BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 6. 723–735
Advance Access published on 17 February 2011

doi:10.1093/bib/bbr002

The what, where, how and why of gene ontology—a primer for bioinformaticians

Louis du Plessis, Nives Škunca and Christophe Dessimoz

Submitted: 28th September 2010; Received (in revised form): 12th January 2011

Abstract

With high-throughput technologies providing vast amounts of data, it has become more important to provide systematic, quality annotations. The Gene Ontology (GO) project is the largest resource for cataloguing gene function. Nonetheless, its use is not yet ubiquitous and is still fraught with pitfalls. In this review, we provide a short primer to the GO for bioinformaticians. We summarize important aspects of the structure of the ontology, describe sources and types of functional annotations, survey measures of GO annotation similarity, review typical uses of GO and discuss other important considerations pertaining to the use of GO in bioinformatics applications.

Keywords: *gene ontology; gene annotation; semantic similarity; gene function; function prediction*

INTRODUCTION

The first attempts at classifying gene functions made use of natural language annotations in databases. Early on it was found that natural language by itself is too vague and unspecific to accurately capture the function of genes [1], as it is difficult to perform searches and establish relationships with natural language annotations. The first efforts towards a structured and controlled annotation of genes were schemes such as the enzyme classification (EC) system representing the function of an enzyme using a four digit sequence of numbers [2]. Such classification schemes are still widely used but were found to be insufficient to accurately describe gene function. This motivated the introduction of the Gene Ontology (GO) [3], which has grown to be the largest resource of its kind.

The 'GO Consortium' consists of a number of large databases working together to define standardized ontologies and provide annotations to the GO. The three ontologies it encompasses are non-redundant and share a common space of identifiers and a well-specified syntax. Apart from providing a standardized vocabulary for describing gene and gene product functions, one key motivation behind the GO was the observation that similar genes often have conserved functions in different organisms. The combination of information from all organisms in one central repository makes it possible to integrate knowledge from different databases and to infer the functionality of newly discovered genes. Originally, the GO was developed for a general eukaryotic cell [3]. The initial GO vocabulary, as well as the available GO term annotations present in the first years of its existence reflects this fact (Figure 1). However, the GO Consortium now includes several annotation groups that focus on prokaryotes [5], further contributing to the expansion of the vocabulary and annotations.

Corresponding author. Christophe Dessimoz, ETH Zurich, Computer Science, Universitätstr. 6, 8092 Zurich, Switzerland. E-mail: cdessimoz@inf.ethz.ch

Louis du Plessis is studying a Masters degree in Computational Biology and Bioinformatics at the ETH Zurich. He completed his undergraduate studies at the University of the Witwatersrand in South Africa. His research interests include computational biology, machine learning and image processing.

Nives Škunca is a PhD student at the Ruđer Bošković Institute in Zagreb. Her research interests include computational functional annotation and machine learning.

Christophe Dessimoz is post-doc and lecturer in the CBRG group at ETH Zurich. He strives to understand the forces that shape genes, genomes and species, using computational and statistical methods.

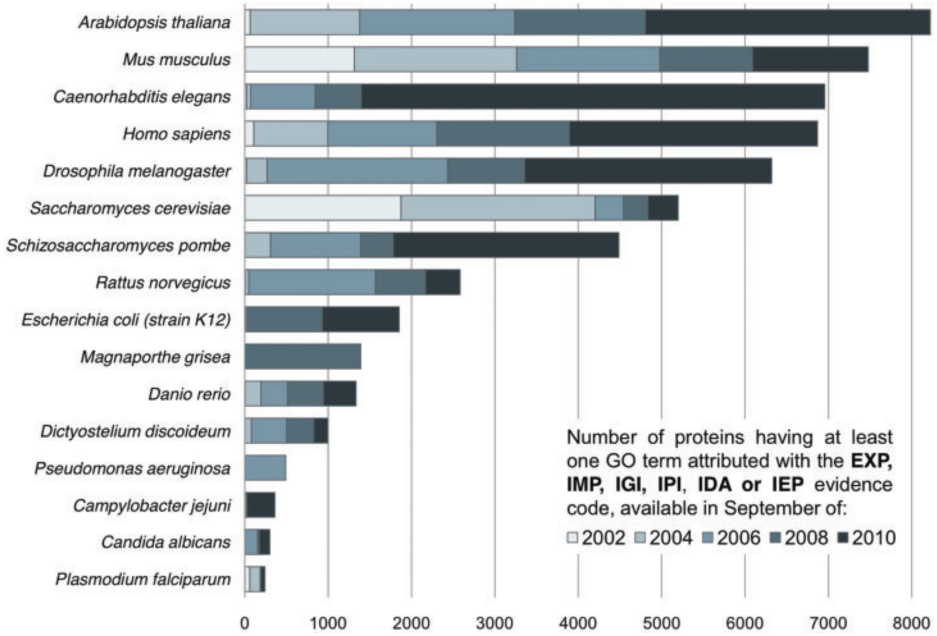

Figure 1: Increase in the number of experimentally verified GO term assignments available for the respective organism between September 2002 and September 2010. The GO consortium was initially focused on Eukaryotes, a fact reflected in the distribution and increase of annotations available in the GO database. Contrast for instance the steady growth of experimentally verified annotations for *A. thaliana*, *S. cerevisiae* or *M. musculus* with the sharp increment in the number of experimentally verified annotations available for *E. coli*: from 33 in 2002 to 1852 in 2010.

The goal of this review is to provide a primer to the GO for bioinformaticians. After a brief introduction to the structure of the ontology, we discuss the different types of annotations associated with the GO. Not all annotations are assigned in the same way and some are more trustworthy than others. Computational inference methods are described in more detail in this section, as they are used to assign a large fraction of GO annotations. The subsequent section discusses common measures of similarity to compare the function of genes quantitatively. The last section reviews typical uses of the GO and common pitfalls for the novice GO user.

WHAT IS THE GO?

The GO is a structured and controlled vocabulary of terms. The terms are subdivided in three non-overlapping ontologies, Molecular Function (MF), Biological Process (BP) and Cellular Component (CC) [6]. Each ontology describes a particular aspect of a gene or gene product functionality, as well as the relations between the terms. These relations are either 'is_a', 'part_of', 'has_part' or 'regulates' relationships. There are two subclasses of the 'regulates' relationships: 'positively regulates' and 'negatively regulates'. The 'is_a' relationship is not used to imply that a term is an instance of another term; instead, it connects a subtype to its more general counterpart (Figure 2). The 'part_of' and 'has_part' relationships are logical complements of each other [7]. The relationships form the edges of a Directed Acyclic Graph (DAG), where the terms are the nodes (Figure 2). This allows for more flexibility than a hierarchy, since each term can have multiple relationships to broader parent terms and more specific child terms. Any path from a term towards the root becomes more general as terms are subsumed by parent terms.

Each gene is associated with the most specific set of terms that describe its functionality. By definition, if a gene is associated with a term, it is also associated with all the parents of that term. The annotation process is discussed in more detail in the next section.

The GO undergoes frequent revisions to add new relationships and terms or remove obsolete ones. If a term is deleted from the ontology, the identifier for the term stays valid, but is labelled as obsolete and all relationships to the term are removed [8]. Changes to the relationships do not affect annotations, because annotations always refer to specific terms, not their location within the GO.

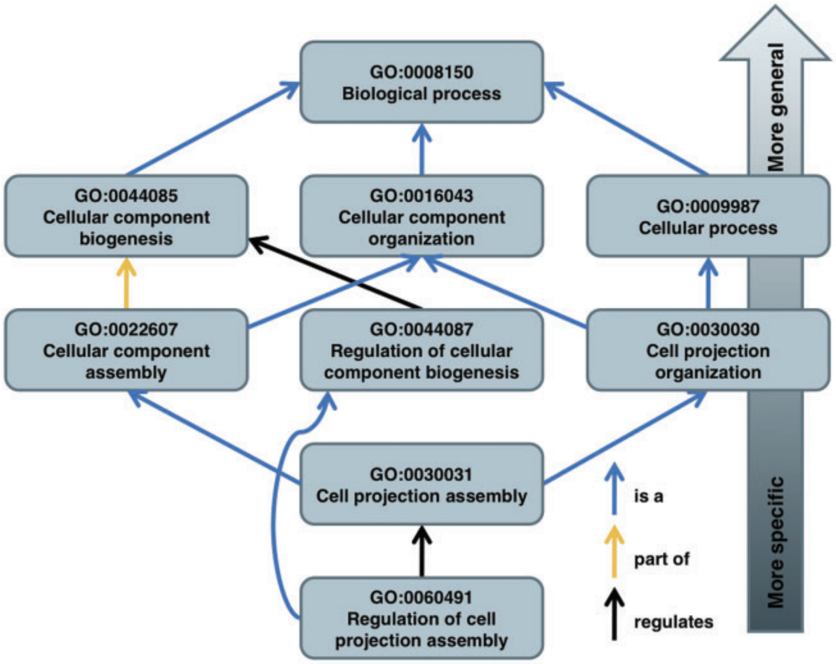
GO:0008150
Biological process
GO:0044085
Cellular component biogenesis
GO:0016043
Cellular component organization
GO:0009987
Cellular process
GO:0022607
Cellular component assembly
GO:0044087
Regulation of cellular component biogenesis
GO:0030030
Cell projection organization
GO:0030031
Cell projection assembly
GO:0060491
Regulation of cell projection assembly
is a
part of
regulates
More general
More specific

Figure 2: The structure of the GO is illustrated on some of the paths of term GO:0060491 to its root term. Note that it is possible for a term to have multiple parents.

It is clear that relationships between the three ontologies exist. For example, an instance of a BP is the execution of one or more MFs [9]. Similarly, relationships exist between the MF and CC ontologies. Recently, these relationships have been integrated into the GO by introducing some inter-ontology links [7]. It should be noted that for the moment there are two concurrent versions of the GO, the filtered and the full GO. The main difference is that the filtered GO does not contain any 'has_part' or inter-ontology relationships. Many of the analysis tools can only use the filtered GO. Thus, the full expressivity of the GO structure is not always available.

WHERE DO ANNOTATIONS COME FROM?

Annotations connect genes and gene products to GO terms. Each annotation in the GO has a source and a database entry attributed to it. The source can be a literature reference, a database reference or computational evidence [4, 6]. In addition, there are three qualifiers used to modify the interpretation of an annotation, 'contributes_to', 'colocalizes_with' and 'NOT', making them an integral part of the annotation [8].

Perhaps the most important attribute of an annotation is the evidence code. The 18 evidence codes available describe the basis for the annotation (Figure 3). These evidence codes are divided into four categories. General guidelines for deciding which evidence code to use are given in Figure 4. It should be kept in mind that one gene can be annotated to the same term with more than one evidence code and that multiple annotations to the same term for the same gene could even share the same reference. This makes it possible to see whether an annotation is supported by more than one type of evidence. However, if the gene is annotated with more than one evidence code and one evidence code is a superclass of another, the annotation with the more general evidence code does not need to be specified explicitly.

INFERRED FROM EXPERIMENT

The most reliable annotations are those inferred directly from experimental evidence. Such annotations are also important to seed the ontology so that the gene function of related genes can be inferred by computational methods [10]. At present, most researchers do not directly add their findings to the GO. The largest fraction of manual annotations are