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 Pesquita et al. (2007), Lin (1998), and Janga and Conrath (1997); and three recently proposed measures: simUL, simGIC by Pesquita et al. (2008); GraSM by Couto et al. (2008); and Couto and Silva (2008). 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 in their descriptions. Several semantic similarity measures have been proposed that have proved to be useful tools in a variety of biological problems.
Therefore, nodes \( M \) and \( N \) should be greater than the similarity between \( I \) and \( L \). 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: (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 [L et al., 2003; Rada et al., 1998]. 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 their terms themselves [e.g. Couto et al., 2003; Jiang, 1999; Liang and Conrath, 1999; Rada and Reusas, 1999; Rada et al., 2004; Wu et al., 2001]. Among these, the methods of Resnik, 1999; Jiang and Conrath, 1999; and Li, 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, 2013; 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, 1999), simGIC and simUI (Pesquita et al., 2008), GraSM (Couto and Silva, 2013; 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 consist 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 ancestors. 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-root node $v$, an extra ‘unknown’ child node $U_v$ is added to the ontology graph (Fig. 3). An edge is then added from $v$ to $U_v$, and is labelled as follows:

$$P(v \rightarrow U_v) = \frac{N_{-v}}{N_v}$$  \hspace{1cm} (1)

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

$$P(v \rightarrow c) = (1 - P(v \rightarrow U_v)) \frac{N_c}{\sum_{v' \rightarrow c} P(v' \rightarrow c)}$$  \hspace{1cm} (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 $P(v \rightarrow c)$ in Equation (2), the larger $N_{-v}$, the smaller the transition probability $P(v \rightarrow c)$.

Step 2: Downward random walk. In this step, a downward random walk is conducted from each non-root node $v_0$ to determine its relationship to the leaf nodes. Let $W^0_{-v}(v)$ denote the probability of the random walker being at node $v$ after $t$ steps when it started from $v_0$. Initially, $W^0_{-v}(v_0) = 1$ 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:

$$W^t_{-v}(l) = W^t_{-v}(v) + \sum_{v \rightarrow c} W^t_{-v}(c) P(v \rightarrow l)$$  \hspace{1cm} (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 $q \rightarrow v$ to arrive at $v$:

$$W^t_{-v}(v) = \sum_{q \rightarrow v} W^t_{-v}(q) P(q \rightarrow v)$$  \hspace{1cm} (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...
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 $W_\infty$ 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 $\mathcal{L}$ 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 $W_\infty$ and $W_\infty^1$.

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 $v_1$ is given by

$$\text{RWC}(v_0, v_1) = \sum_{i \in \mathcal{L}} W_\infty(i) W_\infty^1(i) \times \text{HSM}(i, j)$$

where $\mathcal{L}$ is the set of all leaf nodes except the newly added unknown nodes.

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 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:

$$\text{ISM}(v_0, v_1) = \frac{1}{2} (\text{RWC}(v_0, v_1) + \text{HSM}(v_0, v_1)).$$

We shall now clarify our method through an example. In Figure 2b, there are 100 genes annotated to the terms of the ontology, the exact counts being shown in parentheses next to the node. The ontology extended by the unknown nodes along with the calculated edge transition probabilities are shown in Figure 2c. The stationary distributions $(W_i)_0$ 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 $(0.25, 0.25, 0.25, 0.25)$, respectively; a random walk starting from node $C$ ends up in the same leaf nodes with probabilities $(0, 0.0, 0.6, 0.2)$. The stationary distribution corresponding to starting node $A$ then follows by recognizing that we step to node $B$ with probability $0.56$ 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 $0.36 \times (0.25, 0.25, 0.25, 0.25) + 0.44 \times (0, 0.0, 0.6, 0.2) = (0.09, 0.09, 0.26, 0.18)$. These distributions do not add up to 1 as the remaining probabilities are leaked to the unknown leaves.

Let us now calculate the semantic similarity between node $B$ and $C$ using Resnik’s measure as the HSM (Fig. 2b). 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 $0.25 \times 0.2 + 0.05$, 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 $C$ $0.916 \times 0.05 = 0.0458$. 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

![Figure 2](https://academic.oup.com/bioinformatics/article-lookup/10.1093/bioinformatics/btq185)
Therefore, we chose the six well-known similarity measures of the number of non-zero elements in the transition matrix, then the overall best-match average approach (Pesquita extension is detailed in the Supplementary Material where the algorithm is step, a random walker at gene of genes (e.g. simUI or simGIC). The basic idea of the extension is that the of the above description extends the algorithm for HSMs defined over pairs k were performed on the GO annotations of the yeast cerevisiae which were annotated to several GO terms, their similarity was taken November 12, 2010, and November 11, 2010, respectively. In the equivalent to the ones presented here. 5.1 Comparison with sequence similarity A technique which has been used by several authors to compare the similarity measure is improved when integrated with RWCs. 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 ISMs reported in the Supplementary Material. Following the approach of previous authors [Sevilla et al. 2005; Schlicker et al. 2004], 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

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. 1
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 \( \leq 0.002 \). 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