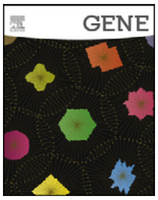
Gene 558 (2015) 108–117

Contents lists available at ScienceDirect

Gene

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



Semantic similarity measurement between gene ontology terms based on exclusively inherited shared information

Shu-Bo Zhang [a,*], Jian-Huang Lai [b,1]

[a] Department of Computer Science, Guangzhou Maritime Institute, Guangzhou, PR China
[b] School of Information Science and Technology, Sun Yat-sen University, Guangzhou, PR China

ARTICLE INFO

Article history:
Received 16 July 2014
Received in revised form 15 December 2014
Accepted 24 December 2014
Available online 28 December 2014

Keywords:
Semantic similarity measurement
Gene Ontology
Information content
Common ancestors
Exclusively inherited common ancestors

ABSTRACT

Quantifying the semantic similarities between pairs of terms in the Gene Ontology (GO) structure can help to explore the functional relationships between biological entities. A common approach to this problem is to measure the information they have in common based on the information content of their common ancestors. However, many studies have their limitations in measuring the information two GO terms share. This study presented a new measurement, exclusively inherited shared information (EISI) that captured the information shared by two terms based on an intuitive observation on the multiple inheritance relationships among the terms in the GO graph. EISI was derived from the information content of the exclusively inherited common ancestors (EICAs), which were screened from the common ancestors according to the attribute of their direct children. The effectiveness of EISI was evaluated against some state-of-the-art measurements on both artificial and real datasets, it produced more relevant results with experts' scores on the artificial dataset, and supported the prior knowledge of gene function in pathways on the Saccharomyces genome database (SGD). The promising features of EISI are the following: (1) it provides a more effective way to characterize the semantic relationship between two GO terms by taking into account multiple common ancestors related, and (2) can quickly detect all EICAs with time complexity of $O(n)$, which is much more efficient than other methods based on disjunctive common ancestors. It is a promising alternative to multiple inheritance based methods for practical applications on large-scale dataset. The algorithm EISI was implemented in Matlab and is freely available from http://treaton.evai.pl/EISI/.

1. Introduction

Comparison of biological entities is important in biological research as it can help to explore the relationship of regulation or function between gene products or genes (collectively called genes hereafter for simplicity), and contribute to the inference of biological roles and functions of genes. Traditional approach to address this issue is based on comparative experiment, which is costly and time consuming, other methods include comparing the sequences or structures between genes (Teng et al., 2013) by means of bioinformatics approaches. The advent of high-throughput technologies has produced a wealth of heterogeneous biological data related to functional annotation of gene. This provides us with a promising way to compare genes on functional level on aspects that could otherwise not be comparable. However, comparing genes based on such huge amount of diverse biomedical datasets is a challenging task, as they are usually constructed in an unconsolidated way. This led to the introduction of various biological ontologies. Gene Ontology (GO) project is one of those that provide consolidated description of gene function for data from different resources. It can be used to explore the functional relationship between two biological entities (Taha, 2012), and has a variety of applications in the fields such as gene function prediction (Nariai et al., 2007; Tao et al., 2007), gene expression data analysis (Alexa et al., 2006; Khatri and Drăghici, 2005), gene clustering (Huang et al., 2007; Yang et al., 2008), disease gene prioritization (Mathur and Dinakarpandian, 2012; Schlicker et al., 2010), analysis of protein interactions (Schlicker et al., 2007; Wang et al., 2010), and so on.

The Gene Ontology comprises two parts, the GO graph and the GO annotation (Camon et al., 2004). The former is composed of controlled terms and organized in three orthogonal aspects: biological process (BP), molecular function (MF), and cellular component (CC), and is structured as a Directed Acyclic Graph (DAG) (Massjouni and Murali,

Abbreviations: BP, Biological Process; MF, Molecular Function; CC, Cellular Component; DGA, Directed Acyclic Graph; GO, Gene Ontology; GraSM, Graph-based Similarity Measure; DiShIn, Dubbed Disjunctive Shared Information; MICA, Most Informative Common Ancestor; DCA, Disjunctive Common Ancestor; DCAs, Disjunctive Common Ancestors; SGD, Saccharomyces Genome Database; EISI, Exclusively Inherited Shared Information; EICA, Exclusively Inherited Common Ancestor; EICAs, Exclusively Inherited Common Ancestors; RF, Robinson–Foulds; PC, Percentage of Correct pathway.

[*] Corresponding author at: Room 803 Building 88, Dashabei Road, Guangzhou, 510275, PR China.
E-mail addresses: 845996912@qq.com (S.-B. Zhang), stsljh@mail.sysu.edu.cn (J.-H. Lai).

[1] Postal address: Room 105 Building 110 East District, 135 Xingangxi Road, Guangzhou, 510275, China.

http://dx.doi.org/10.1016/j.gene.2014.12.062
0378-1119/© 2014 Elsevier B.V. All rights reserved.

2006). In the GO graph, the nodes are controlled terms with specific biological meaning, such as biological process, molecular function, or cellular localization, the edges link nodes and characterize the relationships among terms. The most common relationships are 'is-a' and 'part-of'. The GO annotation builds a bridge between GO terms and genes, it provides annotation information for genes with controlled terms in the GO graph. When a gene is annotated with a GO term, it is also annotated with all the ancestors of that term in the GO graph (Zeng et al., 2008). Moreover, this gene is relevant to other genes that are annotated with the same term, as well as the ancestors and descendants of that term (Taha, 2012). This suggests that we can compare two genes on functional level by measuring the semantic similarity of their GO terms.

In recent years, the research on semantic similarity measurement between two GO terms has drawn more and more attentions from the community of bioinformatics. A variety of metrics have been introduced, and some software tools have been proposed for calculating semantic similarities of GO terms, including Fussimeg (Couto et al., 2003), FunSimMat (Schlicker and Albrecht, 2008), G-SESAME (Du et al., 2009), GFSAT (Xu et al., 2013), GOSemSim (Yu et al., 2010) and SORA (Teng et al., 2013). Traditional approaches to the semantic similarity measurement between GO terms are generally classified into three categories: edge-based methods (Rada et al., 1989; Nagar and Al-Mubaid, 2008; Pekar and Staab, 2002; Jain and Bader, 2010) (also called structure based approaches), which define the semantic similarity based on the conceptual distance derived from the information related with the length or type of edges in the GO graph; node-based methods (Resnik, 1995; Lin, 1998; Couto et al., 2005; Couto and Silva, 2011) (also called annotation-based, information content-based methods), where the nodes and their properties are adopted to compute the information content for similarity; and hybrid methods (Jiang and Conrath, 1997; Wang et al., 2007; Bien et al., 2012; Othman et al., 2008; Wu et al., 2013), that combine the information content with GO graph structure of GO terms for semantic similarity.

Node-based semantic similarity measures are possibly the most frequently mentioned metrics in the literatures (Benabderrahmane et al., 2010). This category of approaches is established on the basis of information theory, and the underlying principle behind is that the more information two concepts have in common, the more similar they are. The information of a concept is quantified by its information content (IC), which rests upon the possibility that it occurs in the GO graph (Othman et al., 2008; Seco et al., 2004) or in a corpus (Resnik, 1995). The information content is an indicator that measures how informative and specific a concept is, and is defined as the negative logarithm of the probability that concept appears. Resnik (1999) proposed a measure based on the information content of the most informative ancestor, which is identified by calculating the IC values of all common ancestors two terms shared and selecting the one with the maximal values. Since the similarity value of Resnik's measurement may be larger than one, Lin (1998) and Jiang and Conrath (1997) proposed their improved schemes to normalize the similarity value to (0,1). Nevertheless, these two kinds of measurements defined similarity based on Resnik's measurement that only consider the information content of a single common ancestor, namely, the Most Informative Common Ancestor (MICA) that inherited by both terms. This is proper in the case that the GO graph is a tree, but it will become problematic in the DAG structure of GO, as a node may have more than one parent nodes and thus some biological information inherited from some ancestors will be neglected.

To address the problem caused by multiple inheritance, Couto et al. employed the concept of disjunctive common ancestors and defined a graph-based similarity measure (GraSM) (Couto et al., 2007), where the information two terms share was derived from all their disjunctive common ancestors by taking the average of their information content. They later updated GraSM and proposed a new method, dubbed Disjunctive Shared Information (DiShIn) (Couto and Silva, 2011), to address the computational complexity problem caused by its recursive definition for disjunctive common ancestors and the problem caused by parallel interpretations shared by two terms. Both GraSM and DiShIn can be directly integrated into any semantic similarity measure based on the MICA (Couto and Silva, 2011). However, the dynamic implementation of GraSM and DiShIn is rather time consuming, as they need to search for the paths between pairs of nodes in the GO graph. To circumvent this problem, they performed a preliminary calculation and stored the results in a database for later computation.

The focus of this paper is to follow the information theoretic vein and propose a novel approach that measures the semantic similarity between two GO terms. We introduced a new measurement based on shared information, exclusively inherited shared information (EISI), which quantifies the information two terms have in common base on some informative common ancestors. The EISI is proposed based on the observation that, only those common ancestors that are inherited exclusively contribute to the shared information of two terms. The common ancestor set was first constructed, each element in which denoted a node that is inherited by both terms. Then all common ancestors were checked, and those had direct descendants inherited by either term exclusively were considered to be exclusively inherited common ancestors (EICAs). Finally, the information content shared by two terms was calculated by taking the average of the information content of all EICAs. Experiments were conducted on artificial dataset and the Saccharomyces genome database (SGD), and the results shown that the similarity measurement based on EISI correlated better with experts' scores on artificial dataset, and supported the prior knowledge of classification information in pathways on SGD. Our measurement has the following advanced properties: (1) it provides a more effective way to characterize the relationship between two GO terms by considering multiple common ancestors, and (2) it can quickly detect all EICAs with time complexity of $O(n)$, which is much more efficient than other methods based on disjunctive common ancestors. It is a promising alternative to multiple inheritance based methods for practical application.

2. Methods

To take into account multiple common ancestors in an effective way, this paper proposes a new measurement to quantify the information shared by two GO terms, based on the exclusively inherited common ancestors (EICAs) they have in common. Like GraSM and DiShIn, EISI takes into account multiple common ancestors two GO terms share, and defines their common information as the average of the information content of their common ancestors. However, EISI considers a common ancestor to be informative only if it has direct descendant that is inherited by either of terms exclusively, this means that not all common ancestors are considered in EISI algorithm, which may reduce the computational complexity for calculating the shared information content.

2.1. Related work

Over the years, great efforts have been devoted to measuring the semantic similarity between GO terms based on information content (Resnik, 1995, 1999; Lin, 1998; Couto and Silva, 2011; Jiang and Conrath, 1997; Couto et al., 2007; Schlicker et al., 2006; Yu et al., 2007). Among these, the methods proposed by Resnik (1999), Lin Lin (1998) and Jiang and Conrath (1997) received much attentions. According to Resnik, the information two terms share is derived from their most informative common ancestors, which can be defined as

$$CA(t_1, t_2) = \{t : t \in parent(t_1) \wedge t \in parent(t_2)\} \tag{1}$$

where $parent(t_1)$ and $parent(t_2)$ are the parent node sets of t_1 and t_2, respectively.

Let t_1 and t_2 are two terms, the information they share can be calculated as the information content of their most informative common