Planteome Gene Annotation Enrichment Analysis

Botong Qu
Jaden Diefenbaugh
Eugene Zhang
School of Electrical and
Computer Engineering
Oregon State University
Corvallis, Oregon 97331

Justin Elser
Pankaj Jaiswal
Department of Botany
and Plant Pathology
Oregon State University
Corvallis, Oregon 97331

Seth Carbon
Chris Mungall
Lawrence Berkeley
National Laboratory
Berkeley, CA 94720

***Abstract*—Annotation enrichment analysis of a gene list helps biologists to identify the potential biological functions associated with it. With the extensions of plant ontology categories, the discovery of significant ontology terms associated with a gene list becomes more and more informative. We introduce a tool to help biologists to find out these terms based on the expanding ontology database of the Planteome project. In addition, we propose some new visualization schemes to help users construct a meaningful interpretation of the results guided by the ontology tree.**

***keywords*—Ontology, Plants, Term enrichment analysis, Visualization**

I. Introduction

Gene annotations are analyzed and explored by gene curators from all over the world. Finding and visualizing the useful information from the annotations has been a hot topic for decades. The *Common Reference Ontologies and Applications for Plant Biology* benefits biologists to be able to discover enriched biological ontology terms among all provided ontologies (*Gene Ontology*, *Plant Ontology*, *Trait Ontology*, *Environment Ontology*, etc.). To assist this analysis process, we provide a gene annotation enrichment analysis tool which uses Fisher's exact and chi-squared methods to statistically analyze all annotation data. Then, we visualize the results two ways: 1) Highlighting the enriched terms among all ontology terms in the database to emphasize relative positions of the enriched terms. 2) Considering the cut-off p-value as a basis of an uncertainty factor when visualizing the tree structure in order to conveniently focus on the interesting terms.

II. Analysis Model and Methods

In our tool, we provide two common analysis methods to find the enriched terms: the Fisher's exact test and the chi-squared test [1]. To apply these statistical analysis methods, the formulation of a contingency table is necessary. In our system, we create the contingency table (table I) similar to ones used in [2] and [3]. For one specific ontology term and n genes, all genes in the database (N) are classified into four categories: the genes annotated to the term and in the input gene list (m), the genes not annotated to the term and in the input gene list ($n - m$), the genes annotated to the term and not in the input gene list ($k - m$), the genes not annotated to the term and not in the input gene list ($N - n - k + m$). This 2 by 2 table is the contingency table used to calculate the p-values for each term. If the p-value is bigger than the user chosen cut-off value (0.01 or 0.05), the term is not enriched by the gene list.

TABLE I
The contingency table for a ontology term

	Input Genes	**Not Input Genes**	**Sum (Ref)**
Annotated	m	k-m	k
Not annotated	n-m	(N-n)-(k-m)	N-k
Sum	n	N-n	N

With the number of genes annotated to the term inside the gene list (m), the total number of genes annotated to the term in the whole database (k), the number of input genes (n) and total number of genes in the database (N), Fisher's exact test is defined as equations 1 and 2. The $H(m, k, n, N)$ represents the hypergeometric distribution.

$$H(m,k,n,N) = \frac{\binom{k}{m} \times \binom{N-k}{n-m}}{\binom{N}{n}} \tag{1}$$

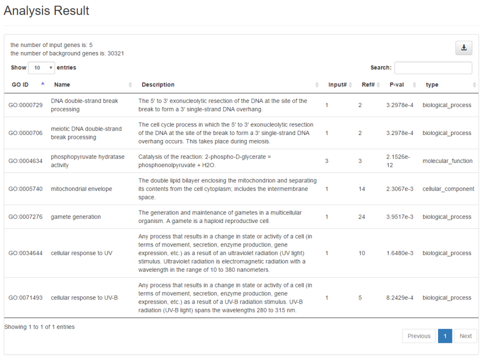
$$p - value = \sum_{i=m}^{k} H(i,k,n,N) \tag{2}$$

Based on the contingency table, we calculate the expected value of the cell that represents the number of genes annotated to the term and inside the input list by $n \times \frac{k}{N}$, then we construct an expected contingency table by fixing the margin values k, n, and N and using the calculated expected value to calculate all other three cells. Then we calculate the χ^2 value with equation 3, and then transfer it to p-value for 1 degree of freedom (a 2 by 2 table always has a freedom of 1).

$$\chi^2 = \sum_{all\ cells} \frac{(expcted - observed)^2}{expcted} \tag{3}$$

Each of these two methods has its own strengths and weaknesses. The Fisher's exact test can be applied when the input genes number is small and provides an exact calculation of the significance of the null hypothesis. But when the sample is large or the data is well balanced, the Fisher's exact test becomes computationally costly for the factorial calculation involved. On the other side, the chi-squared test can be applied to large data samples but can only give an approximation of the significance. Both methods could be used to reject the null hypothesis that the data are independent, i.e. the input genes don't enrich the ontology term.

After inputting an interesting gene list, the server will query graphically among all the annotation data, i.e. transfer all annotations of an ontology term to its parents to make sure the indirect annotations are involved in the analysis. The final analysis is as shows in Fig. 1.



Analysis Result

the number of input genes is: 5
the number of background genes is: 30321

Show 10 entries

Search:

GO ID	Name	Description	Input#	Ref#	P-val	type
GO:0000729	DNA double-strand break processing	The 5' to 3' exonucleolytic resection of the DNA at the site of the break to form a 3' single-strand DNA overhang.	1	2	3.2978e-4	biological_process
GO:0000706	meiotic DNA double-strand break processing	The cell cycle process in which the 5' to 3' exonucleolytic resection of the DNA at the site of the break to form a 3' single-strand DNA overhang occurs. This takes place during meiosis.	1	2	3.2978e-4	biological_process
GO:0004634	phosphopyruvate hydratase activity	Catalysis of the reaction: 2-phospho-D-glycerate = phosphoenolpyruvate + H2O.	3	3	2.1526e-12	molecular_function
GO:0005740	mitochondrial envelope	The double lipid bilayer enclosing the mitochondrion and separating its contents from the cell cytoplasm; includes the intermembrane space.	1	14	2.3067e-3	cellular_component
GO:0007276	gamete generation	The generation and maintenance of gametes in a multicellular organism. A gamete is a haploid reproductive cell.	1	24	3.9517e-3	biological_process
GO:0034644	cellular response to UV	Any process that results in a change in state or activity of a cell (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of an ultraviolet radiation (UV light) stimulus. Ultraviolet radiation is electromagnetic radiation with a wavelength in the range of 10 to 380 nanometers.	1	10	1.6480e-3	biological_process
GO:0071493	cellular response to UV-B	Any process that results in a change in state or activity of a cell (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a UV-B radiation stimulus. UV-B radiation (UV-B light) spans the wavelengths 280 to 315 nm.	1	5	8.2429e-4	biological_process

Showing 1 to 1 of 1 entries

Fig. 1. Analysis result of a set of genes

III. Analysis Visualization

Besides the detail information of enriched ontology terms, there are two other kinds of information to be explored. First, the relationship among enriched terms and their corresponding significant levels, the significant levels are not only limited to the p-values, but also the number of input genes annotated to a particular ontology term. Second, the relationship between enriched terms and the whole reference data. We want to apply two visualization methods to help users efficiently perceive these information.

A. Enriched Ontology Branch Visualization

Biological ontology terms are always organized in a hierarchial structure, i.e. each ontology term inherits the properties of their parents and differs with its siblings in some functionalities. Since each ontology term can have multiple parents and siblings, the research of the enriched ontology branch of a set of genes facilitates biologists to explore the potential functions associated to the genes and can be applied to find featuring genes in it. To visualize the enriched branch, we apply a hairball style visualization (similar to Fig. 2) to all the ontology terms included in the database and highlight the ones that are significant to our input genes.

B. Uncertainty Visualization

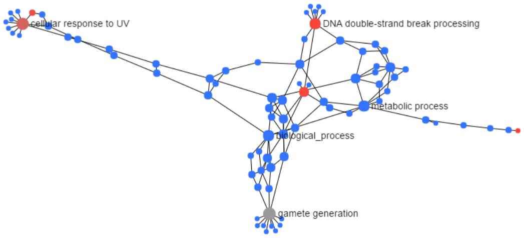


Fig. 2. A hair-ball style visualization of mega data

The hierarchical visualization of the analysis results (e.g. Gene Ontology terms) is a common method to facilitate users to explore the biological meanings behind the gene lists [4]. The tree structure graphs (as Fig. 3a shows) describe the hierarchical structured ontology terms pretty well and have a common use in analysis tools ([4], [5], [6]). However, there are some shortages in these visualization results. For example, in Fig. 3(a), the relative low significant terms (less red ones) could be distractive if the users only want to focus on the most significant terms. Also, the fixed cut off p-values make the visualization results not flexible enough. Therefore, it would be useful if we consider the cut-off p-value as a uncertainty factor and graph it. In this way, users are able to set an interesting significance value range, then the visualization results will re-arrange the focused terms to the center (as shown in Fig. 3b) to help biologists easily study them. The structure relationship between them and the significant levels calculated are always preserved to provide correct hierarchical information.

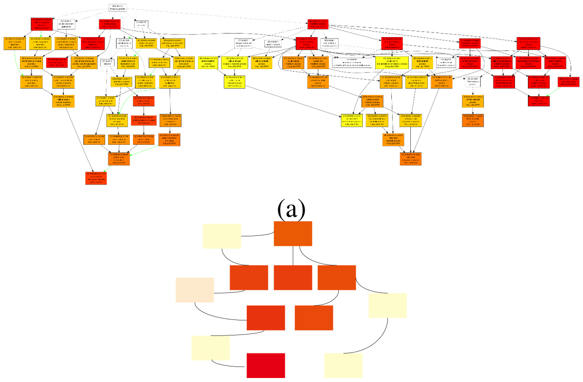


Fig. 3. a) visualization from AgriGO (b) uncertainty visualization with re-arranging the layout

References

[1] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic acids research*, vol. 37, no. 1, pp. 1–13, 2009.

[2] I. Rivals, L. Personnaz, L. Taing, and M.-C. Potier, "Enrichment or depletion of a go category within a class of genes: which test?" *Bioinformatics*, vol. 23, no. 4, pp. 401–407, 2007.

[3] G. Mi, Y. Di, S. Emerson, J. S. Cumbie, and J. H. Chang, "Length bias correction in gene ontology enrichment analysis using logistic regression," *PloS one*, vol. 7, no. 10, p. e46128, 2012.

[4] Z. Du, X. Zhou, Y. Ling, Z. Zhang, and Z. Su, "agrigo: a go analysis toolkit for the agricultural community," *Nucleic acids research*, p. gkq310, 2010.

[5] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, "Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists," *BMC bioinformatics*, vol. 10, no. 1, p. 48, 2009.

[6] S. Maere, K. Heymans, and M. Kuiper, "Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks," *Bioinformatics*, vol. 21, no. 16, pp. 3448–3449, 2005.