Bioinformatics Advance Access published September 23, 2009

Application Note

SimCT: a generic tool to visualize ontology based relationships for biological objects

Carl Herrmann[1,4] Sèverine Bérard[2] and Laurent Tichit[3,4*]

[1]TAGC – U928 Inserm, [2]Université Montpellier 2, UMR AMAP, Montpellier, France, [3]IML – UMR 6206 CNRS , [4]Faculté des Sciences, Université de la Méditerranée, Campus de Luminy, Marseille, France

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Summary: We present a web-based service, SimCT, which allows to graphically display the relationships between biological objects (e.g. genes or proteins) based on their annotations to a OBO ontology. The result is presented as a tree of these objects, which can be viewed and explored through a specific java applet designed to highlight relevant features. Unlike the numerous tools that search for over-represented terms, SimCT draws a simplified representation of biological terms present in the set of objects, and can be applied to any ontology for which annotation data is available. Being web-based, it does not require prior installation, and provides an intuitive, easy-to-use service.

Availability: http://tagc.univ-mrs.fr/SimCT

1 INTRODUCTION

The wealth of data available from large scale experiments in recent years has made the development of efficient tools to visualize, analyze, interpret and share post-genomic data a crucial endeavor. Among these, biomedical ontologies have been increasingly developed to annotate genome wide features, and numerous fields of biology are now covered by a dedicated ontology. If Gene Ontology (GO) appears to be the pioneering project, many other projects are actively pursued (e.g. (Robinson *et al.*, 2008), see http://www.obofoundry.org/ for available biomedical ontologies) and their adoption by genome databases such as MGI, Wormbase and Flybase, will ensure that they will be increasingly used in the community. In this note, we present a generic, web-based tool called SimCT (= Similarity Clustering Tool) which allows the visualization of the relations between biological objects (e.g. genes, proteins,...) based on their annotations to an ontology, in the form of a *clustering tree*. Our clustering procedure is a way to turn the ontology into a simplified tree (which is a subgraph of the ontology) that better represents the terms associated to a list of objects, therefore highlighting their relationships. This representation could neither be obtained by mapping the annotations onto the ontology, due to its complexity, nor by searching for over-represented terms, which by definition overlooks terms that are not statistically relevant. The visualization is done using a dedicated java applet. Although many tools have been developed for GO, very few comparable tools exist for other biomedical ontologies yet.

2 METHODS

To measure the specificity of a term t in an ontology *O*, we have introduced the notion of precision as follows (see Supplementary Text for details, in particular the glossary for definition of terms used):

$$p(t) = -\frac{\log \dfrac{N_d(t)}{NN_a(t)}}{\log NN_a^{\max}} \in [0,1] \tag{1}$$

where $N_d(t)$ represents the number of descendant terms of t, $N_a(t)$ the number of ancestor terms of t, *N* the total number of terms in *O*, and N_a^{max} the maximal number of ancestors a term can have in *O*. Interestingly, our definition of precision only depends on the structure of the ontology and not on annotation statistics like in (Lord et al.,2003). Therefore, it can be applied to any existing ontology. Additionally, precision differs from information content, which gives equal specificity to all leaves of the ontology (Resnik, 1999)(Schlicker et al., 2006)(Wang et al., 2007). Based on precision, we define the similarity of two terms as the precision of their most precise common ancestor.

Given a list of objects annotated to an ontology, we consider the set of (object|ontology term) pairs. If an object has several annotations, it generates several (object|ontology term) pairs. We have implemented an aggregative clustering algorithm that builds the clustering tree based on the similarity between terms. The leaves of the resulting tree are the (object|ontology term) pairs and the internal nodes are ontology terms. We attach to each internal node a numerical index called Subtree Relevance Index (SRI):

$$SRI(T) = p(t) \times N(T) \tag{2}$$

where *T* represents a subtree, *t* the ontology term attached to it, *p(t)* its precision and *N(T)* is the number of leaves of the subtree. It measures the relevance of each term for the list of objects submitted (see Supplementary Text). The topology of the tree respects that of the underlying ontology (i.e. it is included in the DAG of the ontology).

3 IMPLEMENTATION

SimCT can be used in two different ways, depending on the ontology the user is interested in:

(1) with Gene Ontology, the user can input a list of genes/proteins, select the corresponding organism (29 are currently

[*] To whom correspondence should be addressed.

available) and the GO sub-ontologies. The system retrieves available annotations.

(2) for other ontologies, the user must provide a two-columns list of objects associated to their annotations, and select the corresponding ontology among the 25 currently available. The user can also provide custom GO annotations.

In both cases, the (object|ontology term) list is processed by the clustering algorithm. Once done, the user can open a java applet which displays the tree(s). As an example, the clustering of 300 leaves takes about 40 seconds.

4 APPLICATIONS

4.1 Disease ontology

To illustrate the use of SimCT, we have extracted from http://www.genome.gov/gwastudies/ a list of 79 SNPs associated to a disease, described using the Disease Ontology (http://diseaseontology.sourceforge.net/). Fig. 1 shows the resulting tree, in which 2 subtrees are highlighted: *Diabetes Mellitus* and *Noninfectious enteritis and colitis*. Taking the intersection of both trees through the applet menu reveals that 3 SNPs are simultaneously associated to both diseases (rs3024505, rs2542151,rs2476601). Interestingly, the latter two are close to or inside two genes, respectively *PTPN2* and *PTPN22*. Inspection of the OMIM entries related to both genes shows that only the first one is explicitly associated to both diabetes and enteritis, while *PTPN22* is only associated to diabetes. Thus, our result suggests that we could add an additional annotation to *PTPN22*, namely enteritis.

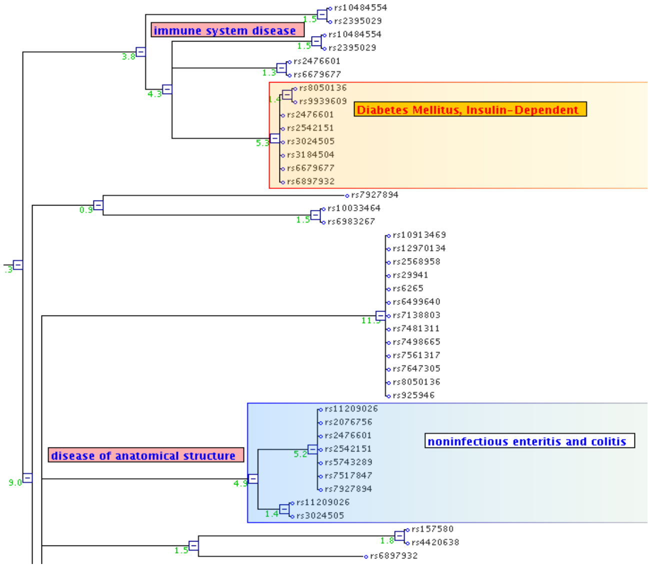

Fig. 1: A clustering tree of SNPs associated to diseases. Highlighted are two subtrees, *Diabetes Mellitus* (red) and *Noninfectious enteritis and colitis* (blue).

4.2 Gene Ontology

We chose a set of 69 co-regulated genes extracted from Transcriptome Browser (Lopez *et al.*, 2008) around the NK gene *NCR3* in human. We compared the p-values of the nodes with SRI≥2.5 with the p-values given by DAVID (Dennis *et al.*,2003) and GO:TermFinder (Boyle *et al.*, 2004). The differences between the SimCT approach and the search for over-represented terms are highlighted in Table S2. In particular, although no term related to biopolymer synthesis is found by DAVID as over-represented, SimCT detects that 5 genes of the list are related to transcriptional (*TBX21, CEBPD, TAF6L, GFI1*) or translational (*EIF5B, RPS8*) processes which are child terms of biopolymer synthesis. These are the effectors at the end of the cascade of natural killer activity, leading for instance to the production of gamma-interferon.

5 CONCLUSION

Our approach can be compared to GOSurfer (Zhong *et al.*, 2004), GOTreePlus (Lee *et al.*, 2008) or GO::TermFinder (Boyle *et al.*, 2004). However, SimCT includes the possibility to work with other biomedical ontologies than GO, and is web-based. Therefore, it provides an intuitive, easy-to-use and immediately available service, allowing to draw a clear picture of the ontological terms represented in a list of biological objects annotated to an ontology, for any OBO biomedical ontology. The viewer applet helps easily exploring and annotating the resulting tree to highlight its most relevant features. As more and more ontologies are being developed, we believe that this tool will prove very useful in working with these.

ACKNOWLEDGEMENTS

We want to thank E. Choron, M. Hermes, P. Bonnaure, L. Baumgaertel and C. Lepoivre for their participation in the developments of the java TreeViewer applet. We thank P. Rihet, C. Brun and B. Jacq for discussions and feedback.

REFERENCES

Robinson PN et al. (2008) The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.*, **83**.

Lord PW et al. (2003) Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**.

Resnik P. (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, **11**.

Schlicker A et al. (2006) A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, **7**.

Wang JZ et al. (2007) A new method to measure the semantic similarity of go terms. *Bioinformatics*, **23**.

Halfon MS et al. (2008) Redfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in drosophila. *Nucleic Acids Res.*, **36**.

Lopez F et al. (2008) Transcriptomebrowser: a powerful and flexible toolbox to explore productively the transcriptional landscape of the gene expression omnibus database. *PLoS ONE*, **3**.

Dennis GJ et al. (2003) David: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**.

Boyle EI et al. (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes, *Bioinformatics* , **20**

Zhong S et al. (2004) Gosurfer: a graphical interactive tool for comparative analysis of large gene sets in gene ontology space. *Appl. Bioinformatics*, **3**.

Lee B et al. (2008) Gotreeplus: an interactive gene ontology browser for proteomics projects. *Bioinformatics*, **24**