# Benchmarking a new semantic similarity measure using fuzzy clustering and reference sets: Application to cancer expression data

Sidahmed Benabderrahmane, Marie-Dominique Devignes, Malika Smail-Tabbone, Olivier Poch, Amedeo Napoli, Wolfgang Raffelsberger, Dominique Guenot, Nguyen Hoan, Eric Guerin

# Benchmarking a new semantic similarity measure using fuzzy clustering and reference sets: Application to cancer expression data

Sidahmed Benabderrahmane*, Marie-Dominique Devignes*, Malika Smail-Tabbone*, Olivier Poch **, Amedeo Napoli*, Wolfgang Raffelsberger**, Dominique Guenot ***, N.Hoan Nguyen **, and Eric Guerin ***.

*LORIA (CNRS, INRIA, Nancy-Université), Équipe Orpailleur, Campus scientifique, 54506 Vandoeuvre-lès-Nancy Cedex, France. Mail: benabdsi@loria.fr
**LBGI, CNRS UMR7104, IGBMC, 1 rue Laurent Fries, 67404 Illkirch, France.
***INSERM U682, 3 avenue Molière, Strasbourg, France.

**Résumé.** Les algorithmes de classification (*Clustering*) reposent sur des mesures de similarité ou de distance qui dirigent le regroupement des objets similaires dans un même groupe et la séparation des objets différents entre des groupes distincts. Notre nouvelle mesure de similarité sémantique (IntelliGO), récemment décrite, qui s'applique à la comparaison fonctionnelle des gènes, est testée ici dans un processus de clustering. L'ensemble de test est composé des gènes contenus dans une collection de classes de référence (*Pathways KEGG*). La visualisation du clustering hiérarchique avec des cartes de densité (*heatmaps*) illustre les avantages de l'utilisation de la mesure IntelliGO, par rapport à trois autres mesures de similarité. Comme les gènes peuvent souvent appartenir à plus d'un cluster fonctionnel, la méthode C-means floue est également appliquée à l'ensemble des gènes de la collection. Le choix du nombre optimal de clusters et la performance du clustering sont évalués par la méthode F-score en utilisant les classes de référence. Une analyse de recouvrement entre clusters et classes de référence est proposée pour faciliter des analyses ultérieures. Enfin, notre méthode est appliquée à une liste de gènes dérégulés, concernant le cancer colorectal. Dans ce cas, les classes de référence sont les profils d'expression de ces gènes. L'analyse de recouvrement entre ces profils et les clusters fonctionnels obtenus avec la méthode C-means floue conduit à caractériser des sous-ensembles de gènes partageant à la fois des fonctions biologiques communes et un comportement transcriptionel identique.

## 1 Introduction

### 1.1 Transcriptomic data analysis

In recent years, DNA microarrays technologies have become an important tool in genomics, allowing the measure of the *expression level* of several thousands of genes in different

biological situations. Using these technologies and clustering approaches, *expression profiles* can be produced by grouping together genes displaying similar expression levels in a set of situations.
Usually a functional analysis is then applied to genes from the same expression profiles in order to associate the profiles with one or more common biological functions, derived from functional annotations. The main purpose of this processing, known as functional profiling, is to identify and characterize genes that can serve as diagnostic signatures or prognostic markers for different stages of cancer.
Among the most commonly used functional annotations of genes are the Gene Ontology terms. The Gene Ontology (GO) is one of the most important tool in bioinformatics, consisting of about 30,000 terms. It is organized as a controlled vocabulary, represented as a rooted Directed Acyclic Graph (rDAG) in which GO terms are the nodes connected by different hierarchical relations (mostly *is_a* and *part_of* relations). This rDAG is covering three orthogonal aspects or taxonomies, namely the *biological process* (BP), *molecular function* (MF), and *cellular component* (CC) aspects of gene annotation (Consortium, 2010). The process of annotating a gene with a given GO annotation is summarized by an evidence code (EC), which reflects the quality of this association (Rogers et Ben-Hur, 2009).
GO annotations are widely used in several complex data mining problems relating to bioinformatics domains. However, it is still a challenge for biologists and computer scientists to analyze and use such a huge amount of data, growing in an exponential way. Authors as (Khatri et Draghici, 2005; Speer et al., 2004; Huang et al., 2009), used gene functional analysis in order to interpret DNA microarrays experiments, using the assumption commonly admitted that genes having similar expression profile should share similar biological function(s). Functional similarity between genes or gene products relies on measuring the similarity between their GO annotation terms. Many GO similarity measures have been described so far (Pesquita et al., 2009), some of which have been used for functional clustering (Speer et al., 2005; Adryan et Schuh, 2004; Brameier et Wiuf, 2007).

## 1.2 Semantic Similarity Measures

The notion of similarity measure is usually applied to objects sharing common attributes or characteristics (Blanchard et al., 2008). In the biological domain, these objects are generally genes or gene products annotated with GO terms. As the GO terms are organized in a rDAG, it is then possible to exploit the relationships between terms and define semantic similarity measures. In (Pesquita et al., 2009; Benabderrahmane et al., 2010) is presented the state of the art of a variety of semantic similarity measures. At the level of the individual GO terms, two categories of measures are reviewed, namely the *edge-based* measures which rely on edge counting in the GO graph, and the *node-based* measures which exploit the information content (*IC*) of both terms of the comparison and of their closest common ancestor (Resnik, 1995). The four semantic similarity measures involved in this study have been described elsewhere (Benabderrahmane et al., 2010). Briefly, they correspond to (i) a *pairwise*, *node-based* approach for Lord measure (Lord et al., 2003) , (ii) a *pair-wise*, *edge-based* approach for Al-Mubaid measure (Nagar et Al-Mubaid, 2008a), (iii and iv) two *group-wise*, hybrid (both *node-based* and *edge-based*) approaches for SimGIC (Pesquita et al., 2008) and *IntelliGO* measures, the latter being our new original *vector-based* method (Benabderrahmane et al., 2010).

### 1.3 Functional Clustering

Clustering algorithms rely on a similarity or distance measure that directs the grouping of similar objects into the same cluster and the separation of distant objects between distinct clusters (Macqueen, 1967). Clustering algorithms have been used in several domains, with the purpose of data reduction, hypothesis testing and prediction (Theodoridis et Koutroumbas, 2006). There are a multitude of clustering algorithms, but all of them are based on the same basic steps : feature selection, choice of the similarity or distance measure, grouping criterion and techniques, validation and interpretation of the results (Theodoridis et Koutroumbas, 2006; Rousseeuw, 1987).
In biology, clustering is often required for grouping genes or gene products with similar functions. The so-called functional clustering relies on a variety of metrics applied to expression levels, GO annotations, etc (Eisen et al., 1998; Huang et al., 2009; Adryan et Schuh, 2004; Wang et al., 2007). Two major categories of clustering algorithms are used in bioinformatics. *Hierarchical clustering* algorithms are popular because the resulting dendrograms are easily interpreted visually (Eisen et al., 1998). *Al Mubaid et al.* used hierarchical clustering for validating their functional similarity measure, by calculating the silhouette index of clusters generated with genes belonging to yeast pathways (Nagar et Al-Mubaid, 2008b). One limitation of this category of algorithms is that they do not allow overlap between clusters.
The second category concerns *partitional clustering* algorithms like the k-means and fuzzy C-means (FCM) algorithms. Gash and Eisen used the FCM algorithm to identify overlaps that may exist between clusters relating to yeast gene expression data (Gasch et Eisen, 2002). In (Speer et al., 2005), the authors presented a functional clustering approach using the k-means method and the functional similarity measure presented in (Jiang et Conrath, 1997).
The two categories of clustering algorithms are used in this paper with the *IntelliGO* semantic similarity measure. Previous results had shown that this measure displays a robust discriminating power between predefined sets of genes (Benabderrahmane et al., 2010). However clustering results obtained with this measure were neither reported nor compared with other measures. In a first step we explore a dataset of genes representing a collection of reference sets (KEGG pathways, Kanehisa et al. (2010)). Using hierarchical clustering and heatmap visualization, we compare the results obtained with the *IntelliGO* measure and those obtained with three other similarity measures. Then we optimize the *IntelliGO*-based FCM clustering using the reference sets and the F-score method (van Rijsbergen, 1979). We also propose an approach called *overlap analysis* that aims to exploit the matching between clusters and reference sets. In a second step, we explore a list of genes selected from a transcriptomic cancer study. We confront *IntelliGO*-based clustering results with the *fuzzy Differential Expression Profiles* (fuzzy DEP) defined in (Benabderrahmane et al., 2009). Overlap analysis of the *IntelliGO*-based FCM clusters leads to the identification of consistent subsets of genes which are further characterized with respect to GO-term enrichment.

## 2 Experimental Design

We decided to evaluate the clustering results using a collection of reference sets composed of 13 human KEGG pathways (Kanehisa et al., 2010). In Table(1) are presented the pathways with the number of genes they contain. The similarity values were calculated by considering

only the BP aspect of GO, assuming that genes belonging to the same pathway are often referring to similar biological process. Let us consider $List1$ the list containing the 280 genes present in the 13 human pathways.

| Pathways (Hsa :) | 00040 | 00920 | 00140 | 00290 | 00563 | 00670 | 00232 | 03022 | 03020 | 04130 | 03450 | 03430 | 04950 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Genes Nb | 26 | 13 | 17 | 11 | 23 | 16 | 7 | 38 | 29 | 38 | 14 | 23 | 25 |

TAB. 1 – *Reference dataset composed of a list of 13 human KEGG pathways. The number of genes present in each pathway is displayed (Gene Nb).*

Another list of genes relating to colorectal cancer was used as an applicative example after the evaluation the *IntelliGO* clustering based method. This list of 128 genes that were found dysregulated in cancer samples, is named $List2$, and corresponds to the 222 genes studied in (Benabderrahmane et al., 2009) from which 94 genes were excluded because they lack GO annotation.

Pair-wise similarity matrices were calculated for $List1$ and $List2$ using the *IntelliGO* measure, and for $List1$ only the three other similarity measures described in section (1.2). These matrices serve as input for the clustering process. The *C++* programming language was used to implement the *Lord, Al-Mubaid* and *IntelliGO* [1] measure, due to its good memory management and calculation speed. The $Sim_{GIC}$ measure is available in the *csbi.go* package within R Bioconductor [2] (Ova). Hierarchical clustering, heatmap visualization and FCM clustering were performed using R Bioconductor. The F-score method and the strategy of overlap analysis were implemented using *C++* programming language.

# 3 Results

## 3.1 Comparison of heat maps obtained with four different functional similarity measure results

Four pairwise similarity matrices were generated from $List1$, using three semantic similarity measures, namely : $SIM_{Lord}$, $SIM_{Al-Mubaid}$, $SIM_{GIC}$, in addition to our measure $SIM_{IntelliGO}$. The heatmaps generated after hierarchical clustering are presented in Figure 1. Color scale ranges from dark red for very similar genes to dark blue for very dissimilar genes. The heatmap visualization obtained with the $SIM_{Lord}$ measure (Panel A) reveals a very fuzzy color distribution associated with a quite imperfect grouping of similar genes in clusters around the diagonal. This confirms our previous observation that the $SIM_{Lord}$ measure does not efficiently discriminate between genes belonging to two different pathways (Benabderrahmane et al., 2010). The situation is globally reversed with the $SIM_{Al-Mubaid}$ and $SIM_{GIC}$ measures (Panels B and C respectively). These two heatmaps present a limited number of small well delineated clusters around the diagonal, with some other clusters displaying weak (light blue color) intra-set similarity and very few if any cross-similarity between clusters. This also

1. http ://bioinfo.loria.fr/Members/benabdsi/intelligo_project/
2. www.bioconductor.org

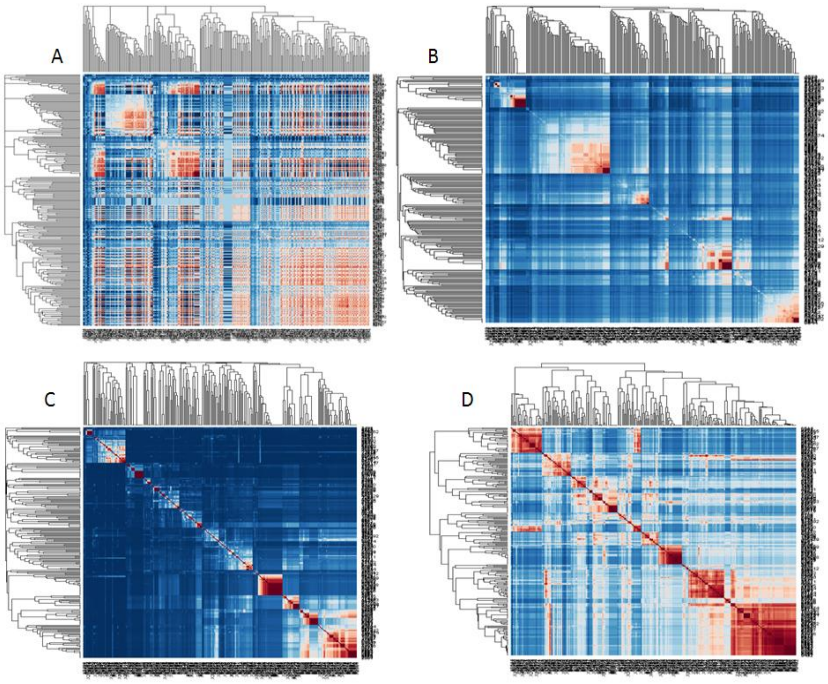


FIG. 1 – *Heatmaps generated after hierarchical clustering using the similarity matrices obtained with (A) : $SIM_{Lord}$, (B) : $SIM_{Al-Mubaid}$, (C) : $SIM_{GIC}$, and (D) : $SIM_{IntelliGO}$ semantic similarity measures. Genes belong to human pathways.*

confirms our previous findings concerning the variations in the discriminative power of these measures depending on the pathway (Benabderrahmane et al., 2010). Finally the heatmap obtained with the $SIM_{IntelliGO}$ measure (Panel D) appears well balanced in terms of color scale usage. More than 10 clusters of various sizes can be clearly identified around the diagonal, as well as cross-similarities between clusters.

Further observation of the heatmaps reveals that for two of them, the cells in the diagonal of the heatmap are seldom of dark red color which means that the self-similarity is rarely maximal with these two measures ($SIM_{Lord}$ and $SIM_{Al-Mubaid}$). On the contrary it can be checked that with the $SIM_{GIC}$ and $SIM_{IntelliGO}$ measures, the self-similarity is always maximal (equal to 1) as expected.

It thus appears that heatmaps constitute an interesting visual mean of estimating the performance of a similarity measure for clustering genes from a collection of reference sets. It should be noted here that very similar results were obtained when using another collection of reference sets, namely yeast instead of human pathways.

In this study we will continue working with the $SIM_{IntelliGO}$ measure which produced the most informative heatmap. However as mentioned above, hierarchical clustering is not the most appropriate method for functional clustering of genes because a given gene often belongs to more than one cluster. In fact, in the collection of genes studied here, several genes are in-

volved in multiple biological processes, and therefore belong to multiple pathways.

## 3.2 Fuzzy clustering approach using a collection of reference sets

The same list of genes ($List1$) was studied for fuzzy clustering using the $SIM_{IntelliGO}$ measure for producing the similarity matrix and the fuzzy C-means (FCM) algorithm. A matching analysis leading to the calculation of an average F-score has to be conducted in order to discover the optimal ($k$) number of fuzzy clusters (Cleuziou, 2010).

Knowing that our $List1$ corresponds to a collection of 13 pathways, we varied the number of generated clusters $k$ from 11 to 17. For each $k$ value, we calculate the precision and the recall of the reference sets in the best matching clusters leading to individual F-scores which are then averaged to give an average F-score reflecting the quality of the fuzzy clustering. The results are presented in Table (2).

| $k$ value | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|
| Average F-Score using IntelliGO | 0.59 | 0.61 | 0.61 | 0.62 | 0.56 | 0.55 | 0.54 |

TAB. 2 – *Variation of the F-Score when varying (k) number of FCM clusters with IntelliGO measure.*

It can be seen that all 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 $List1$ are grouped at best with our measure into $14$ functional clusters. This result can be easily explained by the fact that pathways of the KEGG database do present some overlaps due to genes being involved in multiple biological processes. Similar results have been observed when dealing with genes from 13 yeast pathways (not presented here).

## 3.3 Overlap analysis between cluster and reference sets

A possible exploitation of our fuzzy clustering experiment relies on a careful investigation of cluster content by domain experts. For this purpose, we defined and applied a generic overlap analysis method between cluster ($C_i$) and reference set ($R_j$). The strategy is illustrated in Figure 2.

Each of the $k$ clusters produced with the optimal $k$ value (see above) is compared with each reference set. A recall value is calculated as the ratio between the number of genes from the reference set present in the cluster and the total number of genes in the reference set. A well-matched pair is thus defined as the association of a cluster with a reference set displaying a recall value above a certain threshold. This threshold recall value is set to get at least one well-matched pair for each cluster and each reference set. In consequence more than one well-matched pairs can be produced for some clusters.

Well-matched pairs ($C$, $R$) constitute interesting datasets for further analyses. The intersection $C \cap R$ 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 studied in order to discover missing information. The former difference ($C \backslash R$) contains genes that are similar to genes from the reference set

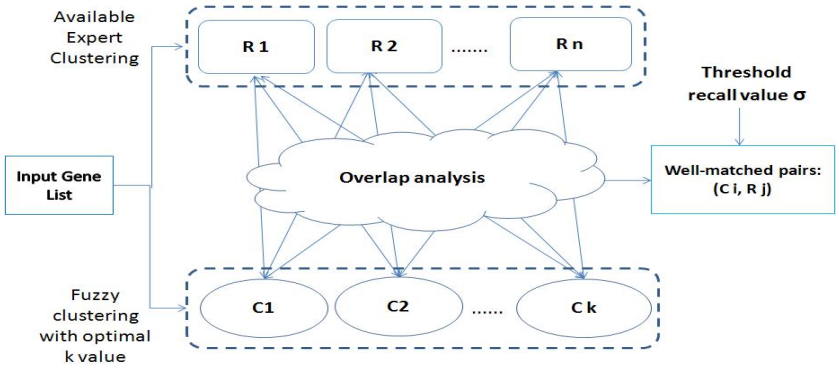


FIG. 2 – *Strategy for overlap analysis between clusters ($C_i$) and reference sets ($R_j$).*

but not counted among its members. This difference content can be presented to an expert in order to check whether some genes from $C\backslash R$ could be missing members of R. The latter difference $(R\backslash C)$ contains genes that are members of the reference set but do not get clustered with most other members on the basis of similar functional annotation. The annotation of genes from $R\backslash C$ can be scruted by an expert in order to check whether some terms could be missing. A GO-term enrichment study of cluster $C$ (Eden et al., 2009) can then be conducted in order to propose the most relevant GO terms for completing gene annotation in $R\backslash C$.

## 3.4 Application to a dataset relating to colorectal cancer

In this section, we present an application of the *IntelliGO*-based clustering and overlap analysis approach using $List2$ which is composed of 128 genes relating to colorectal cancers. The idea here is to confront functional clusters generated with *IntelliG0* measure and *fuzzy Differential Expression Profiles* (fuzzy DEP) obtained from the same list of genes. 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, eight fuzzy DEPs containing genes with GO annotation are retained from our previous study (Benabderrahmane et al., 2009).
The pair-wise similarity matrix was generated for the 128 genes of $List2$. Then, as a first step, the heatmap showing the resulting of hierarchical clustering was produced Figure (3). Despite of a high level of cross similarities in this dataset, several clusters can be distinguished around the diagonal of the heatmap. Fuzzy clustering was applied in a second step. The number of clusters, $k$, was optimized with the F-score method using the 8 fuzzy DEPs as reference sets. Table (3) shows the values obtained for $k$ varying from 2 to 14. The optimal value is $0.4$ for $k = 3$.

The three functional clusters produced for $k = 3$ were then studied by overlap analysis as described above, in order to extract lists of genes displaying both functional similarity and similar expression profile. The list of well-matched pairs maximizing the recall value between clusters and reference sets and *Fuzzy DEP* and the number of genes contained in their inter-

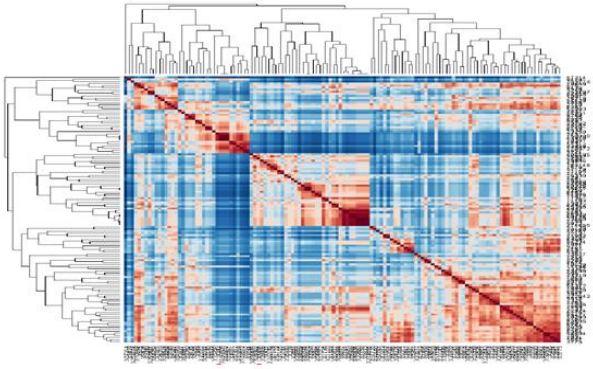

FIG. 3 – *Heatmap generated from IntelliGO pair-wise similarities of colorectal cancer genes (*$List2$*).*

| K generated clusters using IntelliGO | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Average F-Score | 0.39 | 0.4 | 0.37 | 0.32 | 0.26 | 0.27 | 0.29 | 0.29 | 0.29 | 0.28 | 0.26 | 0.25 | 0.16 |

TAB. 3 – *Variation of the F-Score for* $k$ *value varying from 2 to 14. FCM clustering was performed on* $List2$*, a list of* 128 *genes found dysregulated in colorectal cancer samples. The* 8 *fuzzy DEPs previously extracted from these* 128 *genes are taken as reference sets for F-score calculation.*

section are summarized in Table (4).

| | Fuzzy Differential Expression Profiles | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | P1 (34) | P2 (51) | P3 (32) | P13 (6) | P14 (5) | P15 (4) | P20 (31) | P22 (1) |
| Cluster1 (45) | | | 14 | 3 | | | | 1 |
| Cluster2 (64) | 15 | 28 | | | | 2 | 17 | |
| Cluster3 (19) | | | | | 3 | | | |

TAB. 4 – *Overlap analysis between the three functional clusters obtained with FCM and the 8 fuzzy DEPs. Between brackets is indicated the number of genes present in each set. For each well-matched pair, the number of genes of the intersection* $C \cap R$ *is reported in the corresponding cell. The threshold recall value is set here to 0.4 in order to get well-matched pairs for each reference set.*

Various methods exist for characterizing the biological relevance of signature genes obtained from high-throughput experimental results. One of them is the simple GO term enrichment analysis which allows to discover among all GO terms associated with all genes in a given cluster, statistically significant GO terms displaying low P_Value $< 10^{-4}$ or $10^{-5}$. The P_value is calculated for a given gene list versus a background list (here all human genes) displaying GO annotation in the NCBI repository file (GEN), using the *hyper geometric test* (Eden et al., 2009). Only the BP annotations are considered here. Table (5) presents the results obtained

with each well-matched pair. It can be seen that quite specific BP terms are assigned to each subset of genes delineated by the intersection $C \cap R$ of a functional cluster and an expression profile. In the case of Cluster_2 $\cap$ P2 and Cluster_2 $\cap$ P20, the same general GO term (*cell differentiation*) is found at the 1st position, but distinct GO terms appear at the 2nd and 3rd positions which correspond to biological processes which were mixed together in Cluster 2 but are now associated to two distinct expression profiles (P2 and P20). This example illustrates how our overlap analysis appears capable of extracting consistent subsets of genes with respect to biological function and transcriptional behavior.

| Cluster_1 $\cap$ P3 | | Cluster_2 $\cap$ P1 | | Cluster_2 $\cap$ P2 | | Cluster_2 $\cap$ P20 | | Cluster_3 $\cap$ P14 | |
|---|---|---|---|---|---|---|---|---|---|
| GO term | P_Value | GO term | P_Value | GO term | P_Value | GO term | P_Value | GO term | P_Value |
| regulation of transcription, DNA-dependent | 9.95E-04 | chromosome organization | 9.55E-05 | cell differentiation | 7.35E-05 | cell differentiation | 5.97E-05 | Water transport | 2.08E-05 |
| NADH oxidation | 2.98E-04 | strand break repair via homologous recombination | 1.1012E-04 | vascular endothelial growth factor receptor signaling pathway | 1.06E-04 | multicellular organismal development | 9.58E-05 | | |
| | | response to estrogen stimulus | 9.6584E-04 | angiogenesis | 5.83E-04 | insulin secretion | 1.42E-04 | | |

TAB. 5 – *GO term enrichment in the well-matched pairs. Only the top GO terms (with P_Value lower than* $10^{-3}$*) characterizing the genes present in the intersection* $C \cap R$ *are displayed here.*

# 4 Conclusion and Perspectives

In this paper, we have tested our recently described semantic similarity measure *IntelliGO* in various clustering approaches. A collection of reference gene sets composed of selected KEGG human pathways has been used. Heatmap vizualisation of hierarchical clustering has provided visual evidence that the IntelliGO measure is more advantageous than other measures for clustering genes with respect to semantic similarity. Fuzzy C-means clustering was successfully optimized with F-score values reaching a maximum value of $0.62$. A method for overlap analysis between clusters and reference sets has been described and implemented. It has been applied to a set of genes that are dysregulated in cancer using expression profiles as reference sets. It then allows to retrieve at the intersection of functional clusters and expression profiles, relevant subsets of genes that can be meaningfully characterized.

An important motivation of this work was to compare the performance of our *IntelliGO* similarity measure with other measures for clustering purposes. We have illustrated how the visualization with heatmaps of hierarchical clustering results may help to globally appreciate such performance. We intend to make our collections of reference sets of genes available on-line in a comparison tool complementary to the Collaborative Evaluation of GO-based Semantic Similarity Measures (CESSM) tool (Pesquita et al., 2008). Users would download the datasets, produce their own similarity matrices using the measure to be tested and submit these matrices on-line for hierarchical clustering and heatmap generation.

The fuzzy C-means clustering belongs to overlapping clustering methods that attract more and more attention, because of their application to many domains. Recently, some overlapping variants of the K-means algorithms have been proposed (Cleuziou, 2008, 2010), namely *Okm,*

*Okmed*, and *Wokmed*. These algorithms could now be tested with our *IntelliGO* measure and benchmarking collection of genes.
Optimizing fuzzy clustering remains challenging, especially in the absence of reference sets. In our application with cancer genes, we used expression profiles as reference sets. The influence of this choice on clustering results should be tested, an alternative solution being the clustering optimization without any reference sets (Ammor et al., 2008).
The overlap analysis method proposed here leads to a pairing of clusters and reference sets, which may be used for mismatch analysis. Indeed the genes present in a cluster but not in the corresponding reference set may be proposed as missing members of this reference set. Reciprocally, some genes from a reference set that are absent from the corresponding cluster may be enriched with features required for its grouping with other members of this cluster. Thus, the proposed overlap analysis may reveal a mean for discovering missing information.
Applied to a list of genes from a transcriptomic cancer study, our method also 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 DAVID (Huang et al., 2009) and GSEA (Subramanian et al., 2005) tools, could be used to improve the biological interpretation of these subsets of genes.

# Références

The csbl.go package. `http://csbi.ltdk.helsinki.fi/anduril/`.

The NCBI gene2go file. `ftp://ftp.ncbi.nih.gov/gene/DATA/gene2go.gz`.

Adryan, B. et R. Schuh (2004). Gene-Ontology-based clustering of gene expression data. *Bioinformatics 20*(16), 2851–2852.

Ammor, O., A. Lachkar, K. Slaoui, et N. Rais (2008). Optimal fuzzy clustering in overlapping clusters. *Int. Arab J. Inf. Technol. 5*(4), 402–408.

Benabderrahmane, S., M.-D. Devignes, M. Smaïl-Tabbone, A. Napoli, O. Poch, N.-H. Nguyen, et W. Raffelsberger (2009). Analyse de données transcriptomiques : Modélisation floue de profils dexpression différentielle et analyse fonctionnelle. In *INFORSID*, pp. 413–428.

Benabderrahmane, S., M. Smail-Tabbone, O. Poch, A. Napoli, et M.-D. Devignes (2010). Intelligo : a new vector-based semantic similarity measure including annotation origin. *BMC Bioinformatics 11*(1), 588.

Blanchard, E., M. Harzallah, et P. Kuntz (2008). A generic framework for comparing semantic similarities on a subsumption hierarchy. In *18th European Conference on Artificial Intelligence (ECAI)*, pp. 20–24.

Brameier, M. et C. Wiuf (2007). 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.

Cleuziou, G. (2008). An extended version of the k-means method for overlapping clustering. In *ICPR*, pp. 1–4. IEEE.

Cleuziou, G. (2010). Two variants of the okm for overlapping clustering. In F. Guillet, G. Ritschard, D. Zighed, et H. Briand (Eds.), *Advances in Knowledge Discovery and Management*,

Volume 292 of *Studies in Computational Intelligence*, pp. 149–166. Springer Berlin / Heidelberg. 10.1007/978-3-642-00580-0_9.

Consortium, T. G. O. (2010). The Gene Ontology in 2010 : extensions and refinements. *Nucl. Acids Res. 38*(suppl-1), D331–335.

Eden, E., R. Navon, I. Steinfeld, D. Lipson, et Z. Yakhini (2009). Gorilla : a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics 10*(1), 48.

Eisen, M. B., P. T. Spellman, P. O. Brown, et D. Botstein (1998). 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.

Gasch, A. et M. Eisen (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology 3*(11), research0059.1–research0059.22.

Huang, D. W. a. . W., B. T. Sherman, et R. A. Lempicki (2009). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols 4*(1), 44–57.

Jiang, J. J. et D. W. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)*, pp. 9008+.

Kanehisa, M., S. Goto, M. Furumichi, M. Tanabe, et M. Hirakawa (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research 38*(Database issue), D355–360.

Khatri, P. et S. Draghici (2005). Ontological analysis of gene expression data : current tools, limitations, and open problems. *Bioinformatics 21*(18), 3587–3595.

Lord, P. W., R. D. Stevens, A. Brass, et C. A. Goble (2003). Investigating semantic similarity measures across the Gene Ontology : the relationship between sequence and annotation. *Bioinformatics 19*(10), 1275–1283.

Macqueen, J. B. (1967). Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Nagar, A. et H. Al-Mubaid (2008a). A new path length measure based on go for gene similarity with evaluation using sgd pathways. In *Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems (CBMS 08)*, Washington, DC, USA, pp. 590–595. IEEE Computer Society.

Nagar, A. et H. Al-Mubaid (2008b). Using path length measure for gene clustering based on similarity of annotation terms. In *Computers and Communications, 2008. ISCC 2008. IEEE Symposium on*, pp. 637–642.

Pesquita, C., D. Faria, H. Bastos, A. Ferreira, A. Falcao, et F. Couto (2008). Metrics for go based protein semantic similarity : a systematic evaluation. *BMC Bioinformatics 9*(Suppl 5), S4.

Pesquita, C., D. Faria, A. O. Falcao, P. Lord, et F. M. Couto (2009). Semantic similarity in biomedical ontologies. *PLoS Comput Biol 5*(7), e1000443.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pp. 448–453.

Rogers, M. F. et A. Ben-Hur (2009). The use of gene ontology evidence codes in preventing classifier assessment bias. *Bioinformatics 25*(9), 1173–1177.

Rousseeuw, P. (1987). Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math. 20*(1), 53–65.

Speer, N., H. Frohlich, C. Spieth, et A. Zell (2005). Functional grouping of genes using spectral clustering and gene ontology. In *In Proceedings of the IEEE International Joint Conference on Neural Networks*, pp. 298–303. IEEE Press.

Speer, N., C. Spieth, et A. Zell (2004). A memetic co-clustering algorithm for gene expression profiles and biological annotation.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et J. P. Mesirov (2005). 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.

Theodoridis, S. et K. Koutroumbas (2006). *Pattern Recognition, Third Edition*. Orlando, FL, USA : Academic Press, Inc.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth.

Wang, J. Z., Z. Du, R. Payattakool, P. S. Yu, et C.-F. Chen (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics 23*(10), 1274–1281.

## Summary

Clustering algorithms rely on a similarity or distance measure that directs the grouping of similar objects into the same cluster and the separation of distant objects between distinct clusters. Our recently described semantic similarity measure (*IntelliGO*), that applies to functional comparison of genes, is tested here for the first time in clustering experiments. The dataset is composed of genes contained in a benchmarking collection of reference sets. Heatmap visualization of hierarchical clustering illustrates the advantages of using the *IntelliGO* measure over three other similarity measures. Because genes often belong to more than one cluster in functional clustering, fuzzy C-means clustering is also applied to the dataset. The choice of the optimal number of clusters and clustering performance are evaluated by the F-score method using the reference sets. Overlap analysis is proposed as a method for exploiting the matching between clusters and reference sets. Finally, our method is applied to a list of genes found dysregulated in cancer samples. In this case, the reference sets are provided by expression profiles. Overlap analysis between these profiles and functional clusters obtained with fuzzy C-means clustering leads to characterize subsets of genes displaying consistent function and expression profiles.