Proposal for a Gene Enrichment Analysis Assistant Tool

A. Cankaya[1], **and R. Shankar**[1]

[1]Department of Computer & Electrical Engineering and Computer Science,

Florida Atlantic University, Boca Raton, FL, USA

Abstract - *The gene enrichment analysis process involves using a set of genes to generate a list of Gene Ontology (GO) terms that are statistically significant. By performing enrichment using multiple sets of gene terms associated with different risk factors & diseases and examining how the significance of GO terms vary across enrichment result sets, it is possible to gain insight into shared biological pathways.*

Enrichment has traditionally been a process heavy on human-centric analysis. We first provide an example of this manual process involving the enrichment of gene terms related to certain specific focus areas, viz., atherosclerosis regression, obesity, & waist circumference.

We then propose the creation of a software tool that can assist users in semi-automating the enrichment process. The tool will enable comparison of enrichment results by generating combinations of input gene sets, performing enrichment using already available web interfaces, and quickly highlighting the major differences in significance between results lists.

Keywords: Gene Ontology, gene enrichment, automation, bioinformatics, obesity, atherosclerosis

1 Introduction

Gene enrichment analysis is the process of identifying Gene Ontology (GO) terms (biological processes, cellular components, and molecular functions) that are enriched (i.e., statistically significant) for a given set of genes [1]. A set of gene terms related to a topic is provided as an input by the user and a list of GO terms with statistical calculations is returned. Knowing what GO terms are identified as significant can allow a user to generate a functional profile that provides insight into the biological roles influenced by the input set of genes.

This process currently requires a human to manually curate the sets of initial gene terms and then input them into software tools that generate a list of GO terms and significance values (p-values) for each set of gene terms. The human user then manually parses through the lists of GO terms and identifies how the significance varies across the results sets dependent upon the set of genes used for the enrichment.

In Part 1 we explain the concept of the gene enrichment analysis process and our desire to create automation tools supporting it. Part 2 explains the current manual enrichment process and provides an example involving the regression of atherosclerosis. In Part 3 we explain our proposed plan for utilizing existing web technologies for the development of an assistant tool and in Part 4 we summarize our conclusions and development plans.

1.1 Gene enrichment analysis

A set of gene terms can be gathered to represent a disease, pathology, bio-measurement, or other identifiable phenotype. For example, a set of gene terms associated with increased risk of breast cancer or associated with a high body mass index. A set of gene terms can also represent a confounder - a biological measurement or process believed to be independently related to another pathology. For example, a confounder of obesity could be hypertension - having high blood pressure might indicate a higher risk for obesity, but it is not a root cause of obesity.

A user can gather a set of input genes based on a risk factor (e.g., obesity) or a disease (e.g., diabetes), and perform enrichment analysis to gain insight into the biological processes associated with those genes. Multiple sets of gene terms, each set representing a different topic of study (for e.g., obesity, diabetes, hypertension), can be gathered and used to perform enrichment analysis. By comparing how the enrichment results differ for each set of gene terms we can see how the biological processes associated with each topic differ from each other.

Each individual set of gene terms can also be combined with other such sets to form new sets of genes representing the combined effects of more than one topic (obesity + diabetes, obesity + hypertension, diabetes + hypertension, etc.). By then comparing how the results change depending upon how combinations of the gene sets are used as input, it is possible to gather insight into the underlying biological roles of the gene terms, how their related biological pathways interact and how they are regulated in the body.

1.2 Desire for automation

Once sets of gene terms have been gathered, the time intensive part of the current enrichment analysis process is

parsing the GO term results for each set of gene terms and identifying commonalities and differences in significance values, cluster & genome frequencies, & number of genes annotated.

2 Current manual Enrichment process

2.1 Step 1: Collect sets of gene terms

The traditional method for performing enrichment analysis starts with a manual collection of gene terms related to the topic of interest. Gene terms can be gathered by doing a literature review and identifying genes referenced in journal articles or clinical studies related to the topic. There are already some automation tools available for this process such as MEDIE [2], a natural language semantic search for MEDLINE articles.

As an example, to discover shared biological pathways involved in atherosclerosis regression, obesity, and a confounder of waist circumference, three sets of gene terms are curated. These sets of genes are available at our GitHub repository [3]. Using MEDIE we perform a literature search for genes related to atherosclerosis regression that results in a set of 27 gene terms. This set of gene terms are referred to as "Set A".

Separately, a set of 64 gene terms related to obesity, "Set B", is curated by reading a comprehensive literature review of obesity related studies. This set of gene terms is composed of 50 genes identified through GWAS methods [4] and 18 genes identified through pre-GWAS methods [5].

Finally, a third set of gene terms, "Set C", related to the potential confounder of waist circumference measurement (WC), is gathered through the use of the GWAS catalog [6] which referred us to a study [7] identifying 76 gene terms associated with increased WC measurements.

Table I.
The first several gene terms from Set A (atherosclerosis regression), Set B (obesity), and Set C (increased waist circumference). Note that the terms in each set are not in any particular order.

Set A	Set B	Set C
APOB	LEPR	ANKS1A
APOE	POMC	COBLL1
MSR1	FTO	MAP3K1
SCARB1	MC4R	RSPO3
CCR7	NEGR1	ANAPC13
...	...	...

2.2 Step 2: Generate combinations of gene terms

Performing enrichment analysis on Sets A, B, or C could provide us insight into biological pathways related to each set's topic. However, in order to identify potential shared biological pathways between atherosclerosis regression, obesity, and WC measurement, the three sets of gene terms must be grouped together in combinations to form new sets that will then be used for enrichment analysis.

For example, Set A (27 gene term related to atherosclerosis regression) and Set B (64 gene terms related to obesity) are combined to form Set AB (91 gene terms related to atherosclerosis regression and/or obesity). Similarly created are Set AC (103 gene terms related to atherosclerosis regression and/or WC measurement), Set BC (140 gene terms related to obesity and/or WC measurement), and Set ABC, (167 gene terms related to atherosclerosis regression, obesity, and/or WC measurements). These combinations are visualized as a Venn diagram in Figure 1.

For the current manual process, this step involves simply combining the gene sets together in a text document. Note - because we are not currently providing the gene terms in any particular order (i.e., no ranking), the order in which we combine the gene term sets does not matter. This also means that, for example, Sets AB and BA are equivalent.

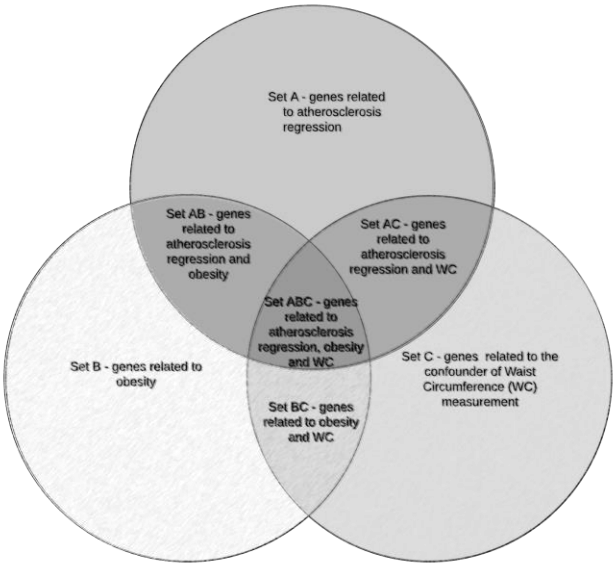

Figure 1. A diagram showing how sets of gene terms related to atherosclerosis regression, obesity, and waist circumference, Set A, B, & C, are combined together to form Set AB, AC, BC, & ABC.

2.3 Step 3: Use Web tools to perform enrichment

The easiest way to perform enrichment analysis is to input a set of gene terms into an online tool such as the "Generic Gene Ontology Term Finder" provided by Princeton [8]. The user can leave the default option values and immediately get back a set of GO terms related to the genes entered as input. Additional

information for each GO term is provided to the user, such as the p-value, cluster & genome frequency, and a list of the input genes annotated to the GO term.

By default, the Princeton GO term finder tool includes associations Inferred from Electronic Annotations (IEA associations). The evidence code IEA is used to indicate computational function annotation, i.e., inferences made without human oversight. Over 99% of all GO annotations are created this way [1]. There are strengths (many more associations, for example) and weaknesses (lack of experimental support, for example) to the use of IEA [1]. The user can choose to exclude IEA associations and perform enrichment analysis. By performing separate enrichment for both IEA included and excluded there is potential for insight by comparing the differences in the significance of GO terms found in the two result sets.

For our example, performing enrichment using the gene terms of Set A (atherosclerosis regression) generates significant GO terms related to lipids - GO:0055088 lipid homeostasis (p-value 3.08e-21), GO:0006629 lipid metabolic process (p-value 6.61e-15), and triglycerides - GO:0006641 triglyceride metabolic process (p-value 6.59e-14).

Performing enrichment using Set B (obesity) generates GO terms related to hormones - GO:0009725 response to hormone (p-value 9.58e-07), GO:0032870 cellular response to hormone stimulus (p-value 3.64e-06) and food & nutrients - GO:0032098 regulation of appetite (p-value 6.14e-06), GO:0007631 feeding behavior (p-value 3.87e-05), GO:0031667 response to nutrient levels (p-value 1.44e-05).

2.4 Step 4: Compare enrichment results

For our example involving atherosclerosis regression (A), obesity (B), and waist circumference (C), we have six combined gene term sets - A, B, AB, AC, BC, ABC. We do not consider Set C alone because we are defining it as a confounder.

When we combine the atherosclerosis regression gene terms with the obesity gene terms (Set AB) and perform enrichment we generate GO terms with annotated genes from both topics. For the most part the significance of these GO terms is less than when enrichment was performed on a single set alone (Set A alone or Set B alone). However, sometimes a GO term will see its significance increase when enriched with additional gene terms associated to a different topic and this can be evidence towards a shared biological pathway.

For our example, when enrichment is performed using Set AB, we find the term GO:0031667 response to nutrient levels to have its significance increased to 1.10e-08 when compared to Set B (obesity) being enriched alone (Set B p-value 1.44e-05). This provides evidence that pathways related to nutrient levels are likely involved in both the atherosclerosis regression and obesity development processes.

Another source for comparison is between results sets where IEA associations are excluded versus included. This is because electronic annotations are usually associated with the maintenance phase of a disease as opposed to the development phase more commonly associated with non-electronic annotation (See [4] and [5] for this comparative focus).

For example, when performing enrichment using Set A (atherosclerosis regression), GO:0006641 triglyceride metabolic process is found to have a p-value significance of 9.30e-07 when IEA associations are excluded (with 6 of 27 genes annotated to it), but when IEA associations are included the significance improves to 6.59e-14 (with 10 of 27 genes annotated). This could be evidence that the metabolism of triglycerides is related to atherosclerosis regression, but not necessarily the development of atherosclerosis.

Enrichment using Set AB (atherosclerosis regression + obesity) leads to some gene terms originally identified as related to obesity exclusively as now being annotated to GO terms previously identified through enrichment of atherosclerosis regression gene terms alone. For example, the most significant GO term found is GO:1905952 regulation of lipid localization. It was previously annotated with 15 of 27 genes when Set A was enriched. Now when the Set A terms are enriched along with Set B gene terms in Set AB it is further annotated with an additional 6 gene terms related to obesity. These same obesity gene terms are annotated to various other lipid related GO terms as well. This provides evidence for a link between lipid pathways, the obesity phenotype and atherosclerosis.

Set AC (atherosclerosis + waist circumference) enrichment show that most of the significant GO terms found when performing enrichment using the atherosclerosis terms (Set A) or the WC terms (Set C) alone are still found to be significant in the combined (Set AC) enrichment. In fact, almost all of the significant GO terms from each individual set is now found to be further annotated with additional gene terms from the other set. Two of the GO terms see significant improvements in the number of genes annotated. First is GO:0043062 extracellular structure organization which is originally annotated with 10 of 27 genes (p-value 1.32e-07), but then when enriched along with the WC gene terms is now annotated with 20 of 103 genes (improved p-value of 2.25e-10). Second is the previously mentioned GO:0006629 lipid metabolic process which is originally annotated with 20 of 27 atherosclerosis gene terms, but is now also annotated with an additional 7 genes from the waist circumference set.

We have yet to perform enrichment analysis on the remaining two combination sets - Set BC (obesity + waist circumference) and Set ABC (atherosclerosis regression + obesity + waist circumference).

3 Enrichment automation plans

3.1 Step 1: Collect sets of gene terms

In the simplest automated process, the user would upload their own sets of gene terms that they had already researched and curated. For each set of gene terms, the user would indicate the pathology associated with the gene terms and this pathology would be used as a title for the result set.

More advanced options would allow the user to indicate if the pathology is to be treated as a confounder, whether the gene terms were gathered using pre-GWAS and/or GWAS methodologies, and whether IEA associations should be included or not.

The tool would take this user input and for each set of gene terms create a JavaScript Object Notation (JSON) encoded object. JSON is an open source web standard for communicating information between servers in a human readable form. The tool would create an array (user_input_array) of these JSON objects where each object (user_input) has a structure such as:

```
user_input = {
        'id': integer,
        'title': string,
        'genes': Array<string>,
        'p-cutoff': number
        'confounder': boolean,
        'IEA_included': boolean
}
user_input_array = Array < user_input_A,
                           user_input_B, user_input_C >
```

where 'id' is auto generated, 'title' is the pathology entered by the user, 'genes' is the set of gene terms entered by the user, 'p-cutoff' is the minimum p-value allowed, 'confounder' is set to true if the user indicates that the set should be treated as a confounder, and 'IEA_included' is set to true if the user indicates that IEA associations should be included.

3.2 Step 2: Generate combinations of gene terms

For our example, we now have an array of JSON objects (user_input_array) with a length of three representing Set A (atherosclerosis regression), Set B (obesity), and Set C (waist circumference measurement as a confounder). To see how GO terms can be enriched or depleted depending upon the gene terms used as input, we generate every possible combination of Sets A, B, & C. We then double the number of sets to also consider IEA associations included and excluded.

These generated combination sets get added to a new object called combined_input_array which now represents sets A, B, AB, AC, BC, & ABC. To represent the input of the IEA associations boolean, a copy is made of each set and the IEA association boolean is toggled. For our notation we add to the set name a suffix of -i for IEA included or -e for IEA excluded, giving us a total of 12 sets.

For our example involving Set A and B, with Set C as a confounder, the combined_input_array becomes:

```
combined_input_array = Array < user_input_A-i,
        user_input_A-e, user_input_B-i,
        user_input_B- e,user_input_AB-i,
        user_input_AB-e,user_input_AC-i,
        user_input_AC-e,user_input_BC-i,
        user_input_BC-e,user_input_ABC-i,
        user_input_ABC-e >
```

3.3 Step 3: Use Web API to perform enrichment

The actual enrichment process can be accomplished using our own server hosted code implementing algorithms of our choosing; or more cheaply, we can take advantage of already available web Application Programming Interfaces (API) provided by online tools such as AmiGO [9] and QuickGO [10].

We simply need to encode our input data as required by the API we plan to use and then it is a matter of using our programming language of choice to generate standard web requests.

For each set of gene terms, we will receive back a set of GO terms along with each GO term's significance (p-value), cluster & genome frequencies, and a list of the gene terms annotated to it. We can then parse and reencode these enrichment results into a JSON encoded object ready for automated analysis.

```
GO_item =  {
        'id': integer,
        'GO_id': string,
        'GO_term': string,
        'p-value': double,
        'annotated_genes': Array<string>
}

total_results = {
        'id': integer,
        'gene_set_id': integer,
        'results': Array<GO_item>
}
```

3.4 Step 4: Perform automated comparison of results and highlight areas of change

The primary comparison calculation that should be performed first are the changes in a GO term's significance value (p-value) between sets. The user should be able to see at a glance how a specific GO term's significance value changes for each set. The list of genes annotated should be displayed with the additions highlighted should also be displayed for the user. The user should be able to select a GO term and quickly see how the list of genes annotated to the term varies across result sets. The change in cluster and genome frequency can also be provided.

Once we have a set of enrichment results, for each set of gene terms, the tool should begin doing a comparison analysis. The clearest user interface to display all these differences in between results sets to the user has yet to be finalized. How exactly this will be visually displayed for the user will require the development of mockups and user story feedback.

We will take inspiration from other attempts at visualization of gene enrichment results. An overview of the various GO annotation visualization methods [11] identifies generic network graphs created by tools such as Cytoscape [12], visual overlays used by tools such as GOrilla [13], word clouds used by ReviGO [14], and tree maps as also seen in ReviGO,

The bubble chart appears to be a standard form for displaying GO annotation relationships across two dimensions. Each bubble is a GO term and each axis represents a different measurement or result set which causes the shape and size of the bubbles to be indicative of the relationship between the given GO term in each of the two dimensions. The measurement or set represented by an axis can be chosen by the user on-demand using toggle buttons.

There should also be a results comparison summary page that shows a list of the big differences between sets. This could take the form of a list of the top GO terms that show the greatest variation in significance, number of genes annotated, and cluster & genome frequencies. A color scheme could be implemented so that the GO term can be highlighted a shade of green if it has gained significance or a shade of red if the significance has decreased.

3.5 Other technical considerations

For a local application we envision the use of a JavaScript framework to create a web based front end for user input and display of results. Meteor [15] is an open source web framework written in Node.JS for creating applications designed to run inside a web browser window. In order to run a Meteor application a user must first install the Meteor framework on their local machine. The user then launches their web browser of choice to interact with the Meteor based server.

Another option is Electron [16] which combines Node.JS with the Chromium browser in a single software framework. Electron applications can be packaged into an executable file that the user can run on demand without installation or setup. Electron natively supports TypeScript which allows it to run other TypeScript based web frameworks such as Angular [17].

An Electron-Angular setup would include the use of the ngrx-store package [18] for maintaining a state object containing the input and output. The actual comparison of GO term values across result sets could also be done in Typescript for simplicity and to provide a proof of concept, however another language may provide better performance for larger scale enrichment analysis. Similarly, we plan to initially use the publicly provided web API's to perform enrichment, but better performance could be made in the future with a privately hosted solution running its own gene enrichment environment and algorithms.

3.6 Additional features

Once the basics of the automated enrichment process has been developed, we have some ideas for additional features. The ultimate goal would be to allow a user to enter two or more search terms for automated gene list gathering. This would involve use of text mining and natural language algorithms to review literature and collect relevant gene terms. The user could also enter one or more species to consider. Results could also automatically be sent to ReviGO to provide the user with visual summaries.

For now, we are also not considering the gene terms to be entered in any specific order or ranking. Some enrichment tools such as GOrilla allow gene terms to be entered in order by decreasing importance. Allowing the user to enter ranked lists opens up an entire new area for computational analysis because we can create a near endless number of permutations of gene rankings.

A user could enter a single list of ranked gene terms and our tool could try to rearrange all of the genes to maximize a specific GO term's p-value or overall number of annotated genes. Doing this might allow insights to be developed by performing the enrichment process "in reverse" - the user enters a list of genes and learns which genes are the most important based on which ranking provides the best result for a known GO term.

3.7 Validation methods

To validate the real-world accuracy of the GO term results found be enriched with our sets of gene terms we plan to make use of the 'NOT' qualifier provided by the GO Annotation File (GAF) format [19]. This qualifier is added to a GO annotation if "a gene product has been experimentally demonstrated not to be able to carry out a particular activity". Because the 'NOT' qualifier has been experimentally validated, we have a higher level of trust in its validity.

For every GO term found to be enriched significantly, a search should be performed (using AmiGO or a similar web tool) to find annotations between the GO term and any of the genes on the list of gene terms we provide. For every annotation found a check will be made to ensure that the 'NOT' qualifier is false. If an annotation between a gene and a GO term is found with the 'NOT' qualifier as true, then that GO term should be removed or otherwise highlighted as likely invalid.

This same principal can be used for testing and validating our software prior to being released. Sets of gene terms with multiple annotations containing the 'NOT' qualifier can be collected and used as input for enrichment. In theory, the enrichment process should not identify any of the 'NOT' qualifier annotations as being significant. If the enrichment algorithms do repeatedly identify as significant these invalid GO terms it can be a warning sign that something is wrong with the algorithm or the tools implementing it.

4 Discussion and conclusions

Gene enrichment analysis demonstrates the potential to discover additional insights into biological processes using data already published. A set of genes related to a pathology can be curated from literature and used to identify statistically significant Gene Ontology (GO) terms.

Sets of genes on different topics can be enriched individually and together in combined sets. By evaluating how the enrichment results vary depending upon the gene terms used as input it is possible to gain insight into the shared biological pathways underlying the pathologies related to the gene terms gathered.

This process of identifying across multiple result sets the GO terms expressing the largest changes in significance, genetic cluster & genome frequencies, or number of genes annotated, has traditionally required hours of human inspection. There is potential to help automate the cumbersome aspects of the enrichment analysis process by developing a software tool that can quickly analyze and identify for users the GO terms that express the largest statistical changes across result sets.

Using the tool, users can provide sets of gene terms related to different topics and then have multiple gene enrichments performed using already available web technology. Users will be able to easily visualize the GO terms expressing the biggest changes between enrichment result sets and spend more of their time trying to gain understanding about biological processes behind the pathologies.

5 References

[1] Dessimoz, C., and Skunca, N., Eds, The Gene Ontology Handbook, Springer Open and Humana Press: New York, 2016.: https://link.springer.com/book/10.1007/978-1-4939-3743-1

[2] "MEDIE - Semantic retrieval engine for MEDLINE". National Center for Text Mining. https://www.nactem.ac.uk/medie/search.cgi

[3] https://github.com/acankaya2017/gea

[4] Hedman, Asa K, et al. The Genetics of Obesity, Chapter 3 - Genome-Wide Association Studies of Obesity. 2014

[5] McCormack, Shana E., The Genetics of Obesity, Chapter 1 - Genetic Variation and Obesity Prior to the Era of Genome-Wide Association Studies. 2014.

[6] "GWAS catalog". The European Bioinformatics Institute. https://www.ebi.ac.uk/gwas/

[7] "Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits". Nat Commun. 2017 Apr 26; doi: 10.1038/ncomms14977.

[8] "Generic Gene Ontology Term Finder". Princeton University. https://go.princeton.edu/cgi-bin/GOTermFinder

[9] "Programmatic Access to Gene Ontology". The Gene Ontology Consortium. http://geneontology.org/docs/tools-guide/

[10] "QuickGO API". The European Bioinformatics Institute. https://www.ebi.ac.uk/QuickGO/api/index.html

[11] F. Supek & N. Škunca. "Visualizing GO Annotations". GO Handbook. https://arxiv.org/pdf/1602.07103.pdf

[12] "Cytoscape: An Open Source Platform for Complex Network Analysis and Visualization". https://cytoscape.org/

[13] 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 2009. http://cbl-gorilla.cs.technion.ac.il/

[14] F. Supek, M. Bošnjak, N. Škunca, T. Šmuc "REVIGO summarizes and visualizes long lists of Gene Ontology terms" PLoS ONE 2011. doi:10.1371/journal.pone.00218

[15] Meteor Development Group. https://www.meteor.com/

[16] Electron. https://electronjs.org/

[17] Angular. https://angular.io

[18] "NGRX – Reactive Extensions for Angular". https://github.com/ngrx

[19] "Introduction to GO Annotations". The GO Consortium. http://geneontology.org/docs/go-annotations/