Gene 586 (2016) 148–157

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Exploring information from the topology beneath the Gene Ontology terms to improve semantic similarity measures

Shu-Bo Zhang [a,*], Jian-Huang Lai [b]

[a] Department of Computer Science, Guangzhou Maritime Institute, Room 803 Building 88, Dashabei Road, Huangpu District, Guangzhou 510275, PR China
[b] School of Information Science and Technology, Sun Yat-sen University, Room 105 Building 110 East District, 135 Xingangxi Road, Guangzhou 510275, PR China

ARTICLE INFO

Article history:
Received 24 November 2015
Received in revised form 28 March 2016
Accepted 8 April 2016
Available online 12 April 2016

Keywords:
Gene Ontology
Semantic similarity measure
Common descendant
Descendant part
Semantic synthesis
Integrated similarity measure

ABSTRACT

Measuring the similarity between pairs of biological entities is important in molecular biology. The introduction of Gene Ontology (GO) provides us with a promising approach to quantifying the semantic similarity between two genes or gene products. This kind of similarity measure is closely associated with the GO terms annotated to biological entities under consideration and the structure of the GO graph. However, previous works in this field mainly focused on the upper part of the graph, and seldom concerned about the lower part. In this study, we aim to explore information from the lower part of the GO graph for better semantic similarity. We proposed a framework to quantify the similarity measure beneath a term pair, which takes into account both the information two ancestral terms share and the probability that they co-occur with their common descendants. The effectiveness of our approach was evaluated against seven typical measurements on public platform CESSM, protein–protein interaction and gene expression datasets. Experimental results consistently show that the similarity derived from the lower part contributes to better semantic similarity measure. The promising features of our approach are the following: (1) it provides a mirror model to characterize the information two ancestral terms share with respect to their common descendant; (2) it quantifies the probability that two terms co-occur with their common descendant in an efficient way; and (3) our framework can effectively capture the similarity measure beneath two terms, which can serve as an add-on to improve traditional semantic similarity measure between two GO terms. The algorithm was implemented in Matlab and is freely available from http://ejl.org.cn/bio/GOBeneath/.

1. Introduction

Exploring the relationship between pairs of genes and gene products (collectively called biological entities) is a fundamental and important problem in biology and biomedicine, as it contributes to our knowledge of gene functions and biological roles of gene products in the organism. Laboratory approaches to this issue are costly, laborious and time-consuming, which makes computational methods for predicting protein function and gene relationship very attractive. Bioinformatics methods such as sequence alignment and structural comparison also suffer from some limits (Devos and Valencia, 2001; Devos and Valencia, 2000; Valencia, 2005; Joshi and Xu, 2007). The introduction of Gene Ontology (GO) (Ashburner et al., 2000) provides us with a promising approach that serves as a complementary to both experimental and sequence-based methods (Yang et al., 2012). Moreover, it provides us with a means to compare biological entities on aspects that would otherwise not be comparable (Pesquita et al., 2009).

GO is composed of two components (Ashburner et al., 2000): the GO graph and the annotation database. The GO graph is structured as a directed acyclic graph (DAG), which includes three orthogonal sub-ontologies: molecular function (MF), biological process (BP) and cellular component (CC). The annotation database contains data that serve as an association between genes and terms in the ontology, it also provides references to the evidence supporting the association (Consortium, 2015). By exploring the semantic similarity between

Abbreviations: ASM, ascending similarity measure; ASMs, ascending similarity measures; AUC, area under the ROC curve; BMA, best match average; BP, biological process; CC, cellular component; CDT, common descendant term; CDTs, common descendant terms; CESSM, collaborative evaluation of semantic similarity measures; DGA, directed acyclic graph; DIP, database of interacting protein; DiShIn, dubbed disjunctive shared information; DSM, descending similarity measure; DSMs, descending similarity measures; ECC, Enzyme Commission number; FN, false negative; FP, false positive; FPR, false positive rate; GO, Gene Ontology; GOA, Gene Ontology annotation; GraSM, graph-based similarity measure; GTP, guanosine triphosphate; IC, information content; IEA, inferred from electronic annotation; ISM, integrated similarity measure; ISMs, integrated similarity measures; MAX, maximum; MF, molecular function; MICA, most informative common ancestor; PPI, protein–protein interaction; ROC, receiver operating characteristic; TN, true negative; TP, true positive; TPR, true positive rate.

* Corresponding author.
E-mail addresses: 845996912@qq.com (S.-B. Zhang), stsljh@mail.sysu.edu.cn (J.-H. Lai).

http://dx.doi.org/10.1016/j.gene.2016.04.024
0378-1119/

pairs of GO terms, one can determine the relationship between two biological entities. In the hierarchical structure of the GO graph, two terms having a common parent implies that they share the biological meaning of that parent term, and this kind of commonality can be used to derive their semantic similarity measure. In the past decade, the information of this kind hiding in the upper part of GO terms has been investigated extensively to characterize their similarity measure, and some semantic similarity measures have been proved to be useful in many fields (Cheng et al., 2014; Khan et al., 2015; Huang et al., 2007; Jain and Bader, 2010). For details about a comprehensive survey of literature and the more recently proposed method that directly quantifies the similarity measure between two genes, the reader can refer to Pesquita et al. (2009), Gan et al. (2013), and Chicco and Masseroli (2015a, 2015b). However, comparing with the ascendant part, the descendant part of GO terms has received less attention. To the best of our knowledge, there are only two semantic similarity measures that resort to the descendant of terms investigated to date (Yang et al., 2012; Bien et al., 2012).

In the hierarchical structure of the GO DAG, a term with multiple parents means that it inherits various biological semantics from each of its parent and synthesizes the parents' semantics into a new instance or component, namely, new function, process or more specific subcellular localization. From the aspect of the parent terms, two parent terms having a common descendant means that they share the specific concept of that descendant term.

From the view point of gene relationship, if two genes are annotated with a pair of terms sharing identical child (or children), they may have similar molecular function, participate in the same biological process or occur in the same cellular compartment, which is represented by the common child (or children), in a more specific sense than those parent terms they are annotated with. In this sense, a common descendant can help to deduce the similarity between two genes.

Moreover, it is quite often that two terms have the same ancestors but share different descendants (Yang et al., 2012; Bien et al., 2012). In this context, if we consider only the upper part of the term pairs, we will obtain identical similarity value, which cannot discern the difference between the term pairs. In order to address this issue, we should take into account both the upper and lower parts of the terms for their semantic similarity measure. A fragment of MF sub-ontology, that demonstrates the semantic synthesis property of the descendant terms in the lower part and supports the motivation of this study, was given in the supplementary material (Fig. S1 in Section 1).

Based on the above observation, we believe that a common descendant represents some kinds of commonality of its ancestors, and this kind of information can also be used to characterize the similarity of an ancestral term pair. In this study, we proposed a framework to derive the similarity measure of a term pair from the lower part of the GO graph. We took into account both the information shared by two ancestral terms from the GO graph beneath them and the probability that they co-occur with their common descendants. The descending common information was derived by a mirror model and the probability of co-occurrence was achieved in an efficient way. Seven existing methods served as benchmark and the evaluation experiments were performed on several datasets, the results suggested better performance of the integration of similarity scores from both the ascending and descending parts of two terms under consideration.

2. Methods

In this section, we shall firstly propose a mirror model to characterize the commonality of two terms with respect to their common descendants, and then quantify the probability that two ancestral terms co-occur with their descendants. After the descending similarity measures corresponding to seven existing ones are introduced, we combine each descending similarity measures with its corresponding ascending similarity into an integrated similarity measure (ISM) to characterize the similarity between two terms.

2.1. The mirror model for descending commonality

From the intrinsic structure of the GO graph, we find that one child term closer to its parent terms indicates more commonality, which will in turn lead to larger similarity score between its parent terms (two examples that support this intuition were given Fig. S2 of Section 2 in the supplementary material). For two given parent terms, this means that the common descendant is closer to the root of the GO graph, and has smaller value of information content (IC) (according to the definition of IC (Resnik, 1995), a term closer to the root has smaller IC value). Hence, the IC value of a common descendant cannot be directly used to quantify the similarity score of two parent terms. To address this issue, we proposed a mirror model to characterize the commonality with respect to a common descendant term (CDT) for the similarity measure of a parent term pair. The mirror model was established on the basis of the mirror symmetry principle, by which we determined the projection of a term in its upper part of the GO graph with respect to a certain parent node above it. In the mirror model, we took the two parent terms investigated as centers and computed their symmetric nodes in the upper part of the GO graph, respectively, and deduced a mirror node of the common descendant from these two symmetric nodes, then computed the descending similarity value of the parent term pair as we did in the ascending part.

The principle of the mirror model is demonstrated in Fig. 1. Suppose that t_1 and t_2 are two terms under consideration, t_0 is their CDT, and $IC(t_1)$, $IC(t_2)$ and $IC(t_0)$ are the IC values of the three terms, respectively. By the definition of semantic Wu's IC-based distance measure (Wu et al., 2013) and Proposition 1 (see Proposition 1 in Section 3 of the supplementary material), the semantic distance between t_0 and t_1 is $dist(t_0,t_1) = IC(t_0) - IC(t_1)$, then the symmetric node of t_0 in the upper part of the GO graph, with respect to t_1, can be defined as $\hat{t}_{01}$, which

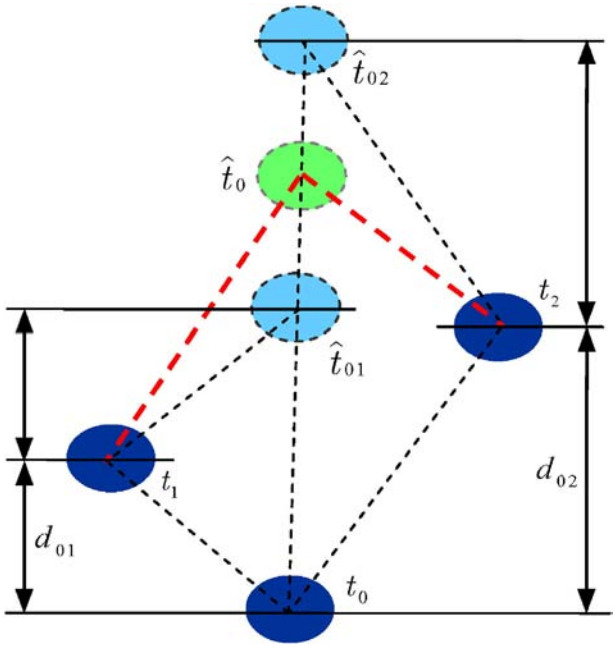

Fig. 1. The mirror model that projects a common descendant into a mirror node in the upper part of the Go graph. The nodes in dark blue color denote the original terms (t_0 is the common descendant of t_1 and t_2), those in light blue color are the symmetric points of t_1 and t_2, and the node in green color is the mirror node of the common descendant node t_0. d_{01} denotes the semantic distance between t_1 and t_0, and d_{02} is the semantic distance between t_2 and t_0.