Contents lists available at ScienceDirect

Atherosclerosis

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

Review

Improvements to cardiovascular Gene Ontology

Ruth C. Lovering[a,*], Emily C. Dimmer[b], Philippa J. Talmud[a]

[a] *Department of Medicine, University College London, Rayne Institute, 5 University Street, London WC1E 6JF, UK*
[b] *GOA Project, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK*

ARTICLE INFO

Article history:
Received 30 June 2008
Received in revised form
29 September 2008
Accepted 14 October 2008
Available online 1 November 2008

Keywords:
Gene Ontology
Cardiovascular science
High-throughput analysis
Chromosome 9

ABSTRACT

Gene Ontology (GO) provides a controlled vocabulary to describe the attributes of genes and gene products in any organism. Although one might initially wonder what relevance a 'controlled vocabulary' might have for cardiovascular science, such a resource is proving highly useful for researchers investigating complex cardiovascular disease phenotypes as well as those interpreting results from high-throughput methodologies. GO enables the current functional knowledge of individual genes to be used to annotate genomic or proteomic datasets. In this way, the GO data provides a very effective way of linking biological knowledge with the analysis of the large datasets of post-genomics research. Consequently, users of high-throughput methodologies such as expression arrays or proteomics will be the main beneficiaries of such annotation sets. However, as GO annotations increase in quality and quantity, groups using small-scale approaches will gradually begin to benefit too. For example, genome wide association scans for coronary heart disease are identifying novel genes, with previously unknown connections to cardiovascular processes, and the comprehensive annotation of these novel genes might provide clues to their cardiovascular link. At least 4000 genes, to date, have been implicated in cardiovascular processes and an initiative is underway to focus on annotating these genes for the benefit of the cardiovascular community. In this article we review the current uses of Gene Ontology annotation to highlight why Gene Ontology should be of interest to all those involved in cardiovascular research.

© 2008 Elsevier Ireland Ltd. Open access under CC BY-NC-ND license.

Contents

1. Introduction

Until recently, the study of specific pathways or individual molecules has been the major approach to understanding the intricate molecular and cellular details associated with cardiovascular processes and disease, with thousands of publications each year adding to our accumulated knowledge of these systems. However, genome-sequencing projects have led to the identification of thousands of genes in higher vertebrates, the majority of which are only characterised by their sequence and genomic location, with their potential involvement in cardiovascular systems awaiting experimental investigation. High-throughput methodologies, such as expression arrays or proteomics are providing substantial information about the properties of these newly identified genes, through the detailed characterisation of the molecular composition of entire tissues, cells or organelles at both specific developmental and specific disease states or through protein binding or cellular location studies. Consequently, such investigations provide researchers with the potential to rapidly increase our understanding of complex interactions and biological functions within the cardiovascular system. However integrating such high-throughput data with the detailed published experimental knowledge about

[*] Corresponding author. Tel.: +44 20 7679 6968; fax: +44 20 7679 6212.
E-mail address: r.lovering@ucl.ac.uk (R.C. Lovering).

0021-9150 © 2008 Elsevier Ireland Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.atherosclerosis.2008.10.014

the function of individual genes is an essential step that is necessary to ensure that all experimental approaches make an impact on current research projects. Fortunately, the Gene Ontology Consortium (GOC) has been developing terms to describe the functional attributes of gene products, across all species, in a consistent and computer-friendly manner to enable the integration of all of these data. This system of terms, called Gene Ontology (GO), enables the accumulated knowledge about individual gene products and their functional domains to be included in individual gene records, in biological sequence databases, and within high-throughput analysis software. This information can then be applied by high-throughput analysis software to aid in the interpretation of large datasets. By providing current functional knowledge in a format that can be exploited by high-throughput technologies, the GOC provides a major freely available public annotation resource that can help bridge the gap between data collation and data analysis [1] (www.geneontology.org).

The success of GO rests on the philosophy behind it; GO was designed by biologists to improve data integration and consequently enables genes to be classified and grouped together according to their functional properties [2–4]. At times the English language can be rather vague, with the majority of words having a variety of subtly different meanings. Similarly, scientific terms or phrases can have dual meanings. Consequently, one of the primary aims of GO is to create a single, explicit definition for each biological term so that these terms can be applied and interpreted consistently by all biologists. All such terms are provided as three structured vocabularies of terms (ontologies) that describe the *molecular functions* that gene products normally carry out, the *biological processes* that gene products are involved in and lastly the *subcellular locations* (*cellular components*) where gene products are active. For example, the annotations for cholesteryl ester transfer protein (*CETP*) include the *Molecular Function* term: 'cholesterol transporter activity', the *Biological Process* term: 'reverse cholesterol transport' and the *Cellular Component* term: 'high-density lipoprotein particle'; whereas the annotations for troponin C type 1 (*TNNC1*) include the *Molecular Function* term: 'troponin I binding', the *Biological Process* term: 'regulation of muscle contraction' and the *Cellular Component* term: 'troponin complex'.

The terms in GO are structured as directed acyclic graphs, where each term can have multiple relationships to broader 'parent' and more specific 'child' terms (Fig. 1). This hierarchical structure produces a representation of biology that allows a greater amount of flexibility in data analysis than would be afforded by a format based on a simple list of terms. Users can manipulate the structure to see either a broad overview of the general functional attributes presented by a set of data, or focus in on specific sections in the ontology to investigate in greater detail.

The second resource supplied by the GOC are datasets of GO terms associated with the appropriate genes and their products, thus providing a resource of diverse detailed functional annotation for many different species [1] (www.geneontology.org/GO.current.annotations.shtml). These annotations are created by 13 different annotation groups, including Gene Ontology Annotation @ EBI (GOA), FlyBase, and the Mouse Genome Database. Depending on the amount of published data available, gene/protein identifiers can be annotated with multiple GO terms from any, or all, of the three gene ontologies (Fig. 1). Annotations can be produced either by a curator reading published scientific papers and manually creating each association or by a software engineer applying computational techniques to predict associations [5]. These two broad categories of techniques have their own advantages and disadvantages, but both require skilled biologists and software engineers to ensure that conservative, high-quality annotations are created. The annotation of each gene is therefore a potentially long laborious process, which for a highly studied gene like *TNF* could take several days (Fig. 2) or, for a more recently described gene like *CDKN2B*, may only take a few hours (Fig. 1).

2. Use of GO in high-throughput studies

As the number of high-throughput methodologies has increased, so has the number of ways in which GO annotation data has been exploited to link experimental results to current functional knowledge.

Proteomes and differentially regulated mRNAs can be analysed with GO data to provide an overview of the predominant activities the constituent proteins are involved in or where they are normally located. For example Ashley et al. [6] used GO to compare the genes up-regulated in *de novo* atherosclerosis with those associated with in-stent restenosis. They found a significant proportion of genes up-regulated following *de novo* atherosclerosis were associated with inflammatory processes, whereas a high proportion of in-stent restenosis up-regulated genes had GO terms indicating an involvement with cell growth and association with the extracellular matrix [6].

Often the generation of hypotheses to explain proteome-wide alterations in response to certain diseases, such as cardiac hypertrophy [7], or stress states, such as hypoxia [8], rely on the use of GO annotation data. In such studies an indication of underlying cellular mechanisms that may account for an observed phenotype can be obtained using GO to cluster subsets of proteins that share related GO annotation, and found to be similarly over- or under-expressed in the disease or stress state. For example, Pan et al. [7] found over-expressed cardiac microsomal membrane proteins in mouse hyper- or hypocontractile hearts were enriched with GO terms describing fat and carbohydrate metabolism and G-protein-dependent signalling pathways [7]. Enrichment of these GO terms validated the investigators proteomic method and was consistent with the suggestion that the deregulation of calcium-dependent cardiac contractility resulted in compensatory growth activities.

The ability to review experimental results with respect to known functional information has also proved useful when investigators need to select a subset of proteins to analyse in greater depth in order to identify new sets of biomarkers for a certain disease. This approach has enabled investigators of buccal carcinoma [9], Parkinson disease [10] and chronic kidney disease [11] to identify new biomarkers for these diseases. Furthermore, in all of these reports the enriched GO categories indicated disease-associated deregulated processes.

GO can also be used to provide a link between the protein binding network and the activities/locations of the participant proteins. Use of cellular component GO annotations can aid data visualisation or confirm whether a particular set of interactions is likely to occur *in vivo*. Dyer et al. [12] used GO data to investigate interactions of human proteins with viral pathogens and found that many different pathogens target the same processes in the human cell, such as regulation of apoptosis, even though they may interact with different proteins. Similarly, many studies have focused on a 'guilt-by-association' hypothesis, where the involvement of proteins in a particular pathway can be hypothesised in relation to the processes their interacting proteins carry out. To this end GO annotations are integrated in the GEOMI [13] and Cytoscape [14] network visualization tools.

A number of proteomic investigations have found that GO data provides an indispensable resource to indicate the success of a particular subcellular enrichment strategy or large scale confocal microscopy analyses [15–18]. Kislinger et al. [15] used GO data to verify that their subcellular fractionation protocol efficiently iso-

Download English Version:

https://daneshyari.com/en/article/5950998

Download Persian Version:

https://daneshyari.com/article/5950998

Daneshyari.com