Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 98 (2016) 275 – 282

Procedia Computer Science

The 6th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2016)

Exploiting Biomedical Web Resources: a Case Study

Nicoletta Dessì * , Barbara Pes

Dipartimento di Matematica e Informatica, Università di Cagliari, Via Ospedale 72, 09125 Cagliari, Italy

Abstract

An increasing number of web resources continue to be extensively used by healthcare operators to obtain more accurate diagnostic results. In particular, health care is reaping the benefits of technological advances in genomic for facing the demand of genetic tests that allow a better comprehension of diagnostic results. Within this context, Gene Ontology (GO) is a popular and effective mean for extracting knowledge from a list of genes and evaluating their semantic similarity. This paper investigates about the potential and any limits of GO ontology as support for capturing information about a set of genes which are supposed to play a significant role in a pathological condition. In particular, we present a case study that exploits some biomedical web resources for devising several groups of functionally coherent genes and experiments about the evaluation of their semantic similarity over GO. Due to the GO structure and content, results reveal limitations that not affect the evaluation of the semantic similarity when genes exhibit simple correlations but influence the estimation of the relatedness of genes belonging to complex organizations.

Peer-review under responsibility of the Program Chairs

Keywords: Information and Knowledge Processing, Bioinformatics; Gene Ontology; Semantic similarity of gene sets.

1. Introduction

In recent years, the advent of high-throughput technologies (such as next generation sequencing) and the consequent production of lists of genes associated with specific conditions is stressing the need of recognizing groups of functionally coherent genes in order to construct networks of genes with high pair-wise similarity[1] and characterize these networks with a particular transcriptional behavior[2]. Given difficulties in establishing these

* Corresponding author. Tel.: +0-39-070-6758758; fax: +0-39-070-6758505.
E-mail address: dessi@unica.it

Peer-review under responsibility of the Program Chairs
doi:10.1016/j.procs.2016.09.043

relationships from comparative experiments on the sequence or structure between genes[2], biomedical researchers started to explore new promising ways to compare genes on functional level, including the development of methods for the exploitation of knowledge from ontologies that provide effective descriptions of biomedical events, avoid the short-comings of natural language descriptions (namely ambiguity, subjectivity and lack of structure) and consequently enable automated annotation and automated reasoning over annotations[3]. One of the main contributions in molecular biology has been Gene Ontology (GO) [4], which is dedicated to the functional annotation of gene products in a cellular context[5].

In this paper we explore the potential and any limits of GO ontology in supporting geneticists for capturing additional knowledge about a set of genes which are supposed to play a significant role in a pathological condition. In particular, we try to give suggestions about the extent to which GO ontology can be trusted for detecting the similarity within a group of functionally coherent genes. Our exploration relies on two web resources made freely and easily available by the large multidisciplinary community of biomedical researchers: the HUGO Gene Nomenclature Committee[6] and Reactome[7]. These resources, namely organizations from now on, provide access to a rich catalogues of biomedical and genomic data including information about functional groups of genes i.e. genes acting through the production of specific products and gene-co-function networks. So, we are trusted that GO should also offer a good support in detecting the functional coherence between genes attributed to the same group by the above organizations. For testing our idea, we carried on experiments on the relatedness of gene groups using two classical and popular similarity measures. Results reveal that GO is effective in detecting functional groups of genes, but the hierarchical structure of its catalogue limits the discovery of complex relationships.

The paper is organized as follows. Section 2 describes the web resources and analyses the semantic similarity measures we considered. Next, in the section 3, we present the datasets used in estimating the semantic similarity, the organization our experiments and we also analyze and discuss the results. The section 4 introduces the related work. Finally, section 5 presents our conclusions and the lines of our future research work.

2. Evaluation of the semantic similarity on GO ontology

This section briefly summarizes the basics about the GO structure and the evaluation of the semantic similarity. GO ontology is the result of a collaborative project to provide a controlled vocabularies of terms that describe specific aspects of a gene product's biology. The structure of GO can be described in terms of a graph where each GO term is a node and the relationships between the terms are edges between the nodes. The GO structure is loosely hierarchical as the relationships between different GO terms can be either is-a (parent-child) or part-of (part-whole) relationships, the leaves being the most specific terms. Fig.1 (left) shows an example of the GO structure. Each node is a term and has a unique identifier. Continue arrows indicate "is-a" relationships and dashed arrows denote "part-of" relationships between the terms. In GO there are three types of gene ontologies each describing a particular biological aspect of genes. Specifically, Cellular Component (CC) ontology represents the structural organization of genes, Molecular Function (MF) ontology depicts the specific activities that the gene entails and Biological Processes (BP) ontology describes the series of events that are influenced by the gene. These ontologies are disjoint meaning that no "is-a" relationship operate between terms from different ontologies.

The graph of GO serves as a platform for annotating a term with genes involved in the event that that term describes. Fig.1 (right) shows a toy example of such annotation where the letters in the rectangles indicate the GO terms and the letters in the oval shapes indicate genes directly annotated to GO terms. In Fig. 2 the gene "g7" annotated to the term **"**GO: 005" and the gene "g8" labeled with the term "GO: 001" are considered as correlated because both are annotated to terms which are semantically alike.

In particular, GO has drawn more and more attention from the bioinformatics researchers as a support for assessing the similarity between two genes by measuring the distance between their respective GO terms. A lot of research work has focused on defining semantic similarity measures tailored to the characteristics of GO[8,9,10] that can be broadly classified into the following categories[11]:

- Information Content (IC) measures - These are the earlier developed methods that evaluate the semantic similarity between two genes by considering the frequencies of their annotations within GO terms and their lower