Chapter 5
GO-Based Term Semantic Similarity

Marco A. Alvarez
Utah State University, USA

Xiaojun Qi
Utah State University, USA

Changhui Yan
North Dakota State University, USA

ABSTRACT

As the Gene Ontology (GO) plays more and more important roles in bioinformatics research, there has been great interest in developing objective and accurate methods for calculating semantic similarity between GO terms. In this chapter, the authors first introduce the basic concepts related to the GO and then briefly review the current advances and challenges in the development of methods for calculating semantic similarity between GO terms. Then, the authors introduce a semantic similarity method that does not rely on external data sources. Using this method as an example, the authors show how different properties of the GO can be explored to calculate semantic similarities between pairs of GO terms. The authors conclude the chapter by presenting some thoughts on the directions for future research in this field.

GENE ONTOLOGY AND GENE ONTOLOGY ANNOTATION

The most successful effort for systematically describing current biological knowledge is the GO project (Ashburner et al., 2000), which maintains a dynamic, structured, precisely defined, and controlled vocabulary of terms for expressing the roles of genes and gene products. The GO is dynamic in the sense that its structure changes as more information is available. The GO consists of three different ontologies describing: 1) biological processes (BP), where a process often involves a chemical or physical transformation (e.g. cell growth); 2) molecular functions (MF), where functions are defined as the biochemical activity of gene products (e.g. enzyme); and 3) cellular components (CC), which refers to places in the cell where gene products are active (e.g. nuclear membrane). Each ontology contains nodes (GO

DOI: 10.4018/978-1-4666-3604-0.ch005

terms) linked to each other through *"is-a"* or *"part-of"* relationships forming a directed acyclic graph. Such organization enables the retrieval and visualization of biological knowledge at different levels.

The Gene Ontology Annotation (GOA) project (Barrell et al., 2009) at the European Bioinformatics Institute (EBI) is a project that aims to provide high-quality electronic and manual associations (annotations) between GO terms and UniProt KnowledgeBase (UniProtKB) entries (Consortium, 2009). Crucial to this project is the integration of different databases, a problem that has been addressed by the GO project. The GO maintains a common vocabulary of terms that can be applied to all organisms enabling the annotation across species and databases. The GOA project associates GO terms to UniProtKB entries using strictly controlled manual and electronic methods where every association is supported by a distinct evidence source. A protein can be annotated to multiple GO terms from any of the three ontologies in GO. Functional annotations of UniProtKB proteins currently consists of over 32 million annotations to more than 4 million proteins (Barrell et al., 2009).

SEMANTIC SIMILARITY BETWEEN GENE ONTOLOGY TERMS

The calculation of semantic similarity between pairs of ontology terms aims to capture the relatedness between the semantic content of the terms. Researchers have made great efforts to develop objective and accurate methods to calculate term semantic similarity. For example, semantic similarity between concepts has been a central topic in natural language processing where several robust methods have been proposed based on the WordNet ontology (Budanitsky & Hirst, 2006). In recent years, ontologies have grown to be a popular topic in the biomedical research community creating a demand for computational methods that can exploit their hierarchical structure, in particular, methods for calculating semantic similarity between terms in the GO. Such methods are designed to reflect the closeness or distance between the semantic content of the terms, in other words, their biological relationships.

Additionally, semantic similarity methods can easily be extended to infer higher level semantic relationships. For example, at the protein level, scores for a given protein pair can be calculated by combining the pairwise semantic similarities for the GO terms associated with the proteins. These scores can be used in a broad range of applications such as clustering of genes in pathways (Wang, Du, Payattakool, Yu, & Chen, 2007, Sheehan, Quigley, Gaudin, & Dobson, 2008, Nagar & Al-Mubaid, 2008, Du, Li, Chen, Yu, & Wang, 2009), protein-protein interaction (Xu, Du, & Zhou, 2008), expression profiles of gene products (Sevilla et al., 2005), protein sequence similarity (Pesquita et al., 2008, Mistry & Pavlidis, 2008, Lord, Stevens, Brass, & Goble, 2003), protein function prediction (Fontana, Cestaro, Velasco, Formentin, & Toppo, 2009), and protein family similarity (Couto, Silva, & Coutinho, 2007). An armada of semantic similarity measures using the GO are available in the biomedical literature. A representative collection of available methods have been reviewed and categorized by (Pesquita, Faria, Falcão, Lord, & Couto, 2009).

SEMANTIC SIMILARITY BETWEEN GENE PRODUCTS

In the research related to biological ontologies, great interest has been seen in exploiting ontological annotations to estimate the relationship between gene products, particularly proteins. The use of ontological annotations to measure the similarities between gene products was first introduced in (Lord et al., 2003), where three different methods (Jiang & Conrath, 1997, Lin, 1998, Resnik, 1995) originally designed for the