BIOINFORMATICS APPLICATIONS NOTE

Vol. 30 no. 7 2014, pages 1043–1044
doi:10.1093/bioinformatics/btt689

Databases and ontologies

Advance Access publication December 1, 2013

LipidGO: database for lipid-related GO terms and applications

Mengyuan Fan[1,2], Hong Sang Low[3], Hufeng Zhou[2], Markus R. Wenk[4] and Limsoon Wong[1,5,*]

[1]Department of Computer Science, National University of Singapore, [2]NUS Graduate School of Integrative Science and Engineering, National University of Singapore, Singapore, [3]National Research Foundation, Singapore, [4]Department of Biochemistry and [5]Department of Pathology, National University of Singapore, Singapore

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Lipid, an essential class of biomolecules, is receiving increasing attention in the research community, especially with the development of analytical technique like mass spectrometry. Gene Ontology (GO) is the de facto standard function annotation scheme for gene products. Identification of both explicit and implicit lipid-related GO terms will help lipid research in many ways, e.g. assigning lipid function in protein function prediction.

Results: We have constructed a Web site 'LipidGO' that facilitates browsing and searching lipid-related GO terms. An expandable hierarchical GO tree is constructed that allows users to find lipid-related GO terms easily. To support large-scale analysis, a user is able to upload a list of gene products or a list of GO terms to find out which of them is lipid related. Finally, we demonstrate the usefulness of 'LipidGO' by two applications: (i) identifying lipid-related gene products in model organisms and (ii) discovering potential novel lipid-related molecular functions

Availability and implementation: LipidGO is available at http://compbio.ddns.comp.nus.edu.sg/%7elipidgo/index.php.

Contact: wongls@comp.nus.edu.sg

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on July 24, 2013; revised on November 5, 2013; accepted on November 20, 2013

1 INTRODUCTION

Lipid and their metabolites are essential in many processes in biology like energy homeostasis, signaling, neurobiology, infectious diseases and so forth. The development of analytical technique like mass spectrometry gives rise to the field of lipidomics (Wenk *et al*., 2005). An important goal in lipidomics research is to identify genes and proteins involved in lipid synthesis, metabolism and other relevant processes. As with other subfield of biology, lipid researchers are primarily interested in elucidating functions of gene products. The Gene Ontology (GO) (Ashburner *et al*., 2000) is the most widely used control vocabulary in functional annotation of genes products, and it is being widely used by wet lab researchers and bioinformaticians alike. Thus, the identification of lipid-related GO terms can facilitate lipid research, either directly or with other bioinformatics tool involving GO terms. For example, lipid-related GO terms were used to identify novel lipid-associated complexes involved in cancer progression (Goh *et al*., 2013).

GO consists of three independent subontologies: biological process (BP), cellular component (CC) and molecular function (MF). GO is a directed acyclic graph with multiple inheritance property. In GO, there are several relationship types. However, in the LipidGO project, where the focus is on the lipid-relatedness property, only 'is-a' relationship is considered. More on term relationship is elaborated in the appendix and in the documentation section of our Web site

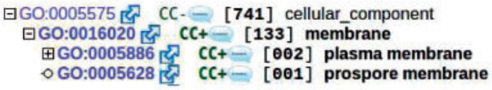
The lipid-related GO terms in LipidGO are discovered by curation and computational approach. Detailed procedures and issues are given in Section 3. All the lipid-related GO terms found are subject to stringent predefined curation rules. The rules can be found in the documentation at our Web site or appendix. In our Web site, the following notations are used to describe class labels of GO terms: 'BP+' for lipid-related BP terms, 'BP−' for non–lipid-related BP terms and 'BP?' for BP terms whose lipid relatedness has not been examined by a curator. Similar notations are defined for CC terms (CC+, CC−, CC?) and MF terms (MF+, MF−, MF?). We will talk about the *unknown* class labels in Section 3.

2 DATABASE AND FEATURES

2.1 Browse function

An expandable hierarchical GO tree (Fig. 1) is constructed. It allows users to find lipid-related GO terms easily. The structure of the tree is reflected in the indentation, with lower-level (more specific) terms indented to the right. By default only the root node and its children are shown. Users need to click the expand button (⊞) or collapse button (⊟) to unfold/fold the tree. Leaf nodes (◇) cannot be expanded. As our focus is on lipid-related terms, we make them easier to be spotted in two ways: first, the term is bold if it is lipid-related; second, for each GO term, the number of lipid-related terms in the subtree rooted at that term is indicated inside a square bracket '[]'. To help users obtain more information, internal link (GO:xxxxxxx itself) and external link (↗) to the official GO browser Amigo (Carbon *et al*., 2009) are provided. Furthermore, if a user does not agree with the class label of a particular term, he or she can give a comment by clicking the comment balloon (💬). We welcome feedbacks so that we can make better classification of GO terms in terms of lipid relatedness. Instead of browsing in

*To whom correspondence should be addressed.



```
⊟GO:0005575  CC-  [741] cellular_component
  ⊟GO:0016020  CC+  [133] membrane
    ⊞GO:0005886  CC+  [002] plasma membrane
    ○GO:0005628  CC+  [001] prospore membrane
```

Fig. 1. A snippet of expandable hierarchical GO tree

hierarchical format, flat lists of lipid-related GO terms are available for viewing or downloading. The number of lipid-related GO terms are relatively small (4674 terms among a total of 37 462 terms for the April 2013 version of GO), and they are usually near the bottom of the GO hierarchy. Therefore, clicking the expand button at higher level nodes several times are required to reveal lipid-related terms. For those interested, the number of lipid-related GO terms for each subontology can also be found on the browse page of the Web site.

2.2 Search function

Users are able to search in two ways. Searching by key words will return the list of GO terms whose term or definition contains that phrase. Searching by GO accession number will give detailed information about the term and its related gene products. Users are able to download the GO subtree rooted at a term and see which branches are lipid related.

2.3 Batch query function

All the lipid-related gene products can be downloaded on the batch query page of our Web site. Moreover, users are able to upload a list of gene products to find out which of them is lipid related and what the associated lipid-related GO terms are. The users are also able to upload a list of GO terms to identify the ones that are lipid related.

2.4 Applications

Here, we provide two applications; for more details and download, please visit application page on our Web site. The data sources for both applications (gene product, annotation information) are obtained from the official GO database.

2.4.1 Lipid-related gene products in model organisms In recent years, lipid research is being carried out extensively in many model organisms. There are a few questions frequently asked by lipid researchers: What percentage of gene products in the studied genome are lipid related, what these gene products are and so on. Given a comprehensive list of lipid-related GO terms, the above questions can be answered using GO annotations. We have calculated the percentage of lipid-related gene products in several major genomes and prepared lists of lipid-related genes for users to download.

2.4.2 Discovering novel lipid-related molecular functions Because a BP is a series of events accomplished by one or more ordered assemblies of molecular functions, gene products that participate in a lipid-related BP are likely to have some sort of lipid-related molecular function. In other words, gene products annotated to some BP+ terms but not annotated to any MF+ term yet are likely to have some undiscovered MF+ function. Thus, based on these gene products and their associated GO terms, a hypothesis can be formulated by an expert lipid biologist where these gene products are predicted to have certain lipid-related molecular functions, which can be validated by wet lab research. We have predicted such gene products for user to download. For more information, please check our Web site or appendix.

3 DISCUSSION

Because of the sheer number (~40 000 as of April 2013) and specificity of GO terms, giving each GO term a lipid-related or non-lipid-related label is a non-trivial task, especially if performed manually. To overcome the issue, we adopt the following incremental expansion strategy. With the help of an expert curator on lipid, we obtained a list of high-quality manually curated lipid-related terms and non–lipid-related terms as the initial gold standard (GO version: June 2009). The observation that GO terms annotated to similar sets of gene products are likely to be correlated (Lord *et al.*, 2003) make it possible for us to develop a machine learning method that can be used to expand the list of lipid-related terms from the gold standard. Those terms predicted most likely to be lipid related were examined by a human curator following the curation rules (which is available at the Web site or in the appendix) so that the class labels could be confirmed. The procedure was repeated until no more lipid-related terms were found. The rest of the terms, likely to be non–lipid-related, were not manually examined and were assigned the class label ‘unknown’. For those terms with the class label ‘unknown’, there could be some missed lipid-related terms (i.e. false negatives). We did sample one hundred terms randomly but no hit was found (GO version: April 2013).

4 CONCLUSION

LipidGO serves as a platform allowing lipid researchers to get information about lipid-related GO terms and gene products. The usefulness of lipid-related GO terms is demonstrated in the applications given earlier in the text, and we believe there is much more, especially when used in conjunction with other bioinformatics tools.

Funding: Singapore National Research Foundation CRP grant (NRF-G-CRP-2007-04-082(d)) (in part) and a Singapore Ministry of Education tier-2 grant (MOE2012-T2-1-061).

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Carbon,S. *et al.* (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.

Goh,W.W. *et al.* (2013) Enhancing the utility of Proteomics Sig-nature Profiling (PSP) with Pathway Derived Subnets (PDSs), performance analysis and specialised ontologies. *BMC Genomics*, **14**, 35.

Lord,P.W. *et al.* (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.

Wenk,M.R. (2005) The emerging field of lipidomics. *Nat. Rev. Drug Discov.*, **4**, 594–610.