Deciphering gene sets annotations with ontology based visualization

A. Ayllón-Benítez\textsuperscript{1,2}, P. Thébault\textsuperscript{1}, J.T. Fernández-Breis\textsuperscript{3}, M. Quesada-Martínez\textsuperscript{3}, F. Mougin\textsuperscript{1,2} and R. Bourqui\textsuperscript{1}
\textsuperscript{1} Univ. Bordeaux, Inserm, Bordeaux Population Health Res. Center, ERIAS team, UMR1219, France
\textsuperscript{2} Univ. Bordeaux, LaBRI CNRS UMR5800, France
\textsuperscript{3} Univ. Murcia, TECNOMOD group, Spain
{aaron.ayllon-benitez, thebault, fmougin, bourqui}@labri.fr, {jfernand,manuel.quesada}@um.es

Abstract—Nowadays, one of the main challenges in biology is to make use of several sources of data to improve our understanding of life. When analyzing experimental data, researchers aim at clustering genes that show a similar behavior through specific external conditions. Thus, the functional interpretation of genes is crucial and involves making use of the whole subset of terms that annotate these genes and which can be relatively large and redundant. The manual expertise to clearly decipher the main functions that may be related to the gene set is time-consuming and becomes impracticable when the number of gene sets increases, like in the case of vaccine/drug trials. To overcome this drawback, it may be necessary to reduce the dataset with the aim to apply visualization approaches. In this paper, we propose a new pipeline combining enrichment and annotation terms simplification to produce a synthetic visualization of several gene sets simultaneously. We illustrate the efficiency of our method on a case study aiming at analyzing the immune response in diseases.

I. INTRODUCTION

In clinical science, [1], the discovery of new medical drugs or vaccines involves carrying out large scale testing of new therapies. The evaluation of their impacts requires to compare experimental data from at least two populations, one for the control cases (e.g., healthy population) and one for a population with a chronic disease. More formally, researchers are interested in analyzing the relation between genotype and phenotype while comparing two populations. The genotype of an organism is given by its collection of genes while its phenotype corresponds to the observable features that are encoded by those genes. The comparison of these datasets is crucial to identify the genes involved in the immune response by comparing the expression levels of genes in both populations.

The key idea of such studies is then to identify groups of impacted genes having common biological functions in order to better understand their roles in the phenotype. These groups (gene sets) are formed on the basis of experimental data by using clustering algorithms or statistical approaches [2]. Then, their analysis involves associating them to one or several annotation terms that describe their function. The annotation terms can come from different databases (for an illustration, see the NCBI biomedical portal [3] that provides easily accessible information and services related to biomolecules). Focusing on Gene Ontology Annotation (one of the most popular annotation database), on average, each human gene is described by 10 terms, giving an information with varying degrees of detail. At the gene level, the attribution of terms to a gene mostly requires computational methods that make use of experimental and proved information. The terms of specific genes can be transferred to genes related to them from an evolutionary point of view. At the gene set level, automatic enrichment methods [4] make use of lists of terms (coming from each gene) to produce a large amount of annotation terms that hinder the analysis, and therefore the understanding of their biological roles. In particular, two gene sets with similar biological roles may be annotated with different, yet similar (and even synonymous) annotation terms. A common way to reduce the complexity of such data is to organize these annotation terms using a knowledge model, or ontology, which organizes them according to their semantic meanings. Among biological ontologies, one of the most famous and also exhaustive is the Gene Ontology (GO) [5]. While structuring the annotation terms, using such an ontology increases the amount of information to be apprehended by biologists. This data complexity/size therefore urges the need of dedicated visualization techniques. In that context, such techniques mainly refer to directed acyclic graph (DAG) visualization as the aforementioned ontology has that particular topology. Many tools for the enrichment of gene sets can be found in the literature and use different approaches (the reader can refer to [4] for a review). To explore the results, some of these tools propose enhanced visualizations which can be mainly classified into four categories (according to [5], we can also consider the word cloud category): node-link diagrams, space-filling, heatmaps and semantic similarity space. In the node-link diagrams category, some methods [6], [7], [8] extract a DAG from the computed annotation terms and their relations within the ontology and draw it using a layered/hierarchical drawing algorithm. Others (e.g., [9]) link the annotation terms according to their semantic meaning (computed using the ontology) and make use of a force-directed algorithm.
that takes that similarity into account. Methods of the space-filling category use space-filling visualization techniques to render the DAG (computed using similar strategies). For instance, [9] uses treemaps that only represent two levels of the hierarchy and avoid annotation term duplication by making use of the computed p-values. And, [10] preferred using a navigable circular chart where the angular sectors represent the number of genes annotated with the term or according to a p-value. The semantic similarity space category uses PCA/MDS to represent annotation distances in scatterplots [9]. Finally, the heatmaps category proposes a binary association between genes and annotation terms [11] or groups of similar annotation terms [12] in a matrix-based diagram where a row represents a gene and a column an annotation term or a group of terms. While these methods support the visualization of annotation terms of a (narrowed) list of genes, they are not designed to handle a list of dozens or more gene sets. They are therefore not suitable for the context of this paper where the response of the organism to a vaccine, a drug or disease has to be understood as a global mechanism.

In this paper, we propose a complete pipeline that supports the analysis of gene sets for identifying their biological roles. Our method combines enrichment and annotation terms simplification to produce a synthetic visualization of several gene sets simultaneously. We believe that the exploration of data produced from drug/vaccine trials is facilitated and helps researchers to interpret their experimental results. The remainder of this paper is structured as follows. In section II, we describe in detail each step of the proposed analysis pipeline that produces a simplified hierarchy of GO terms from a list of gene sets. We then present and motivate the design of our visualization prototype in section III. We illustrate the efficiency of our method with a case study provided by analyses of immune response in diseases in section IV. Finally, we draw a conclusion and give directions for future work.

II. ANALYSIS PIPELINE

The workflow consists of three main steps (Figure 1). The first step aims at annotating gene sets with a limited number of annotation terms. Then, these terms are lexically processed in order to identify close terms within the Gene Ontology. Finally, as the Gene Ontology is large, a simplification step is performed to select only terms which are relevant for the interpretation of the input gene sets.

A. Annotating gene sets

One popular solution in biology to interpret gene sets consists in using enrichment analysis. This method is based on statistical approaches that run tests to identify the most statistically significant terms in the subset versus a reference subset (for instance, the whole terms that are used for annotating the complete collection of human genes). With enrichment tools, biomedical scientists get lists of over-represented annotation terms that: (i) guide their exploration of the repertoire of genes according to their functionality given by annotation and (ii) facilitate their interpretation of genes that were selected early in the immune response. To carry out this stage, several bioinformatics tools exist (for a review of these solutions, see [4] and about their advantages and pitfalls, see [13]).

To achieve that step of our analysis pipeline, we chose the g:Profiler [14] tool whose main advantage is its capacity to take into account several annotation databases. Using different databases allows to obtain the functional roles of the gene sets from different points of views (as gene annotations may have been done at different cell organization levels). However, such a tool provides a given gene set with various annotations even if the latter describe the same (or similar) biological functions. That drawback is handled by the next step of our pipeline, which reconciles the output annotation terms with terms from a reference ontology. Note that this
step also makes possible to use any enrichment tool for the same reason.

B. Relating annotation terms to GO terms

As previously said, the GO is the most widely used resource for annotating genes. It contains more than 40,000 terms that describe the roles of genes of any biological organism [5]. These terms are organized according to the three following disjoint categories: Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). Within each category (also called ontology), terms are connected through is_a relations, which link a given GO term to its parent term(s) and are oriented. Thus, each ontology is actually a DAG where nodes are GO terms and edges are is_a relations.

For reconciling the annotation terms provided by the enrichment method with GO terms, a lexical approach implemented within the OntoEnrich framework was applied [15]. Primarily, each word of annotation terms is normalized. Then, an exact match is searched for each annotation term. If no GO term corresponding to the annotation term is found, the latter is split in order to find partial matches. Both exact and partial matches are calculated using labels and their synonyms. Finally, partial matches are filtered by removing: (i) GO terms which are ancestors of other candidate GO terms and (ii) GO terms which are lexically included in other candidate GO terms (e.g., blood coagulation is lexically included in regulation of blood coagulation). It is noteworthy that either one, multiple or no GO terms can be obtained at the end of this step. As an illustration, if a given gene set is annotated by cell cycle, ATP bindings, OntoEnrich identifies four partial matches in the GO: cell, cell cycle, ATP binding and binding but only cell cycle and ATP binding are kept.

C. GO structure simplification

A filtering stage is then performed because only some GO terms are to be displayed, so it is not necessary to visualize the entire GO. In practice, for each GO term obtained by OntoEnrich, its most informative parent term is recovered, and this process is recursively applied until the root term is reached. For determining the most informative parent, the following IC (information content) measure is computed for each selected GO term $t$ [16]:

$$IC_{zhou}(t) = k \cdot (1- \frac{\log(f_{req}(t))}{\log(f_{req}(root))}) + (1-k) \cdot \frac{\log(d(t))}{\log(MDO)}$$

where $f_{req}(t)$ is the number of descendant terms of $t$ and $d(t)$ is the depth of $t$ within the ontology. $MDO$ is the maximal depth in the ontology and $k$ is an adjustable factor providing a weight for each item of the equation. This factor impacts the equation relying more on descendants ($k$ near 1) or depth ($k$ near 0). In our experiments, we chose 0.5 as proposed in [16]. Once this subgraph of the ontology is generated, only GO terms that are ancestors of at least two annotating GO terms are kept. These intermediate ancestors can be removed thanks to the transitivity of the is_a relationship, which makes possible to create an is_a relation between a given term and any of its ancestors.

III. Visualization

The output of the analysis pipeline described in the previous section is a hierarchical tree of GO terms whose leaves are associated to the input gene sets. Considering that each gene set is linked to its GO terms in that tree (therefore forming a DAG), the key idea of our method is to duplicate each gene set to unfold that DAG into a tree whose leaves are gene sets. Such an idea has already been used in [17], [18] to support the visualization of DAGs. This technique offers the advantage of simplifying the representation although the duplication of leaves (here gene sets) may lead to misinterpretation. In the following, we present how the produced tree is visually represented, how we support/ease the identification of duplicated gene sets and the interaction tools supported by our prototype.

A. Drawing the gene sets/GO terms association

Inspired from [17], our prototype uses two interactive and linked views (developed with D3 library [19]): a circular treemap and an indented tree. Unlike [17], [18], circular treemap has been preferred over rectangular or Voronoi
treemaps. Indeed, circular treemaps present the following advantages in comparison to other treemap approaches [20], [21]: (i) there is a clear separation between nodes, and (ii) the different levels of hierarchical data are easily interpretable. Nevertheless, this is made at the expense of a larger space usage. Thanks to our annotation strategy, the generated tree contains few hundred nodes which counterbalances that drawback. In the context of gene sets annotation, the number of genes in each set is an important information as a large gene set (in term of number of genes) may have a more generic biological role in the cell overall functioning. Therefore, larger gene sets are emphasized by setting their size proportionally to their number of genes. A collapsible intended tree view is also provided to show additional information to what is displayed within the circular treemap. In that view, each node (GO term and gene set) is represented as a rectangle and clicking on one of these rectangles expands it and displays its child nodes. That view helps the user when seeking for a GO term of interest and thus for a particular functional role that could have been impacted during the experiment (see figure 2). Such a visualization technique has been preferred over other techniques like node-link diagrams (as in [17]) because it has been shown to improve user experience/performance in usability studies of ontologies [22], [23].

![Figure 3. Zoom representing the gene sets that contain a complete or partial relation with the GO term signaling.](image)

B. Overcoming duplication issue

As mentioned above, gene sets have been duplicated and linked to each of their annotating GO terms. Such duplication may hinder a good understanding of the gene set biological roles. For facilitating the identification of duplicated gene sets, we combined a specific tree node coloring algorithm and bar charts within gene set nodes. First, tree nodes were colored using the Tree colors algorithm [24] that assigns similar colors to close nodes. The basics of this algorithm are to use the HCL (Hue-Chroma-Luminance) color space and to recursively divide a hue interval associating a node to its children (the hue interval of the tree root is set to $[0, 359]$). Then, increasing the chroma and reducing the luminance according to node depth improve the perception of depth within the tree. This algorithm is applied to the entire tree, except from the leaves that represent the gene set nodes. By setting their color to an unused color facilitates the identification of gene set nodes in the tree. Second, within each gene set node, a bar chart is displayed representing its annotation terms (using their assigned colors) and the level of confidence one could have on that annotation term (according the p-value computed at the first step of the analysis pipeline). It allows to identify gene sets annotated with several GO terms, as well as to provide a cue about the positions of these terms in the hierarchy as these positions are visually encoded by their colors. This is an important feature of our visualization because it emphasizes gene sets annotated with several yet similar terms that could improve the level of confidence of the user.

C. Interaction tools

To support the exploration in the proposed visualizations, our prototype integrates several interaction tools. Besides a classical zoom and pan interaction tool that supports basic exploration, it also integrates an interaction that allows to focus on a GO term or a gene set of interest. Clicking on a node (either a GO term or a gene set), in the treemap view, automatically zooms on that particular node of the tree. As mentioned in section III-A, the two views are linked, it therefore also expands the indented tree to display the entire path between the root node and the focused one (and the remaining expanded paths of the tree are collapsed). If the clicked node represents a gene set, all paths between the root node and all duplicates are expanded. Clicking on a node in the indented tree view also allows to zoom on the corresponding gene set duplicate in the treemap view. When considering large trees, some of the node colors may be perceptually similar (as mentioned in [24]). A last interaction tool thus tries to alleviate that drawback and highlights all duplicates in the treemap view when pointing a gene set annotated with several GO terms.

IV. CASE STUDY

To illustrate the usefulness of our pipeline, we experimented an analysis making use of the data provided by Chaussabel et al. [1] concerning the immune response within
the human population. The objective of such analysis is to identify the key regulators that manage the immune defense system. The resulting dataset is composed of 260 gene sets grouped according to their similar expression profiles through one or several of 15 experimental works (each of them corresponds to the study of a disease) using a co-expressed network approach [1].

Applying our analysis pipeline, the enrichment steps produced 1296 annotation terms while the alignment with the GO and simplification steps allowed to reduce drastically their number to 157. Figure 2 shows the treemap view representing these gene sets together with the simplified GO structure which corresponds to a tree of 782 nodes (including all gene sets duplicates). In immunology, one of the most important biological activities comes from the interaction and communication between cells. Both processes can be investigated by studying the signaling pathway from a global point of view (which corresponds to the signal transduction by which a signal is transmitted from cell to cell). The first step of our analysis was therefore to search and select the signaling annotation from the indented tree view. The resulting focused view (see figure 3) displayed seven gene sets. The indented tree shows the annotation terms computed by the enrichment tool and our reconciliation step has related them to the signaling GO term which seems relevant here. Looking at the bar charts within the gene set (in the treemap view), one can identify some gene sets where the level of confidence is high (e.g., M1.2 and M3.4) while other are low (e.g., M5.1 and M5.9). This is confirmed by the study of [1] that manually annotated M1.2 and M3.4 as interferon (which is closely related to signaling) while M5.1 was annotated inflammation and M5.9 protein synthesis. As the M1.2 gene set is known to be involved at an early stage of the immune response [1], we further focused our study on that gene set. Clicking on it in the treemap view allowed to obtain more details on that particular gene set. From the indented tree view of figure 4, one can easily identify the five annotation terms computed by our pipeline. Among these annotation terms, three out of five are from the BP ontology (with high levels of confidence) and two are from the MF ontology (with low levels of confidence). It provides a good cue on the biological role of the M1.2 gene set thanks to the three annotation terms: signaling, response to virus and negative regulation of viral genome replication.

With very few user interactions, we could retrieve the right biological role of some of the gene sets, in particular, those involved at the early stage of the immune response.

**CONCLUSION**

In this paper, we propose a new analysis pipeline that combines enrichment, GO terms alignment and simplification whose results are then displayed with a visualization prototype. That prototype uses two linked views: a treemap view that provides an overview but also display detailed information about gene sets, and an indented tree view that facilitates the search of annotation terms of interest. We illustrated the efficiency of our approach with a case study on immune response data. Our method presents an advantage in
term of pertinence of information and might help biologists and clinicians to interpret the function of gene sets by navigating in a simplified tree. As future work, we plan to implement a complete web application that will provide all services described in this paper, from the enrichment to tool to the visual exploration of the results. In this work, we used the GO as knowledge model of reference and aligned the annotation terms produced by the enrichment method to that model. Another future work is instead to align several knowledge models. Such approach should reduce the number of gene sets without associated GO terms.

REFERENCES


