) is identical. The difference between pairs (I,L) and (M,N) now lies in the number of gene products annotated to A, B, E and F. Ten genes are annotated to each of E and F, thus making node M completely determined. On the other hand, there is a great uncertainty about node I, as only two genes are accounted for in A and B. This means that relatively little is known about node I: genes for this functional category are currently not well characterized, possibly some of its descendant nodes have not been characterized yet [See the Annotation Conventions (http://www.geneontology.org/GO.annotation.conventions.shtml) for a better understanding]. This makes the similarity between the pair I and L much uncertain, and therefore we would like to assign a greater semantic similarity to nodes M and N which are instead completely determined. Further discussions on the role of uncertainty are given in the Supplementary Material where we also show the importance of including uncertainty by comparing results obtained with and without taking uncertainty into account.

Nodes with multiple parents and genes annotated to non-leaf terms appear very prominently in GO (Table 1). This provides us with a strong motivation for developing methods which can take into account the knowledge about the descendants of the terms being considered and the uncertainty in the annotation and ontology structure.

In this article, we shall describe how both these factors can be quantified using downward random walks. This measure, which we call the random walk contribution (RWC) can be integrated with any standard semantic similarity measure, which we call host similarity measure (HSM), to yield an integrated similarity measure (ISM) that takes into account the whole ontology structure. In other words our random walk similarity measure is a kind of ‘add on’ to one’s favourite underlying similarity measure. The random walk calculations can be done very efficiently—for the RWC we only need to calculate the random walk stationary distribution probabilities which can be easily obtained from the transition equations presented in the sequel. We shall show results obtained by integrating our random walk measure onto six commonly used semantic similarity measures. These experiments will quantify the advantage of including into semantic similarity calculations the ontology structure beneath the terms under consideration and the uncertainty in the ontology structure and annotation.

3 RELATED WORK

Several authors have provided methods for quantifying the semantic similarity between terms in an ontology. These methods can roughly be classified into three categories (Pesquita *et al.*, 2009): (i) edge-based methods which use the edges (relations) in the ontology and their types as the primary data source; (ii) node-based methods, in which the main data sources are the terms, their properties and the number of entities annotated to the given terms; and (iii) hybrid methods which exploit the properties for both edges and nodes. Edge-based semantic similarity measures are defined as some function of the length of the paths linking the terms being considered and the global position of the terms themselves within the ontology structure (Li *et al*., 2003; Rada *et al*., 1989). Node-based methods recognize the fact that the terms in the ontology are not equivalent: some terms have more associated entities whereas others have less, and the number of entities associated to a term may give an indicator of the term importance or specificity. Therefore, these methods generally define the semantic similarity between two given terms as some function of the information content of their ancestors and optionally of these terms themselves [e.g. (Couto *et al*., 2005; Jiang and Conrath, 1997; Lin, 1998; Resnik, 1999; Schlicker *et al*., 2006; Yu *et al*., 2007)]. Among these, the methods of Resnik (1999), Jiang and Conrath (1997) and Lin (1998) received much attention in the past few years. Given two terms, these measures are constituted by (a normalization of) the information content of their most informative common ancestor. Other more recent approaches consider more than one ancestor, such as simUI and simGIC (Pesquita *et al*., 2008) or GraSM (Couto and Silva, 2011; Couto *et al*., 2007). For more details on different semantic similarity measures, see (Pesquita *et al*., 2009).

There has been indeed much debate regarding which measure should be preferred over the others for biological problems. However, no clear consensus has been reached and it seems that different measures are best suited for different domains. For this reason our method, where the random walk similarity measure is 'added on' to a given underlying HSM, is suitable for different applications.

In this article, we shall test the efficacy of our procedure on six of these semantic similarity measures which we shall refer to as: Resnik (Resnik, 1999); Lin (Lin, 1998); Jiang (Jiang and Conrath, 1997); simUI and simGIC (Pesquita *et al*., 2008); GraSM (Couto and Silva, 2011; Couto *et al*., 2007). These measures are summarized in the Supplementary Material.

4 METHODS

We begin by giving an intuitive description of our method, followed by a more formal definition. Our basic assumption is that a HSM can generally be considered accurate when comparing two leaf terms in GO—leaf terms have no further children that these measures would ignore. On the other hand, we propose that the semantic similarity of two non-leaf terms (or one leaf term and one non-leaf term) would consists of two components: (i) a similarity that depends on the ancestors of the terms being considered and (ii) a similarity that depends on the (possibly shared) descendant terms and their similarity scores.

To understand how our method accounts for these descendant terms, let us consider Figure 2a which shows a simple hypothetical ontology consisting of seven terms. When a gene is annotated to a term, it is also annotated to all of its parents; therefore we know that 10 out of the 50 genes annotated to C are *not* annotated to any of C's descendants (F and G). We call these genes *partially annotated* as they cannot currently be assigned to F or G or to a yet uncharted new GO term (situated below C) due to the limitation of our current biological knowledge. Given a randomly selected gene annotated to C, this gene is annotated to F with probability 0.6 and to G with probability 0.2. The remaining probability of 0.2 corresponds to the event that the gene is only partially annotated. Therefore, when one compares a gene annotated to C to some other gene annotated to a term X, there is a 60% chance that one is comparing a gene annotated to F and a 20% chance that one is comparing a gene annotated to G. The semantic similarity between C and X thus may be approximated by weighting the semantic similarities between the pairs (F,X) and (G,X) by the factors 0.6 and 0.2, respectively. In other words, our idea is to decompose the semantic similarity of the two terms being compared into a weighted sum of the semantic similarities of their descendant leaf terms, and in this way we take into account both the ontology structure beneath the terms under consideration and the uncertainty in the current annotation. Note that, due to the possibility of partially annotated genes assigned to a yet uncharted new GO term, the weights assigned to the children of a node do not necessarily sum to 1, thus accounting for the uncertainty in the ontology structure.

In order to obtain the weights, we need to estimate the probability of a gene annotated to a general non-leaf term T to actually belong to an arbitrary leaf term L. This is done by conducting downward random walks on the ontology structure: we start a random walker from T, let it move downward towards leaf terms by following the edges and we observe the fraction of walks that terminate in L. Note that, considering the possibility of partially annotated genes assigned to a yet uncharted new GO term when calculating these probabilities amounts to introducing some fictional extra nodes in the ontology structure—this is our model for the uncertainty in the ontology structure. At the same time, the higher the uncertainty of a node, the smaller will be the fraction of genes assigned to its descendants—this is our model for the uncertainty in the annotation. Formally, the method consists of four major steps as follows.

Step 1: Initialization. Let N_v be the number of genes annotated to node v, and N_v^* the number of genes annotated to v but not to any of its children. For each non-leaf node v, an extra 'unknown' child node U_v is added to the ontology graph (Fig. 2b). An edge is then added from v to U_v and is labelled as follows:

$$P(v \to U_v) = \frac{N_v^*}{N_v}. \tag{1}$$

Each parent–child edge $v \to c$ in the ontology graph is then labelled by a transition probability:

$$P(v \to c) = (1 - P(v \to U_v)) \frac{N_c}{\sum_{u:\exists v \to u} N_u}. \tag{2}$$

The above two equations ensure that the transition probabilities define a downward random walk on the graph as each edge points downwards in the tree and the transition probabilities of the outgoing edges of a node add up to 1. Note, how N_v^* in Equation (1) quantifies the amount of the uncertainty in the annotation of node v. According to Equation (1), $N_v^* \neq 0$ implies a non-zero transition probability to the unknown child node U_v. This affects the transition probability $P(v \to c)$ in Equation (2): the larger N_v^*, the smaller the transition probability $P(v \to c)$.

Step 2: Downward random walk. In this step, a downward random walk is conducted from each non-leaf node v_0 to determine its relationship to the leaf nodes. Let $W_t^{v_0}(v)$ denote the probability of the random walker being at node v after t steps when it started from v_0. Initially, $W_0^{v_0}(v) = 1$ if $v = v_0$ and zero otherwise. The probabilities at step $t+1$ can be determined based on the probabilities at step t given the transition probabilities. The exact rules are different for leaf and non-leaf nodes. We know that if we were at a leaf node in step t, we will stay there in step $t+1$. Therefore, the probability of being at some leaf node l in step $t+1$ is equal to the probability of being there in step t plus the probability of arriving there from one of its parents:

$$W_{t+1}^{v_0}(l) = W_t^{v_0}(l) + \sum_{v:\exists v \to l} W_t^{v_0}(v) P(v \to l). \tag{3}$$

Similarly, the probability of being at a non-leaf node v at step $t+1$ is equal to the probability of being at one of its parents q at step t, multiplied by the probability that we have chosen edge $q \to v$ to arrive at v:

$$W_{t+1}^{v_0}(v) = \sum_{q:\exists q \to v} W_t^{v_0}(q) P(q \to v). \tag{4}$$

(for the root node the summation becomes empty and we set its value to zero). Since, we are always stepping downward from a non-leaf node towards

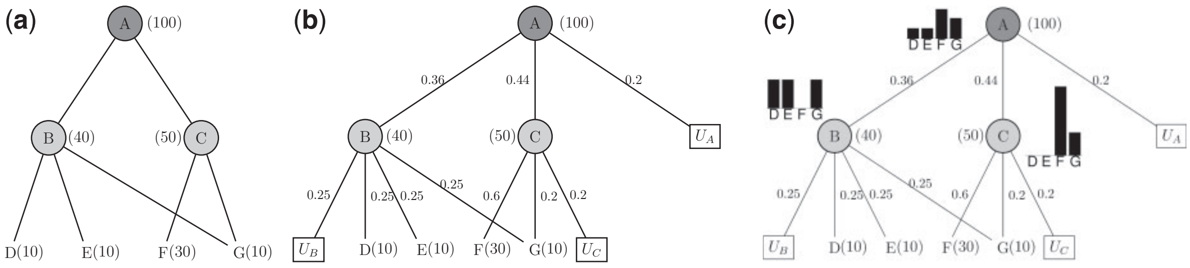

Fig. 2. Illustration of the proposed measure on a hypothetical ontology. (**a**) The ontology as a DAG. Nodes are represented by a letter, non-leaf nodes are circled; the number of genes associated to each node is shown in parentheses next to the node. (**b**) The ontology graph extended by the unknown nodes as described in Step 1. The transition probabilities shown as edge labels were calculated using Equations (1) and (2). (**c**) Histograms of the random walk distributions over the leaf nodes for nodes *A*, *B* and *C*.

one of its children, we will always end up in one of the leaf nodes sooner or later. Once we entered a leaf node, there is no escape, therefore the stationary distribution $W_\infty^{v_0}$ of the random walk will attain zero probabilities on non-leaf nodes and some non-zero probabilities on leaf nodes (including the newly introduced unknown nodes). This distribution depends solely on v_0 and it contains the information about the relationship between v_0 and the leaf terms of the ontology.

Step 3: Calculating the RWC. Given two nodes, v_0 and v_1, the RWC can be calculated based on a given host measure HSM and their stationary distributions $W_\infty^{v_0}$ and $W_\infty^{v_1}$.

As we described earlier, we assume that the similarity between two leaf nodes is given by the HSM. The RWC between two non-leaf terms is the HSM between all their leaf descendants weighted by their probabilities. These probabilities are given by the stationary distribution of the random walks started at these non-leaf nodes. The two random walks are assumed to be independent, therefore the RWC for nodes v_0 and v_1

$$\mathrm{RWC}(v_0,v_1)=\sum_{i,j\in\mathcal{L}} W_\infty^{v_0}(i)W_\infty^{v_1}(j)\mathrm{HSM}(i,j) \qquad (5)$$

(where $\mathcal{L}$ is the set of all leaf nodes except the newly added unknown nodes) gives the expected semantic similarity according to the host measure between the descendants of v_0 and v_1, assuming that those descendants are reached according to the probabilities in $W_\infty^{v_0}$ and $W_\infty^{v_1}$.

Note, how the RWC takes into account both the ontology structure beneath the terms under consideration and the uncertainty in the ontology structure and annotation. In fact, given two terms, the existence of common descendants will influence the RWC: the greater the number of common descendants, the more similar the distributions obtained on their descendant leaves and as a result, the greater the contribution of the RWC to their similarity score. At the same time, uncertainty affects the RWC as follows. Since the transition probabilities encode the uncertainty in a given node *T* itself and each of *T*'s descendants, the uncertainty information is finally transmitted to the distribution on leaves as the random walker starting from *T* moves down. Therefore the greater the uncertainty in a given node *T* itself and each of its descendants, the smaller will be the total probability mass accumulated on its descendant leaves, and consequently the smaller the contribution of the RWC to the similarity score.

Step 4: Combining the HSM and the RWC. The RWC now needs to be combined with the HSM. In fact, the RWC only considers the hierarchy below the terms being examined while the HSM only accounts for the hierarchy above the given terms. By combining the two, we are able to consider both the higher parts and the lower parts of the hierarchy relative to the given terms. The ISM is then as follows:

$$\mathrm{ISM}(v_0,v_1)=\tfrac{1}{2}\left(\mathrm{RWC}(v_0,v_1)+\mathrm{HSM}(v_0,v_1)\right). \qquad (6)$$

We shall now clarify our method through an example. In Figure 2a, there are 100 genes annotated to the terms of the ontology, the exact counts being shown in parentheses next to the nodes. The ontology extended by the unknown nodes along with the calculated edge transition probabilities are shown in Figure 2b. The stationary distributions (W_∞) can be calculated in a bottom-up manner for each non-leaf node (Fig. 2b). The probability of a random walk starting from node *B* and ending in node *D*, *E*, *F* and *G* is $(0.25, 0.25, 0, 0.25)$, respectively; a random walk starting from node *C* ends up in the same leaf nodes with probabilities $(0, 0, 0.6, 0.2)$. The stationary distribution corresponding to starting node *A* then follows by recognizing that we step to node *B* with probability 0.36 and node *C* with probability 0.44 in the first step and then the random walks are the same as in the above cases, hence the final stationary distribution follows by taking $0.36\times(0.25, 0.25, 0, 0.25)+0.44\times(0, 0, 0.6, 0.2)=(0.09, 0.09, 0.26, 0.18)$. These distributions do not add up to 1 as the remaining probabilities are leaked to the unknown leaves.

Table 2. Resnik's measure (HSM_Resnik), together with RWC and ISM calculated using Resnik's measure as HSM (RWC_Resnik and ISM_Resnik) for some pairs of nodes in the example in Fig. 2

	HSM_Resnik	RWC_Resnik	ISM_Resnik
A-A	0	0.333	0.166
B-B	0.916	0.775	0.846
B-C	0	0.311	0.156
C-C	0.693	0.692	0.693
C-E	0	0.183	0.092
D-D	2.302	2.302	2.302
D-E	0.916	0.916	0.916
D-F	0	0	0
D-G	0.916	0.916	0.916
E-E	2.302	2.302	2.302
E-F	0	0	0
E-G	0.916	0.916	0.916
F-F	1.204	1.204	1.204
F-G	0.693	0.693	0.693
G-G	2.302	2.302	2.302

Let us now calculate the semantic similarity between node *B* and *C* using Resnik's measure as the HSM (Fig. 2c). The RWC is obtained by taking the expected HSM between the pairs of leaf nodes in which two random walkers will end up if the first walker is started from node *B* and the second one is started from node *C*. For instance, the probability of the first random walker ending up in node *E* and of the second one in node *G* is $0.25\times0.2=0.05$, since the two random walkers are independent. This has to be multiplied by the HSM of node *E* and *G* (0.916) yielding the contribution of the pair *E*–*G* to the overall RWC of *B* and *C*: $0.916\times0.05=0.0458$. Such contributions have to be calculated for every pair of leaves, and the sum of these contributions gives us the RWC of node *B* and *C*, which is 0.311. Finally, in Step 4, the

RWC is combined with the HSM. Since the host measure happens to be zero for node B and C if we are using Resnik's measure, the final ISM will be equal to half of the random walk similarity.

It is worthwhile to compare Resnik's original similarity measure with our ISM (Table 2). The two similarity measures are equivalent for leaf nodes, since we always trust the HSM unconditionally for leaves. However, the combined measure is larger in those cases, in which there are overlaps between the descendants of the nodes being considered. For instance, the ISM yields a similarity of 0.092 for nodes C and E since the measure assumes that nodes annotated by C are also annotated by either F or G with some probability, and the HSM is not zero for the E–G pair. Similar reasoning applies for the similarity of B–C: the increase is due to the fact that node G is a child of both B and C. However, sometimes the ISM is smaller than the HSM, as in the case of the semantic similarity of B with itself, where the combined measure considers that there is a fairly wide branching at node B and genes annotated by B may end up in different branches.

It is important to point out that the algorithm actually explores the ontology in more than just the upward and downward directions. For instance, in Figure 2, the fact that E and G are siblings increases the similarity between E and C, as C is a parent of G. Since E and C do not belong to either one's ancestry, it follows that the measure is able to take more from the ontology than the set of ancestors and descendants.

The calculations for the random walk are computationally inexpensive. The random walks from each non-leaf node to all the leaves can easily be calculated using Equations (3) and (4), which amount to k matrix-vector multiplications where k is an integer number smaller than the maximum depth of the tree. Let us assume that k is proportional to $\log(n)$, where n is the number of nodes. If the cost of each multiplication is $O(m)$, where m is the number of non-zero elements in the transition matrix, then the overall cost for a random walk for all the non-leaf nodes is $O(nm\log(n))$—which is equivalent to $O(n^2\log(n))$ since m is proportional to n.

The algorithm description given in this section details the case in which the HSM is defined over pairs of terms (e.g. Resnik). A slight modification of the above description extends the algorithm for HSMs defined over pairs of genes (e.g. simUI or simGIC). The basic idea of the extension is that the random walkers will start from each gene instead of each term. In the first step, a random walker at gene i will jump randomly to one of the terms. This extension is detailed in the Supplementary Material where the algorithm is also represented in matrix form.

5 RESULTS

In order to evaluate our method we need to show that standard similarity measures are improved when integrated with RWCs. Therefore, we chose the six well-known similarity measures of Resnik, Jiang, Lin, GraSM, simUI and simGIC, and we compared their performance with the performance of the ISM which used these as host measures—we shall indicate these as ISM_Resnik, ISM_Jiang and so on, respectively. The performance was compared using sequence similarity data, co-expression data derived from microarray and protein–protein interaction data. All experiments were performed on the GO annotations of the yeast *Saccharomyces cerevisiae*. The versions of the GO and of the annotation file dated November 12, 2010, and November 11, 2010, respectively. In the results reported here, annotations with evidence code IEA, NR, ND and IC were excluded. Also, when comparing two gene products which were annotated to several GO terms, their similarity was taken as the maximum of their pairwise similarities.[1] Results using the best-match average approach (Pesquita *et al.*, 2009; Schlicker *et al.*, 2006) are shown in Supplementary Material. Following previous authors, [e.g. (Pesquita *et al.*, 2009; Schlicker *et al.*, 2006)], in this article we used only the 'is-a' and 'part-of' relationships between GO terms.

5.1 Comparison with sequence similarity

A technique which has been used by several authors to compare the performance of different semantic similarity measures is to test how well these measures correlate with sequence similarities (Lord *et al.*, 2003). The idea is that semantic similarity should correlate, to some extent, with sequence similarity.

We used the protein similarity scores published by the SGD project (http://downloads.yeastgenome.org) which were calculated using the Smith–Waterman algorithm (Smith and Waterman, 1981) for every pair of yeast ORF protein sequences. Figure 3 shows the correlation between these scores and the different semantic similarity measures on each of three GO DAGs separately. We can see that our proposed method greatly improves all the three host semantic similarity measures—the value of the correlation of the ISMs improves the corresponding HSMs for all the GO DAGs in all cases except two.

We note that some authors have evaluated their semantic similarity measure by first binning the sequences and semantic similarity scores, and then calculating the correlation between the bins [e.g. (Lord *et al.*, 2003)]. Our experiments employing this binning procedure gave results similar to the ones shown here and they are reported in the Supplementary Material.

5.2 Comparison with gene co-expression patterns

Another method which has been used by several authors to compare the performance of different semantic similarity measures is by testing how well these measures correlate with gene expression similarities (Sevilla *et al.*, 2005; Wang *et al.*, 2004). The idea behind this is that since genes involved in the same process tend to exhibit similar expression patterns, we could expect good semantic similarity measures calculated on the GO biological process ontology to be correlated with the expression similarity.

For our experiments, we used the yeast cell cycle data from (Spellman *et al.*, 1998), and tested our procedure both on the four independent experiments (α factor, CDC15, CDC28 and elutriation) and on a combined dataset obtained by concatenating these four microarrays.

Following the approach of previous authors (Sevilla *et al.*, 2005) we first measured the gene expression similarity using the Pearson's correlation coefficient between the gene profiles. We then calculated the correspondence between such expression similarity and the semantic similarity again using the Pearson's correlation coefficient. Results are shown in Figure 4.

We can see that our approach, which combines a given HSM with the random walk measure, improves the correlation between co-expression and semantic similarity in all the cases except one. Experiments using the binning procedure also gave the same conclusion (see the Supplementary Material).

5.3 Comparison with protein–protein interactions

Finally, we compared the performance of our ISMs by investigating the relationship between semantic similarity and protein–protein interactions. Our idea was to formulate this as a classification

[1] We also tried different settings, e.g. including IC, taking the average value of the similarity between groups of GO terms, and the results obtained were equivalent to the ones presented here.

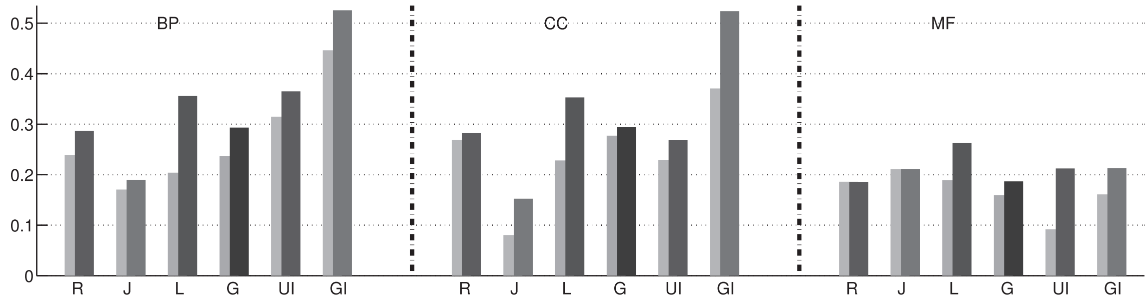

Fig. 3. Correlations between sequence similarity and different semantic similarity measures in yeast, shown separately for the three DAGs of the GO (BP, biological process; CC, cellular component; MF, molecular function). For a given HSM (R, Resnik; J, Jiang; L, Lin; G, GraSM; UI, simUI; GI, simGIC) the lighter colour shade represents the original HSM and the darker colour the corresponding ISM.

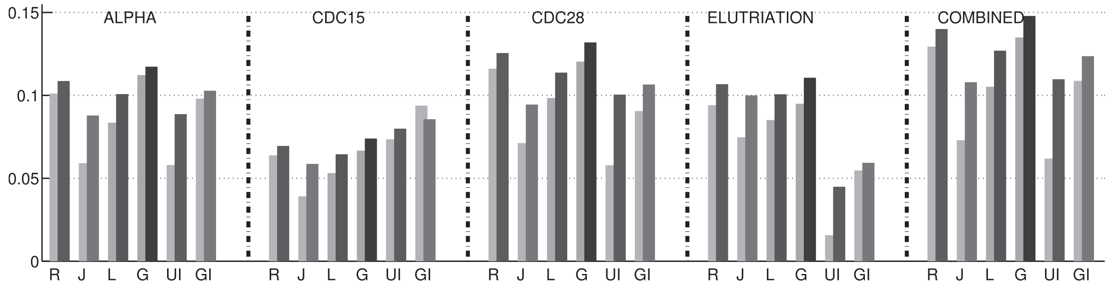

Fig. 4. Correlations between gene co-expression scores and different semantic similarity measures on the Biological Process DAG of the GO. Measures are calculated for the four individual cell cycle microarray experiments for yeast (α factor, CDC15, CDC28 and elutriation) as well as for the combined dataset, which is obtained by concatenating the four experiments. Notation and colours are the same as in Figure 3.

problem and to check how well the different semantic similarity measures perform at predicting protein–protein interactions.

We built a gold standard dataset of interacting and non-interacting pairs of proteins (positive and negative pairs) taken from all possible yeast protein pairs. Following the approach of previous authors [e.g. (Krogan *et al*., 2006)] positive pairs were obtained from the MIPS protein complex database (Mewes *et al*., 2006), whereas negative pairs were constituted by pairs of proteins known to have different subcellular localization. The final gold standard set contained 9324 positive and 2 341 019 negative pairs.

Results of the prediction were evaluated using receiver operating characteristic (ROC) curves—the best semantic measures are the ones for which the ROC curve steeply rises towards the top left corner and the area under the curve (AUC) is greatest. As noticed by previous authors [e.g. (Collins *et al*., 2007)], due to the imbalance between positive and negative examples, the relevant part of the ROC curve is on the far left end of the X-axis. Therefore, we restrict our analysis to this part of the ROC curve only—following the setting of (Collins *et al*., 2007), we used the part of the ROC curves where the false positive rate (FPR) is $\leq$ to 0.002. Figure 5 shows the AUC scores, and Figure 6 shows the ROC curves for the cellular component (CC). We can see that our approach always improves the reliability of all tested semantic similarity measures when predicting protein interactions using CC and that it is better in the majority of the cases when using biological process or molecular function.

6 DISCUSSION

Existing semantic similarity measures have two important limitations. First, these methods assess the similarity between two terms by examining only the part of the hierarchy that is above these two terms while they do not consider the hierarchy below the terms being examined. Second, existing measures do not model the uncertainty in the GO structure and existing gene annotation. In this article, we proposed a novel approach for measuring the semantic similarity among terms on DAGs. The method is based on downward random walks and it can be used to improve existing semantic similarity measures in order to overcome the above two limitations. We extensively tested our approach by using three different perspectives based on gene expression data, sequence similarity data and protein–protein interaction data. Results consistently show that semantic similarity measures are improved when they are combined with downward random walks.

A few aspects of our method should be further investigated. For example, we are currently mixing HSM and RWC in equal proportion, while one could optimize the balance between the two components of the ISM for different problems. Also, for ISM_simUIC and ISM_simGIC instead of using a uniform jump to go from gene to GO terms one could attempt using a non-uniform jump which could be weighted, for example, by the information content.

ACKNOWLEDGEMENT

We thank the anonymous reviewers for their valuable suggestions, and Prajwal Bhat, Alfonso E. Romero, Emilio Ferrara, Horacio Caniza Vierci and Sandra Smieszek for useful discussions.

Funding: Biotechnology and Biological Sciences Research Council (BBSRC) New Investigator Grant (grant number BB/F00964X/1 to

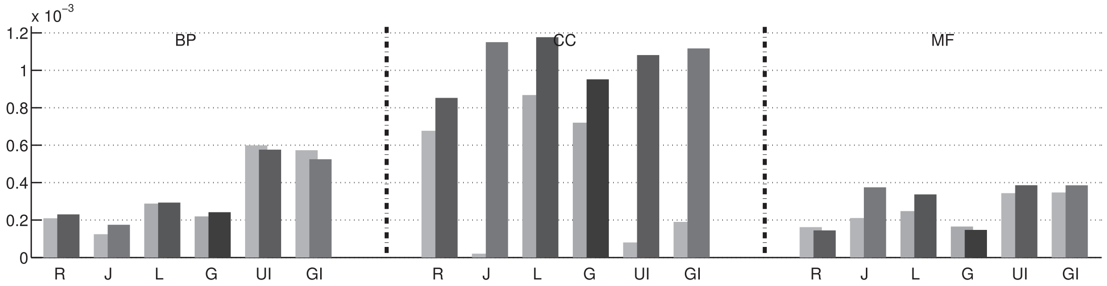

Fig. 5. AUC scores for FPR ≤ 0.002 comparing the different semantic similarity measures on the three DAGs of the GO for predicting protein–protein interactions. Notation and colours are the same as in Figure 3.

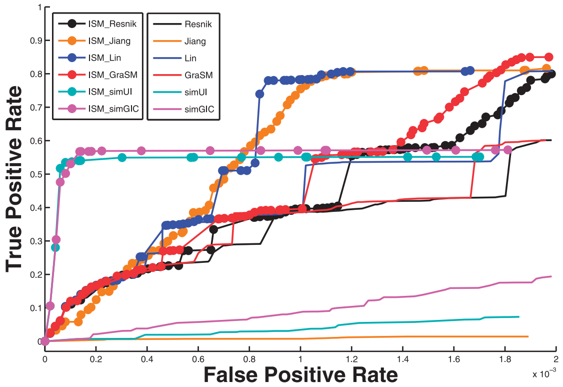

Fig. 6. ROC curves comparing the different semantic similarity measures on the Cellular Component DAG of the GO for predicting protein-protein interactions.

A.P.) and Newton International Fellowship Scheme of the Royal Society (grant number NF080750 to T.N.).

Conflicts of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Collins,S.R. *et al.* (2007) Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol. Cell Proteomics*, **6**, 439–450.

Couto,F. and Silva,M. (2011) Disjunctive shared information between ontology concepts: application to gene ontology. *J. Biomed. Semantics*, **2**, 5.

Couto,F.M. *et al.* (2007) Measuring semantic similarity between gene ontology terms. *Data Knowl. Eng.*, **61**, 137–152.

Couto,F.M. *et al.* (2005) Semantic similarity over the Gene Ontology: family correlation and selecting disjunctive ancestors. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management.* ACM Press, New York, NY, pp. 343–344.

Guo,X. *et al.* (2006) Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, **22**, 967–973.

Jain,S. and Bader,G. (2010) An improved method for scoring protein–protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics*, **11**, 562.

Jiang,J.J. and Conrath,D.W. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)*, Taiwan, pp. 9008–9022.

Krogan,N.J. *et al.* (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440**, 637–643.

Li,Y. *et al.* (2003) An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans. Knowl. Data Eng.*, **15**, 871–882.

Lin,D. (1998) An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA, pp. 296–304.

Lord,P.W. *et al.* (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.

Mewes,H.W. *et al.* (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucl. Acids Res.*, **34**, D169–172.

Othman,R. *et al.* (2008) A genetic similarity algorithm for searching the Gene Ontology terms and annotating anonymous protein sequences. *J. Biomed. Inform.*, **41**, 65–81.

Pesquita,C. *et al.* (2008) Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, **9** (Suppl. 5), S4.

Pesquita,C. *et al.* (2009) Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.*, **5**, e1000443+.

Rada,R. *et al.* (1989) Development and application of a metric on semantic nets. *IEEE Trans. Syst. Man Cybern.*, **19**, 17–30.

Resnik,P. (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.*, **11**, 95–130.

Riley,M. and Space,D.B. (1996) Genes and proteins of Escherichia coli (GenProtEc). *Nucl. Acids Res.*, **24**, 40.

Schlicker,A. *et al.* (2006) A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, **7**, 302–317.

Sevilla,J.L. *et al.* (2005) Correlation between gene expression and GO semantic similarity. *IEEE ACM Trans. Comput. Biol. Bioinformatics*, **2**, 330–338.

Smith,T. and Waterman,M. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Wang,H. *et al.* (2004) Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, San Diego, CA, pp. 25–31.

Yu,H. *et al.* (2007) Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications. *Bioinformatics*, **23**, 2163–2173.