New approach for biological clustering based on Gene Ontology

Guillermo Leale*[1] , Diego H. Milone[2], Ariel Bayá[3], Pablo M. Granitto[3], Georgina Stegmayer[1,2]

[1]CIDISI, UTN-FRSF, CONICET, Lavaise 610, (3000) Santa Fe, Santa Fe, Argentina
[2]SINC, UNL-FICH, CONICET, (3000) Santa Fe, Santa Fe, Argentina
[3]CIFASIS, UPCAM (France)/UNR, CONICET, (2000) Rosario, Santa Fe, Argentina

Email: Guillermo Leale*- guileale@santafe-conicet.gov.ar;
*Corresponding author

Background

Clustering algorithms are applied on gene expression data to unravel information about biological processes which are hidden in the data. The knowledge and relations extracted from the data are later validated by the domain experts (in this case, biologists) [1]. As a common practice, gene clustering is performed using Euclidean distance or correlation on gene expression data [2]. This approach does not include other explicit biological information in the process. Recently, several semantic measures based on Gene Ontology (GO) have been developed to include direct biological knowledge into the calculus of distances between biological objects [3]. In this work, we propose the combination of both types of distances, which can be used within a clustering algorithm, leading to better results from a biological perspective. The proposal has been tested on two real datasets and validated with classical and biological performance measures.

Main proposal

This work presents the Gamma distance, defined as $d_\gamma(\mathbf{g}_i, \mathbf{g}_j) = \gamma d_{GO}(\mathbf{g}_i, \mathbf{g}_j) + (1-\gamma) d_e(\mathbf{g}_i, \mathbf{g}_j)$; $0 \leq \gamma \leq 1$; $\mathbf{g}_i, \mathbf{g}_j \in X$, where $d_e(\mathbf{g}_i, \mathbf{g}_j)$ is an expression distance such as Euclidean or Pearson, and $d_{GO}(\mathbf{g}_i, \mathbf{g}_j)$ is a semantic distance between the genes $\mathbf{g}_i$ and $\mathbf{g}_j$ from a set of genes X. Common choices for calculating the semantic distance are Resnik, Lin and Relevance [4–6]. The γ indicates the importance given to the semantic similarity between genes. In this work, we propose to take into account the number of common GO annotations between a pair of genes as a measure of their closeness or similarity from a biological point of view. A value of $\gamma = 0$ corresponds to a pure expression-based distance, and a $\gamma = 1$ corresponds to a pure semantic-based distance between genes. This distance is calculated pairwise among genes and it is incorporated in the training process of the Partitioning Around Medoids (PAM) clustering algorithm.

Materials and methods

The Gamma distance was evaluated on gene expression datasets from two species, *Arabidopsis thaliana* and the budding yeast *Saccharomyces cerevisiae*. Several number of clusters were selected according to the dataset sizes. Values of gamma ranging from 0 to 1 were used. Several validation measures were calculated to assess the results. Classical data mining external measures have been used, such as Compactness, Silhouette (calculated upon a clustering with the combined distance matrix and a clustering with the Euclidean distance only, the latter denoted as Sil_e) and Davies-Bouldin index. Biological measures have been used as well, such as Biological Homogeneity Index [7] and z-score [8]. The Global Measure for Linked Clustering (G) [1] was also calculated. This measure takes into account internal cohesion and separation from clusters as well as their biological homogeneity (measured in terms of the number of common metabolic pathways). We have also calculated the Biological Compactness (BC) of the resulting clusters as the average of the mean pairwise distances $d_{GO}(\mathbf{g}_i, \mathbf{g}_j)$ within each element of each cluster, for each value of gamma.

Results

Table 1 shows the results with different values of gamma for a sample of five genes from the *Saccharomyces cerevisiae* dataset. Four values of gamma have been used for this example: 0, 0.1, 0.4 and 0.75. The corresponding GO annotations are also shown. It can be seen that given a group of genes having annotations closely located within the GO as depicted in Figure 1, a distance with a low value of gamma groups those genes separately. On the contrary, distances considering increasing values of gamma group the genes together. Table 2 comprises the results of the validation measures applied to both datasets for $k = 100$. The measures were calculated upon considering the Relevance distance as the d_{GO} term and the Euclidean distance as the d_e term. On the columns, five different values of gamma have been evaluated for each index: 0, 0.25, 0.5, 0.75 and 1. The best value for each index is underlined in each table. In Table 3, it can be noticed that the compactness measure tends to have higher values for increasing values of γ. The silhouette shows a low increase for values of γ under 0.50, and a high increase for values of γ from 0.50 to 1, which indicates better quality clusters for increasing γ. The Sil_e has values under 0, which is related to an overall low quality of the obtained clusters from an expression-based point of view, and has a decreasing trend for values of γ closer to 1. Values for the DB index raise in a slow manner for increasing values of γ, reaching its maximum at $\gamma = 0.75$, and then decreasing. These low quality results were expected for the measures based only on expression as gamma increases and the expression distance is disregarded. However, it should be noticed that the differences in indexes values are quite low. BHI starts raising at a high rate for values of $\gamma \leq 0.50$. For $\gamma > 0.50$, the

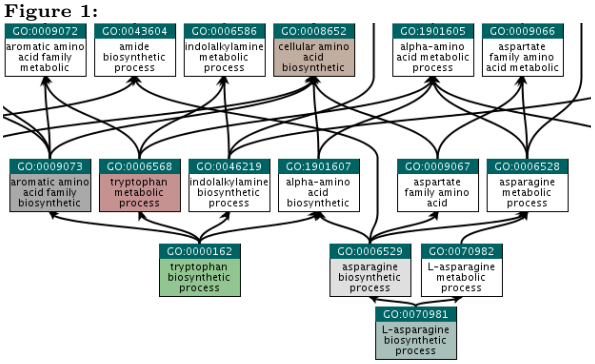
sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
G. Leale, D. H. Milone, A Bayá, P. Granitto & G. Stegmayer; "New approach for biological clustering based on Gene Ontology"
4to. Congreso Argentino de Bioinformática y Biología Computacional (4CAB2C), 2013.

Table 1:

Gene	Description	Cluster ID for each γ γ=0	γ=0.1	γ=0.4	γ=0.75	GO annotations
YGL026C	TRP5 TRYPTOPHAN BIOSYNTHESIS TRYPTOPHAN SYNTHASE	47	41	53	2	GO:0000162 GO:0006568 GO:0008152 GO:0008652 GO:0009073
YER090W	TRP2 TRYPTOPHAN BIOSYNTHESIS ANTHRANILATE SYNTHASE COMPONENT I	61	46	53	2	GO:0000162 GO:0008652 GO:0009058 GO:0009073
YPR145W	ASN1 ASPARAGINE BIOSYNTHESIS ASPARAGINE SYNTHETASE	62	60	75	2	GO:0000162 GO:0006568 GO:0008152 GO:0008652 GO:0009073
YLR146C	SPE4 SPERMINE BIOSYNTHESIS SPERMINE SYNTHASE	37	46	53	2	GO:0000162 GO:0008652 GO:0009058 GO:0009073
YPR069C	SPE3 POLYAMINE BIOSYNTHESIS PUTRESCINE AMINOPROPYLTRANSFERASE	54	69	53	2	GO:0006529 GO:0006541 GO:0008152 GO:0008652 GO:0070981
Validation measure	*BC*	0.39	0.34	0.22	0.18	

Example from the budding yeast *Saccharomyces cerevisiae*, showing a sample of genes and the obtained clusters for several values of gamma. In this example, the value of $k = 100$ was used, and the chosen combined distance was Euclidean as d_e and Relevance as d_{GO}. Clusters including more than one gene from the sample are underlined.

Figure 1:



Names and location of GO annotations for the genes of Table 1.

Table 2:

	$k = 100$				
$\gamma \rightarrow$	0.00	0.25	0.50	0.75	1.00
C	3.91	4.22	4.60	4.98	5.31
Sil	0.14	0.13	0.18	0.25	0.29
Sil_e	-0.02	-0.07	-0.16	-0.20	-0.21
DB	3.03	3.36	3.52	3.62	3.43
BHI	0.09	0.26	0.33	0.34	0.35
BC	0.39	0.24	0.20	0.18	0.18
z	11.60	11.60	13.50	10.00	11.80

Validation measures comparison for the *Saccharomyces cerevisiae* dataset. The best values for each validation measure are underlined.

Table 3:

	$k = 100$				
$\gamma \rightarrow$	0.00	0.25	0.50	0.75	1.00
C	2.89	3.12	3.77	4.43	6.19
Sil	0.14	0.10	0.15	0.24	0.35
Sil_e	0.51	0.47	0.42	0.40	0.34
DB	1.71	1.96	2.40	2.91	4.34
BHI	0.07	0.17	0.24	0.26	0.27
BC	0.59	0.40	0.30	0.26	0.23
G	5.06	4.85	3.43	4.50	4.67

Validation measures comparison for the *Arabidopsis thaliana* dataset. The best values for each validation measure are underlined.

values maintain the rising trend but at a lower rate. This shows the real improvement of the results from a biological point of view. The BC index decreases for increasing values of γ, achieving their best value at $\gamma \geq 0.75$. This means that the biological compactness of the cluster is effectively modified through the use of the Gamma distance. The z-score does not show a clear trend for increasing values of γ, reaching a maximum for $\gamma = 0.50$. In Table 3, it can be clearly seen that very similar trends are achieved with respect to Table 2 for all the measures. In this particular case, as the z-score is available for the budding yeast only, we used the G measure for Arabidopsis. Since G measures biological connectivity in terms of metabolic pathways, it is not directly improved by adjusting γ, obtaining the best score for $\gamma = 0.50$. In summary, it can be stated that the obtained results were consistent across all the validation measures, indicating better semantic quality for the clusters found with the new algorithm and increasing gamma values, in comparison to standard clustering ($\gamma = 0$).

Conclusions

We addressed the problem of incorporating biological information into clustering during training. In order to achieve this goal, we combined expression and semantic based measures into a new distance measure, which was evaluated on two real datasets. The obtained results showed that the γ parameter can be effectively used to control the biological quality of the partitions obtained by a clustering algorithm, by taking into account related biological annotations during training. This approach appears promising for the development of new biological clustering algorithms. As future work, we aim to extend this approach to other algorithms, in order to provide the ability to work with our combined measure. Furthermore, we intend to incorporate other biological validation measures into the training process in order to obtain clusters with better semantic quality from different biological points of view.

References

1. Stegmayer G, Milone DH, Kamenetzky L, López MG, Carrari F: **A biologically inspired validity measure for comparison of clustering methods over metabolic data sets**. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 2012, **9**(3):706–716.
2. de Souto MC, Costa IG, de Araujo DS, Ludermir TB, Schliep A: **Clustering cancer gene expression data: a comparative study**. *BMC bioinformatics* 2008, **9**:497.
3. Pesquita C, Faria D, Falcao AO, Lord P, Couto FM: **Semantic similarity in biomedical ontologies**. *PLoS computational biology* 2009, **5**(7):e1000443.
4. Resnik P: **Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language** 1999, **11**:95–130.
5. Lin D: **An information-theoretic definition of similarity.** In *ICML*, *Volume 98* 1998:296–304.
6. Schlicker A, Domingues FS, Rahnenführer J, Lengauer T: **A new measure for functional similarity of gene products based on Gene Ontology.** *BMC bioinformatics* 2006, **7**:302, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1559652&tool=pmcentrez&rendertype=abstract].
7. Datta S, Datta S: **Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes**. *BMC bioinformatics* 2006, **7**:397.
8. Gibbons F, Roth F: **Judging the quality of gene expression-based clustering methods using gene annotation**. *Genome research* 2002, (617):1574–1581, [http://genome.cshlp.org/content/12/10/1574.short].

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
G. Leale, D. H. Milone, A Bayá, P. Granitto & G. Stegmayer; "New approach for biological clustering based on Gene Ontology"
4to. Congreso Argentino de Bioinformática y Biología Computacional (4CAB2C), 2013.