Incorporation of Ontology-driven Biological Knowledge into Cardiovascular Genomics

Huiru Zheng[1], Haiying Wang[1], Francisco Azuaje[2]

[1]School of Computing and Mathematics, University of Ulster, UK
[2]Laboratory of Cardiovascular Research, Public Research Centre for Health (CRP-Santé), Luxembourg

Abstract

This study presents a system that enables the incorporation of similarity knowledge extracted from the Cardiovascular Gene Ontology (CGO) into cardiovascular research. The implementation of the system is based on the combination of biological function annotations provided by the CGO for more than 4000 genes associated with cardiovascular processes and topological features encoded in the Gene Ontology (GO). Using cardiovascular-related annotations provided by CGO, term-term similarity within each of the GO hierarchies, i.e., molecular function, biological process and cellular component, is computed using three GO-driven similarity measures (Resnik's, Lin's and Jiang's metrics). These provide the foundation for the estimation of semantic similarity between cardiovascular-associated genes. The system allows users to retrieve between-gene similarity using a single query or batch query mode. This study contributes to the development of automated methods for supporting annotation tasks, such as the generation of new annotations for partially-characterized genes associated with cardiovascular disease.

1. Introduction

1.1. The Gene Ontology

The Gene Ontology (GO) is perhaps one of the best known ontologies within the bioinformatics community. Being able to provide a set of controlled, structured vocabularies to describe key domains of molecular biology that can be applied to all organisms, the GO is becoming the de facto standard for annotating gene products. It comprises three hierarchies: Molecular function (MF), biological process (BP), and cellular component (CC). MF represents information on the role played by a gene product. BP refers to a biological objective to which a gene product contributes within an organism. CC stands for the cellular localization of the gene product, including cellular structures and complexes. For example, the annotations for human BIRC6 protein include the BP term: 'apoptosis', the MF term: 'protein binding' and the CC term: 'intracellular'. In the April 2011 GO release, there are 21105 BP terms, 9834 MF terms and 2942 CC terms.

The annotation terms within each hierarchy are structured to allow both assignment and querying at different levels of granularity (from very general functional categories to more specific categories). An important feature of the GO is that the terms and their relationships within each hierarchy are represented by *directed acyclic graphs* (DAGs), in which each component may be linked to more than one parent node.

Recent developments in GO includes the introduction of new relationships and new links between its three hierarchies [1] as illustrated in Figure 1. In addition to two relationship types (*is_a* and *part_of*) initially introduced, *regulates, positively_regulates,* and *negatively_regulates* relationships have been added to GO to describe regulatory terms and their regulated parents. For instance, the link between 'phosphorylation' and 'regulation of phosphorylation' two BP terms is now described by the *regulates* link as shown in Figure 1. Examples of the cross-ontology links can be found in Figure 1, in which, a *regulates* relation is used to link the BP term: 'regulation of kinase activity' and the MF term: 'kinase activity'.

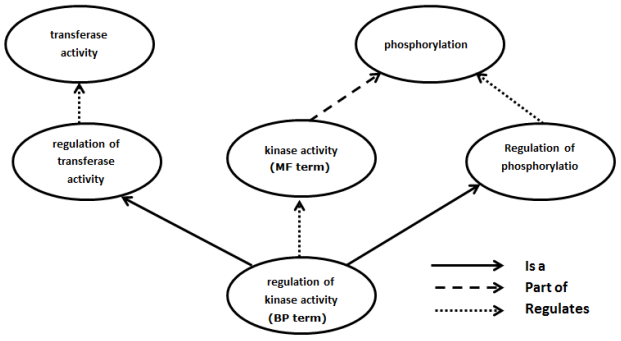

Figure 1. Examples of the new relationships and new links between GO hierarchies

The widespread use of GO annotations greatly facilitates cross-species/cross-database queries. However, the significance of the GO has been increasingly recognized and it is being used beyond the primary goal for which it is designed: functional annotation [2].

1.2. Cardiovascular GO

Funded by British Heart Foundation (BHF), the Cardiovascular GO Annotation Initiative (http://www.ebi.ac.uk/GOA/CVI/) aims to provide comprehensive functional annotation for genes implicated in heart development and cardiovascular processes and disease for the cardiovascular research community, both in the UK and internationally. This is the first time that a physiological process-centered approach has been used for human protein GO annotation. The implementation of the project is based on the collaboration between University College London (UCL) and the European Bioinformatics Institute (EBI).

In an attempt to support high-throughput cardiovascular research, the BHF-UCL tem has initiated an effort to fully describe heart development and cardiovascular process in GO [3], [4]. For example, the number of GO terms used to describe heart development has been refined and expanded from 12 to over 280. Twenty-six cell type annotation terms specific to the heart, ranging from general cell type terms like cardiac endothelial cell differentiation (GO:0003293) to specific cell type terms such as cardiac Purkinje fiber cell differentiation (GO:0003168), have been included in GO. As illustrated in Figure 2, among 17 child terms associated with the BP term: 'heart development', nine are newly introduced by the BHF-UCL initiative, including GO:0003205, GO:0003161, GO:0003204, GO:0060973, GO:0061311, GO:0060976, GO:0003197, GO:0003157, and GO:0003170. With the structure of the ontology associated with heart development established by the BHF-UCL team, it is expected that all aspects of heart development process can be interpreted from both anatomical and cellular prospective.

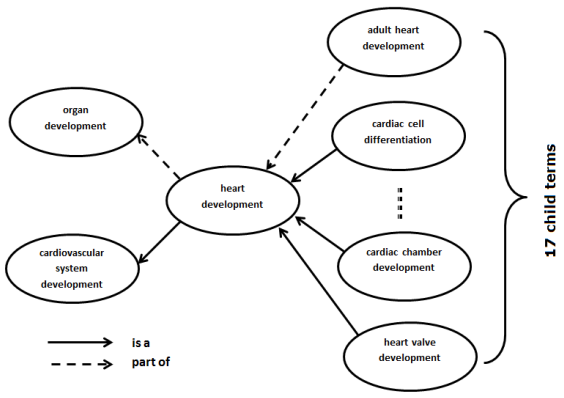

Figure 2. Partial view of GO term: heart development

More than 4100 human proteins have been identified as associated with cardiovascular processes in the annotation file published in April 2011. A total of 5513, 2299, and 706 GO terms associated with biological process, molecular function and cellular component have been used respectively. Altogether there are 133,722 annotations, of which 59510 were inferred from electronic annotation with *IEA* as evidence code. It appears that not all the annotations included in the BHF-UCL annotation file available at ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/bhf-ucl/ are listed in the GO database (http://www.geneontology.org/GO.database.shtml). For example, based on the BHF-UCL annotation file released in April 2011, there are 18 GO terms used to annotate the *BBS1*, *Bardet-Biedl syndrome 1 protein*, while only 6 GO terms are found in the GO database associated with this protein, as shown in Table 1.

Table 1. The annotations of BBS1 human protein included in both GO database and BHF-UCL annotation file.

Source	Ontology	Annotations
GO database	BP	GO:0001895;GO:0035058; GO:0042384;GO:0045494;
	MF	GO:0005515;
	CC	GO:0034464;
BHF-UCL	BP	GO:0060271;GO:0060296; GO:0021766;GO:0021987; GO:0045444;GO:0045494; GO:0033365;GO:0040015; GO:0048854;GO:0007288; GO:0021756;GO:0008104; GO:0030534;GO:0032402;
	MF	GO:0005515;
	CC	GO:0034464;GO:0005932; GO:0031514;

1.3. The objectives of this study

The assessment of semantic similarity between gene products using the annotation provided by the GO database has been widely studied. However, the similarity between gene products derived from the Cardiovascular Gene Ontology (CGO) has not been rigorously studied yet. There is a lack of publicly-available, user-friendly tools for supporting CGO-driven similarity assessment tasks. Based on the exploitation of the CGO, this study presents a system that enables the incorporation of CGO similarity knowledge into cardiovascular research. The implementation of the system is based on the combination of biological function annotations provided by the CGO for more than 4000 genes associated with cardiovascular processes and topological features encoded in GO. It will provide the foundation for the estimation of semantic similarity between cardiovascular-associated genes.

2. Methodology

The implementation of the system is based on the exploitation of both topological features of the GO (i.e., between-term relationships in the hierarchy) and statistical features of the annotations provided by the BHF-UCL initiative (i.e., the frequency of GO terms used) to assess functional similarity among gene products associated with heart development.

2.1. Information content-based approach to assessing between-term similarity

As a first step, the number of cardiovascular-related genes associated with each GO term and its child terms was computed. By calculating the probability of finding a child of a GO term in the CGO annotation file, the information content (*IC*) associated with each GO term, *c,* was then established using the following equation.

$$IC(c) = -\log(p(c)) \tag{1}$$

where *p(c)* is the probability of finding term *c* and a child of *c* in the annotation database under analysis. In the context of cardiovascular-related annotations, term-term similarity within each of the GO hierarchies, i.e., molecular function, biological process and cellular component, is computed using three GO-driven similarity measures (Resnik's, Lin's and Jiang's metrics) defined as follows. $sim(c_i, c_j)$ and $d(c_i, c_j)$ represent the semantic similarity and distance between terms c_i and c_j. $S(c_i, c_j)$ represents the set of parent terms shared by c_i and c_j.

- Resnik's metric:

$$sim(c_i, c_j) = \max_{c \in S(c_i, c_j)} [-\log(p(c))] \tag{2}$$

- Lin's metric:

$$sim(c_i, c_j) = \frac{2 \times \max_{c \in S(c_i, c_j)} [\log(p(c)]}{\log(p(c_i)) + \log(p(c_j))} \tag{3}$$

- Jiang's metrics:

$$d(c_i, c_j) = 2 \times \max_{c \in S(c_i, c_j)} [\log(p(c)] - [\log(p(c_i)) + \log(p(c_j))] \tag{4}$$

The reader is referred [5] and [6] to for a more detailed description of these metrics.

2.2. Between-gene similarity

Once semantic similarity between annotation terms is established using equations (2) to (4), the semantic similarity between genes can be calculated using their annotations provided by the BHF-UCL initiative. Let A_i and A_j be two sets of annotations describing the gene pair g_i and g_j, m and n be the number of GO terms included in A_i and A_j respectively, and $sim(c_k, c_p)$ be the similarity between terms c_k and c_p. The following two approaches [5], [7] have been implemented to estimate the similarity between the genes g_i and g_j.

- *average inter-set similarity*

$$SIM(g_i, g_j) = \frac{1}{m \times n} \times \sum_{c_k \in A_i, c_p \in A_j} sim(c_k, c_p) \tag{5}$$

- *maximum between-term similarity*

$$SIM(g_i, g_j) = \frac{1}{m+n} \times \left(\sum_{k \in A_i} \max_{p \in A_j} (sim(c_k, c_p)) + \sum_{p \in A_j} \max_{k \in A_i} (sim(c_k, c_p)) \right) \tag{6}$$

3. Implementation

The system was implemented as a standalone Java-based system in which all the information including the annotations provided by the BHF-UCL is stored in a MySQL database. It includes the following components:

- Login MySQL database. When starting the program, only *Login MySQL* menu is enabled. This allows authorized users to access the database and to run the system.
- Establishing the information content for each GO term based on the BHF-UCL annotations. It involves the following steps: (1) finding out the number of gene products associated with each GO term, *c*, and its child terms, *freq(c)*; (2) calculating the probability of finding a child of c in the annotation database being analysed; and (3) computing the information content for each GO term using Equation (1).
- Estimating the similarity value for each term pair, c_i and c_j. with Resnik's, Lin's and Jiang.s metrics as shown in Equations (2) to (4)
- Calculating the similarity between genes associated with heart development and cardiovascular process, which are identified by the BHF-UCL team, using Equations (5) to (6). The system allows users to retrieve between-gene similarity using a single query or batch query mode. Only manual annotations are considered.

In addition, the system allows users to input a gene expression file to simultaneously estimate expression correlation and semantic similarity between a cardiovascular-associated gene pair, as illustrated in Figure 3.

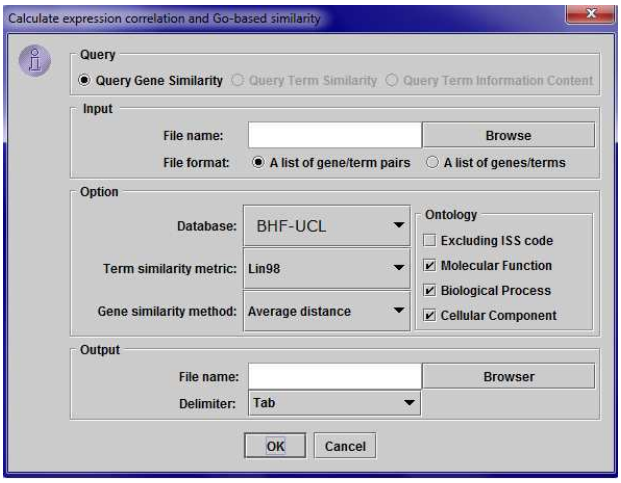

Figure 3. A screenshot for calculating expression correlation and semantic similarity between genes

4. Discussion and conclusions

Similarity assessment between gene products is important to implement predictive models for large-scale functional genomics. It is widely accepted that ontology-based similarity knowledge can be used to support both structural and functional classification problems. The system presented in this study can be used to support various functional genomics tasks, such as the generation of new annotations for partially-characterized genes associated with cardiovascular disease. The system can also be used to support functional prediction tasks in cardiovascular genomics, such as the validation of gene expression analyses and the identification of false positives in protein interaction networks. As a case study, we investigated the relationship between gene expression correlation and functional similarity for a list of 247 priority cardiovascular genes identified by the BHF-UCL team. The expression data used in this preliminary study was published by Barth et al. [8], which includes 7 samples with dilated cardiomyopathy and 5 samples with non-failing hearts. Pearson correlation coefficient (PCC) and Lin's metric-based semantic similarity (SS) were calculated and compared for each gene pair. No significant association between functional similarity and expression correlation were observed in this study as illustrated. One possible reason is the limited size of the expression data, which only consists of 12 samples under two physiological conditions. Using other large datasets to further investigate the relationship between gene expression correlation and functional similarity knowledge extracted from the CGO would be an important part of our future work. Relationships between semantic similarity derived from CGO and other functional properties such as protein-protein interactions will also investigated. In terms of the system presented in this paper, several expansions are on the way including the development of an online system and the incorporation of the functionalities into other open-source, public available biological resources such as Cytoscape [9].

References

[1] Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res. 2010, 38(Database issue): D331-D335.

[2] Chagoyen M, Pazos F. Quantifying the biological significance of gene ontology biological processes--implications for the analysis of systems-wide data. Bioinformatics. 2010 Feb 1;26(3):378-84.

[3] Khodiyar VK, Hill DP, Howe D, Berardini TZ, Tweedie S, Talmud PJ, Breckenridge R, Bhattarcharya S, Riley P, Scambler P, Lovering RC. The representation of heart development in the gene ontology. Developmental Biology. 2011, 354(1):9-17.

[4] Lovering RC, Dimmer EC, Talmud PJ. Improvements to cardiovascular gene ontology. Atherosclerosis. 2009, 205(1):9-14

[5] Lord P, Stevens R, Brass A, and Goble C. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics, 2003, 19, 1275-1283.

[6] Wang H, Azuaje F, Bodenreider O, and Dopazo J. Gene expression correla-tion and gene ontology-based similarity: an assessment of quantitative relation-ships. In Proc. of IEEE 2004 Symposium on Computational Intelligence in Bioinformatics and Computational Biology, La Jolla, CA, USA, 25-31.

[7] Azuaje F., Wang H., and Bodenreider O. Ontology-driven similarity ap-proaches to supporting gene functional assessment. In Proc. Of The Eighth An-nual Bio-Ontologies Meeting, Michigan, 25 June 2005, http://bio-ontologies.man.ac.uk/

[8] Barth AS, Kuner R, Buness A, Ruschhaupt M, Merk S, Zwermann L, Kaab S, Kreuzer E, Steinbeck G, Mansmann U, Poustka A, Nabauer M, Sultmann H. Identification of a Common Gene Expression Signature in Dilated Cardiomyopathy Across Independent Microarray Studies. Journal of the American College of Cardiology 2006, 48: 1618-20.

[9] Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011, 27(3): 431-432.

Address for correspondence.

Name. Huiru Zheng
Full postal address: School of Computing and Mathematics, University of Ulster at Jordanstown, BT37 0QB, UK

E-mail address: h.zheng@ulster.ac.uk.