Ontology-based gene set enrichment analysis using an efficient semantic similarity measure and functional clustering.

Sidahmed Benabderrahmane[1] and Mekami Hayet[2]

(1): King Abdul Aziz University,
Faculty of Computing and Information Technology.
Jeddah 21589, Saudi Arabia.
sidahmed.benabderrahmane@gmail.com,
(2): Djillali Liabes University,
Laboratory of communication and
multimedia architecture,
Bel Abbes 22000, Algeria.

Abstract. Gene set enrichment analysis allows to extract specific biological functions relative to a group of genes. To this aim, we propose here a novel approach for mining biological data, using the Gene Ontology (GO) as main source of genes annotation terms. Firstly, we will use our new semantic similarity measure (*IntelliGO*) in a clustering process, for grouping genes sharing similar biological functions described in GO. Secondly, the clustering results are evaluated using the F-score method and public genes reference sets. After that, an overlap analysis is presented as a method for exploiting the matching between clusters and reference sets. This method is then applied to a list of genes found dysregulated in cancer samples. In this case, the reference sets are replaced by gene expression profiles. Consequently, overlap analysis between these profiles and functional clusters obtained with the *IntelliGO*-based clustering, leads to characterize subsets of enriched biological functions of genes displaying consistent functions and similar expression profiles.

1 Introduction

In the last decade, DNA microarrays were used for measuring the expression levels of thousands of genes under various biological conditions [11, 8]. Thus, gene expression data analysis proceeds in two steps: Firstly, expression profiles are produced by grouping genes displaying similar expression levels under a given set of situations [13]. Secondly, a functional analysis, based on functional annotations, is applied on genes sharing the same expression profile, in order to identify their relevant biological functions [6, 16, 18, 19]. In fact, one important purpose of this functional analysis is to identify and characterize genes that can serve as diagnostic signatures or prognostic markers for different stages of a disease. One of the most interesting ontology in the biological domain is the Gene Ontology (GO), which is one of the most commonly used source of functional annotations

of genes [5, 1, 2].
This ontology of about 30,000 terms is organized as a controlled vocabulary describing the *biological process* (BP), *molecular function* (MF), and *cellular component* (CC) aspects of gene annotation, also called GO aspects [17]. The GO vocabulary is structured as a rooted Directed Acyclic Graph (rDAG) in which GO terms (concepts) are the nodes connected by different hierarchical relations (mostly *is_a* and *part_of* relations). The *is-a* relation describes the fact that a given child term is a specialization of a parent term, while the *part-of* relation denotes the fact that a child term is a component of a parent term. By definition, each rDAG has a unique root node, relationships between nodes are oriented, and there are no cycles, *i.e.* no path starts and ends at the same node. The GO Consortium regularly updates a GO Annotation (GOA) Database [2] in which appropriate GO terms are assigned to genes or gene products from public databases.
GO is widely used in several complex biological data mining problems. Authors in [20, 7] used GO for gene functional analysis in order to interpret DNA microarrays experiments, by exploiting the commonly accepted assumption that genes having similar expression profile should share similar biological functions. In such analysis, an enrichment study based on statistical P-value calculation are applied to genes sharing the same expression profile [10]. The results usually consist in sets of GO terms characterizing the biological function predominantly represented in a list of genes, thereby suggesting which function or process is affected when the behavior of this group of genes varies. However, the main limitation of these kinds of methods is that they consider the input list of genes in the enrichment analysis, as functionally homogeneous. Nonetheless in practice, genes present in the same expression profile could be involved in multiple biological processes. Thus the statistical tests for extracting specific GO terms could be biased. Moreover, the already proposed methods for gene functional enrichment analysis do not consider exclusively the three aspects of GO, that is not important for the biologists. To overcome this problem, we proposed here a new approach for analyzing gene expression data, by refining and creating subgroups of functionally homogeneous genes. The enrichment analysis could be then applied on each subgroup of genes, thus assuring the extraction of specific biological functions (GO terms) for those genes. The creation of subgroups of genes is performed using a clustering method based on our recently described semantic similarity measure called *IntelliGO* [4], that applies functional comparison between genes annotated by GO terms.
This paper is organized as follows. The next section, outlines the utilization of *IntelliGO* in a functional clustering approach and presents the evaluation results, using the F-score method and collections of reference sets. In second stage (Section III), we present and overlap analysis method that exploits the matching between functional clusters and reference sets. This method is then applied to a list of genes found dysregulated in cancer samples by replacing the reference sets by gene expression profiles. An enrichment analysis is then applied on overlapping genes, and leads to characterize subsets of enriched biological functions of

genes displaying consistent functions and similar expression profiles. Finally in the last section the relevance of the obtained results of the proposed algorithms are are discussed.

2 The IntelliGO-based gene functional classification

2.1 Presentation of the datasets

Gene functional clustering aims to regroup genes sharing common biological functions. We used four datasets for evaluating functional classification of genes, already presented in our past study[4]. In each dataset we prepared a collection of reference sets. Each reference set represents a group of genes grouped by an expert due to their shared biological functions. We selected a total of 13 KEGG pathways from the KEGG database [15] for the Biological Process aspect of GO for human (total of 280 genes) and yeast (total of 185 genes) species. For the Molecular Function aspect of GO we chose 10 Pfam clans from the Sanger Pfam database [12] for both species(100 genes for human and 118 for yeast species).

2.2 Calculating similarity matrices and clustering

For performing a gene functional clustering for a given list of genes, the first step is to compute a matrix representing the semantic similarity values between all genes in the input list. This similarity matrix will be then used as parameter of a clustering algorithm. In our case we used both hierarchical and fuzzy clustering. The first method allows to have a global overview of the distribution of genes on different clusters, while the second method allows to a gene to belong to multiple clusters at one time. In fact, one gene could be involved in multiple biological process simultaneously. These two algorithms are available in R-Bioconductor package[1].
Pairwise similarity matrices were calculated for all genes present in the four datasets using our recently proposed similarity measure (*IntelliGO*) [4]. This measure is represented in an innovative vector space model (VSM), and takes into account both information content of annotation terms and their positions in the ontology rDAG [4]. With *IntelliGO VSM*, each gene is represented as a vector $\boldsymbol{g}$ in a k-dimensional space where the basis vectors $\boldsymbol{e_i}$ correspond to the k annotation terms. To measure the semantic relationships between terms, we defined a term similarity product as:

$$\boldsymbol{e_i}.\boldsymbol{e_j} = \frac{2 * Depth(LCA)}{MinSPL(t_i, t_j) + 2 * Depth(LCA)}. \quad (1)$$

Moreover, we included in the *IntelliGO* VSM a novel weighting scheme in which a coefficient α_i is assigned to each $\boldsymbol{e_i}$ so that the gene representation becomes: $\boldsymbol{g} = \sum_i \alpha_i.\boldsymbol{e_i}$. The coefficients ($\alpha_i$) combine a weight $w(g, t_i)$ which depends

[1] www.bioconductor.org

on the evidence code tracking the annotation of gene g with a GO term t_i and on the *Inverse Annotation Frequency* ($IAF(t_i)$) which is an estimation of the information content IC of the term t_i. Thus, the similarity between $\boldsymbol{g_1}$ and $\boldsymbol{g_2}$ is given by the following generalized cosine formula:

$$IntelliGO(g_1, g_2) = \frac{\boldsymbol{g_1}.\boldsymbol{g_2}}{\sqrt{\boldsymbol{g_1}.\boldsymbol{g_1}}\sqrt{\boldsymbol{g_2}.\boldsymbol{g_2}}}, \quad (2)$$

with: $\boldsymbol{g_1}.\boldsymbol{g_2} = \sum_{i,j} \alpha_{1i}\alpha_{2j}\boldsymbol{e_i}.\boldsymbol{e_j}$.
Remark that *IntelliGO* is a *pair-wise* measure involving both *node-based* and *edge-based* similarities. The measure, the clustering algorithms and the used datasets are available at http://intelligo.loria.fr.

2.3 Evaluation of the clustering using the F-score method

When reference sets are available, the best method for optimizing the number of classes produced by unsupervised classification approaches is the F-score method [22]. This method relies on pairing each reference set with the best-matched cluster and provides a quantitative estimation of the pairing efficiency (precision and recall). We decided to extend the F-score method in order to further investigate the pairing between reference sets and clusters in a so-called overlap analysis. Our approach is outlined in the following algorithms. Algorithm 1 describes unsupervised clustering optimization with reference sets and global F-score measure. We applied fuzzy C-means clustering on the gene-gene pairwise

Algorithm 1 Clustering optimization with reference sets and F-score measure.

Require: Σ=$\{R_1, R_2, .., R_p\}$: a collection of reference sets, (n_1,n_2) such that n_1<p<n_2. The pairwise similarity matrix of all elements of Σ.
Ensure: The optimal number of generated clusters $\widehat{K}$, $Global\ \widehat{F-score}(\widehat{K})$.
1: **for** each K in [n_1,n_2] **do**
2: Generate K clusters Φ=$\{C_1, C_2, ..., C_K\}$, using all elements in $\bigcup R_i$
3: **for** each reference set $R_i \in \Sigma$ **do**
4: **for** each cluster $C_j \in \Phi$ **do**
5: Precision(R_i,C_j)=$|R_i \cap C_j|/|C_j|$
6: Recall(R_i,C_j)= $|R_i \cap C_j|/|R_i|$
7: $F-score(R_i, C_j)$=$\frac{2*Precision(R_i,C_j)*Recall(R_i,C_j)}{Precision(R_i,C_j)+Recall(R_i,C_j)}$
8: **end for**
9: $\widehat{F-score}(R_i) = Max_{\forall C_j \in \Phi}(F-score(R_i, C_j))$
10: **end for**
11: $Global\ F-score(K) = \frac{\sum_{i=1}^{p}(|R_i|*\widehat{F-score}(R_i))}{\sum_{i=1}^{p}|R_i|}$
12: **end for**
13: $Global\ \widehat{F-score}(\widehat{K}) = Max_{\forall K} Global\ F-score(K)$
14: **return** $\widehat{K}$, $Global\ \widehat{F-score}(\widehat{K})$.

similarity matrices calculated with *IntelliGO* for the four datasets. We used this

clustering algorithm since some genes can be involved in multiple biological processes or molecular functions. The same evaluation procedure was performed on a tool representing the state of the art for gene classification methods (DAVID: Database for Annotation, Visualization, and Integrated Discovery classification tool) [9, 14]. Each clustering result together with the corresponding collection of reference sets served as input to Algorithm 1 for determining global F-scores.
Concerning the *IntelliGO*-based fuzzy clustering of Datasets 1 and 2, we varied the number of generated clusters, K, between 11 and 17 in steps of 1 since these datasets are composed of 13 pathways for human and yeast species. For Datasets 3 and 4, the values of K were taken between 8 and 14 with a step of 1, since these two datasets are composed of 10 Pfam clans for both species. For each K, the global F-score(K) value was calculated. Concerning the DAVID functional classification of the same datasets, we varied the *Kappa similarity threshold* between 0.3 to 0.7 with a step of 0.1 in order to obtain different numbers of clusters, since DAVID does not allow the number of clusters to be specified *a priori*. As in the previous case, the K clusters were matched with the input reference sets, and the $Global\ F-score(K)$ value was calculated. The results are presented in Table 1.

Regarding the results obtained with Dataset 1 (13 human KEGG pathways) using our similarity measure, it can be seen that all global F-Score values are greater than 0.5, with a maximum value of 0.62 for $K = 14$. This means that the genes of the 13 human pathways considered in Dataset 1 are best grouped with our measure into 14 functional clusters. This result can reflect the fact that one pathway of the KEGG database encompasses two biological processes and/or that the clustering process has grouped together genes from various pathways sharing common BP annotations.
With DAVID (Table 1), the maximum global F-score (0.67) is reached when $Kappa = 0.3$, giving 10 functional clusters. At higher threshold, the number of genes excluded from the clustering increases, revealing one limit of the DAVID tool. Similar results are obtained with Datasets 2, 3 and 4 and are detailed in Table 1.
In summary these results indicate that *IntelliGO*-based clustering appears as a valuable alternative to DAVID classification tool. It is noteworthy that with DAVID classification tool all maximum values of global F-score are obtained for the minimal Kappa similarity threshold (0.3) which corresponds, according to DAVID, to the poorest quality of clustering. Moreover, the calculation of the global F-score is somewhat biased with DAVID as a certain number of genes are excluded from the classification results.

	A. IntelliGO		B. DAVID tool			
Dataset	K	Global F-score	Kappa Thr.	K	Global F-score	% excl.
1	11	0.59	**0.3**	**10**	**0.67**	20.7
(13 reference sets, total of genes =280)	12	0.61	0.4	11	0.63	31.4
	13	0.61	0.5	14	0.66	38.2
	14	**0.62**	0.6	11	0.41	75.9
	15	0.56	0.7	8	0.31	68.2
	16	0.55				
	17	0.54				
2	11	0.59	**0.3**	**9**	**0.68**	17.8
(13 reference sets, total of genes =185)	12	0.62	0.4	8	0.65	31.4
	13	0.64	0.5	9	0.55	43.2.
	14	**0.67**	0.6	6	0.39	56.8
	15	0.66	0.7	7	0.20	69.2
	16	0.62				
	17	0.62				
	8	0.70	**0.3**	**11**	**0.64**	27.0
3	9	0.64	0.4	11	0.51	52.0
(10 reference sets, total of genes =100)	10	0.68	0.5	8	0.31	66.0
	11	**0.75**	0.6	3	0.09	93.0
	12	0.66	0.7	2	0.01	96.0
	13	0.66				
	14	0.64				
	8	0.79	**0.3**	**10**	**0.70**	40.7
4	9	0.77	0.4	9	0.47	61.0
(10 reference sets, total of genes =118)	10	0.78	0.5	9	0.39	69.5
	11	**0.82**	0.6	5	0.28	84.7
	12	0.78	0.7	3	0.21	91.5
	13	0.78				
	14	0.71				

Table 1. Variation of the global F-score values when (A) varying the number of generated fuzzy clusters K with the fuzzy C-Means algorithm using *IntelliGO* similarity measure and (B) varying the Kappa threshold (Kappa thr.) with DAVID functional classification tool. In B the percentage of genes that are excluded from the classification is indicated (% excl.). Results are shown for the four datasets used in this study (total number of genes between parentheses). The optimal K value and the corresponding maximal global F-score value are in bold.

3 Overlap analysis between functional clusters and reference sets

3.1 Overlap analysis algorithm

In order to refine our comparison, we decided to look at the matching between the reference sets and the clusters obtained with the optimal K value. We used Algorithm 2 to extract the top-ranked cluster from each list of clusters assigned to each reference set. This algorithm explains how clusters (C) are assigned to reference sets (R) according to the F-score values, allowing the identification of best-matching pairs ($R \cap C$).

The intersection $R \cap C$ is expected to display a highly homogeneous content composed of genes known as members of a reference set and found most similar by clustering. Alternatively, the two set-theoretic differences $C \backslash R$ and $R \backslash C$ can be considered in order to discover missing information. In our study, we are interested by genes present in $R \cap C$. Indeed, we apply an enrichment analysis on genes present in such intersection, in order to extract specific functions.

3.2 Application to cancer expression data

In this section, we present an application of the *IntelliGO*-based clustering and overlap analysis approach using a list composed of 128 genes relating to human

Algorithm 2 Assignment of clusters to reference sets according to the $F-score$ values.

Require: Σ=$\{R_1, R_2, .., R_p\}$: a collection of reference sets, Φ_K=$\{C_1, C_2, ..., C_K\}$: a collection of clusters, $\forall(i,j) \mid 1 \leq i \leq p, 1 \leq j \leq K$, $F-score(R_i, C_j)$ (see Algorithm 1).

Ensure: A ranked list of clusters, ordered by decreasing $F-score$, assigned to each reference set.

1: **for** each reference set $R_i \in \Sigma$ **do**
2: $List_i \longleftarrow (C_j, F-score(R_i, C_j))$: A list of clusters C_j ordered by decreasing values of F-score(R_i,C_j)
3: **print** $List_i$.
4: **end for**

colorectal cancers. The idea here is to confront the *IntelliGO* functional clusters of the 128 genes, and to consider as reference sets the *fuzzy Differential Expression Profiles* (fuzzy DEP) obtained from the same list of genes [3]. Here, each DEP represents a group of genes having similar expression profile. We believe that overlap analysis may lead to discover hidden relationships between gene expression and biological function. Fuzzy DEPs are considered here as a collection of reference sets for overlap analysis. More precisely, 8 fuzzy DEPs containing genes with GO annotation are retained from our previous study [3]. The pair-wise similarity matrix was generated for the 128 genes, and the number of clusters, k, was optimized with the Algorithm 1 using the 8 fuzzy DEPs as reference sets (Σ=$\{DEP_i\}, i = 1..8$). The optimal number of cluster was obtained for $k = 3$ with and F-score value equals to 0.4.

After that, Algorithm 2 was used to extract lists of genes present in $C \cap DEP$, *i.e.* displaying both functional similarity (C) and present in one of the eight fuzzy DEPs (R). The enrichment analysis could be then applied on these signature genes, to discover among theme statistically significant GO terms displaying low P_Value. In our case, the P_value is calculated for genes present in $C \cap DEP$ versus a background list (here all human genes) displaying GO annotation in the NCBI repository file[2], using the *hyper geometric test* [10].

Preliminary results have shown that very specific biological functions with inferior P_Values ($\leq$10E-04) were extracted for genes present in $C \cap DEP$. For example, genes in $Cluster\_1 \cap PED3$ have "regulation of transcription DNA-dependent" and "NADH oxidation", as very specific functions (non exhaustive). Genes of $Cluster\_2 \cap PED2$ have the following functions: "cell differentiation", "multicellular organismal development", "insulin secretion". Genes of $Cluster\_3 \cap PED14$ have the "Water transport" as specific function. The "Transport" processes are very important in the physiology of the digestive system. This function was found for the $AQP8$ (Aquaporine 8) human gene, which is found in the literature under expressed in the tumoral tissues. This gene belongs to $PED14$ which regroups genes under expressed in cancer versus normal tissue [3]. This observation could be considered as a positive witness of our strategy. Other similar results were obtained for the remaining PED, are not reported here.

[2] ftp://ftp.ncbi.nih.gov/gene/DATA/gene2go.gz

4 Conclusion and perspectives

In this paper, we have presented a gene set enrichment analysis based on functional clustering with the *IntelliGO* semantic similarity measure. In a first step, we proposed an algorithm for evaluation the clustering approach using reference sets and the F-score method. Very encourageous results were obtained with *IntelliGO* when compared with a well known classification method (DAVID tool). Beyond clustering *per se*, we have presented an overlap analysis method which leads to a pairing of clusters and reference sets and may be used for set-difference analysis. Applied to a list of genes from a transcriptomic cancer study, our method leads to identify subsets of genes displaying consistent expression and functional profiles. Promising results have been obtained using a simple GO term enrichment procedure. More sophisticated tools such as GSEA [21] could be used to improve the biological interpretation of these subsets of genes.

References

1. Michael Ashburner, Catherine Ball, Judith Blake, David Botstein, Heather Butler, Michael Cherry, Allan Davis, Kara Dolinski, Selina Dwight, Janan Eppig, Midori Harris, David Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel Richardson, Martin Ringwald, Gerald Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.
2. Daniel Barrell, Emily Dimmer, Rachael P. Huntley, David Binns, Claire O'Donovan, and Rolf Apweiler. The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucl. Acids Res.*, 37(suppl1):D396–403, 2009.
3. Sidahmed Benabderrahmane, Marie-Dominique Devignes, Malika Smaïl-Tabbone, Amedeo Napoli, Olivier Poch, Ngoc-H. Nguyen, and Wolfgang Raffelsberger. Analyse de données transcriptomiques: Modélisation floue de profils dexpression différentielle et analyse fonctionnelle. In *INFORSID*, pages 413–428, 2009.
4. Sidahmed Benabderrahmane, Malika Smail-Tabbone, Olivier Poch, Amedeo Napoli, and Marie-Dominique Devignes. Intelligo: a new vector-based semantic similarity measure including annotation origin. *BMC Bioinformatics*, 11(1):588, 2010.
5. Olivier Bodenreider. Special issue: Biomedical ontology in action. *Applied Ontology*, 4(1):1–4, 2009.
6. J-F Boulicaut and O. Gandrillon. *Informatique pour l'analyse du transcriptome*. Hermes Lavoisier, Paris, 2004.
7. Markus Brameier and Carsten Wiuf. Co-clustering and visualization of gene expression data and gene ontology terms for saccharomyces cerevisiae using self-organizing maps. *J. of Biomedical Informatics*, 40(2):160–173, 2007.
8. Alvis Brazma, Jaak Vilo, and Edited Gianni Cesareni. Gene expression data analysis. *FEBS Letters*, 480:17–24, 2000.
9. Glynn Dennis, Brad Sherman, Douglas Hosack, Jun Yang, Wei Gao, H Lane, and Richard Lempicki. David: Database for annotation, visualization, and integrated discovery. *Genome Biology*, 4(9):R60, 2003. A previous version of this manuscript was made available before peer review at http://genomebiology.com/2003/4/5/P3.
10. Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics*, 10(1):48, 2009.

11. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–14868, December 1998.
12. Robert D. Finn, Jaina Mistry, Benjamin Schuster-Bockler, Sam Griffiths-Jones, Volker Hollich, Timo Lassmann, Simon Moxon, Mhairi Marshall, Ajay Khanna, Richard Durbin, Sean R. Eddy, Erik L. L. Sonnhammer, and Alex Bateman. Pfam: clans, web tools and services. *Nucl. Acids Res.*, 34.
13. Audrey Gasch and Michael Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3(11):research0059.1–research0059.22, 2002.
14. Da Huang, Brad Sherman, Qina Tan, Jack Collins, W Gregory Alvord, Jean Roayaei, Robert Stephens, Michael Baseler, H Clifford Lane, and Richard Lempicki. The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, 8(9):R183, 2007.
15. Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, 38(Database issue):D355–360, January 2010.
16. Purvesh Khatri and Sorin Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, 2005.
17. P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.
18. David Martin, Christine Brun, Elisabeth Remy, Pierre Mouren, Denis Thieffry, and Bernard Jacq. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biology*, 5(12), 2004.
19. Brad Sherman, Da Huang, Qina Tan, Yongjian Guo, Stephan Bour, David Liu, Robert Stephens, Michael Baseler, H Clifford Lane, and Richard Lempicki. David knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics*, 8(1):426, 2007.
20. Nora Speer, Christian Spieth, and Andreas Zell. A memetic co-clustering algorithm for gene expression profiles and biological annotation. 2004.
21. Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.
22. C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 1979.