A novel clustering approach for biological data using a new distance based on Gene Ontology

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

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

Abstract. When applying clustering algorithms on biological data the information about biological processes is not usually present in an explicit way, although this knowledge is later used by biologists to validate the clusters and the relations found among data. This work presents a new distance measure for biological data which combines expression and semantic information, in order to be used into a clustering algorithm. The distance is calculated pairwise among all pairs of genes and it is incorporated during the training process of the clustering algorithm. The approach was evaluated on two real datasets using several validation measures. The obtained results are consistent across all the measures, showing better semantic quality for clusters with the new algorithm in comparison to standard clustering.

1 Introduction

Clustering methods aim to find an appropriate division of a set of patterns into groups that present high similarity among them [1]. The clustering process is naturally associated to knowledge discovery, since its goal is to find interesting properties which are previously unknown on a given problem. In the biological domain, clustering is performed under the assumption that compounds involved in common biological processes behave similarly [2]. This is known as the *guilt-by-association* principle [3]. Therefore, if a compound with unknown function varies in a similar way than a known one, a strong assumption that both compounds are involved in the same regulatory process could be inferred [4].

The process of finding clusters in a dataset can be divided into three stages: 1) measuring the similarity of objects under analysis; 2) grouping objects according to this similarity; and 3) evaluating the quality of the clusters formed. Regarding the first stage, particularly in bioinformatics, most studies use the Pearson correlation coefficient and the Euclidean distance, mainly because of their ease of use and wide availability [5]. However, recent studies propose new metrics to better measure the similarity between points, for example through the use of perceptual organization, path-based dissimilarities and graphs [6,7,8]. The second stage has been widely studied in literature [9] and several clustering algorithms have been introduced and used in bioinformatics [5,10]. The last stage, validation of the results, has received less attention until recent years, although there is a growing interest in the problem [11].

In fact, after the application of one or several clustering algorithms on a biological database, validation of the clusters found is a common (and necessary) practice. This is often performed manually, based primarily on visual inspection of the patterns, according to a priori biological knowledge. This knowledge, however, is not present in the training patterns. The results obtained after examining each cluster may indicate functionally related patterns, leading to the generation of new biological hypotheses, and therefore, of new knowledge. For example, clusters grouping genes can provide evidence on the regulatory processes associated with them. This approach, which identifies process characteristics, results

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; "A novel clustering approach for biological data using a new distance based on Gene Ontology"
Argentine Symposium on Artificial Intelligence (ASAI 2013) - 42º JAIIO, 2013.

in greater biological information which has to be proved through the design of "wet" experiments to confirm the results [12].

As stated before, when applying clustering algorithms the information about biological processes is not usually present in an explicit way in the training patterns, although this knowledge is later used by biologists to validate the clusters and the relations found between data. However, data on regulatory processes involving known compounds are available and could be readily associated to each data pattern. For this reason, it could be useful to consider this (biological) domain information while clusters are being formed. With this information, new clustering techniques should be proposed, in order to use metrics that take into account biological information. This could be achieved through a new way of assessing the importance of the biological information within the groups during training, using a more efficient metric than the ones traditionally used for similarity calculation [8].

In fact, several new measures have been proposed recently, particularly on the basis of a controlled vocabulary such as the Gene Ontology (GO) [13], which considers the *semantics* of each biological element instead of just the traditional experiment-based measures [13,14]. GO provides concepts or *terms* in order to assign or *annotate* biological knowledge to a structured set of descriptions. Intuitively, the key to the similarity of two terms inside such a vocabulary could be interpreted as the extent to which they share information in common, indicated by the specificity of the term that subsumes them, or *information content* (IC) [15]. The more specific a subsumer term, the higher its IC and vice versa [16]. Based on IC as a means of comparing terms, several similarity measures have been developed for biological data. A detailed study can be found in [17]. Recent studies show that the use of a combined distance which includes both semantic measures based on ontologies and measures based on gene expression data can lead to stable, biologically relevant and still representative clusters [18,19]. However, these works do not consider nor study in deep the several alternatives that exist for a semantic measure. Furthermore, these proposals only focus on one single species and do not consider classical data mining validation measures to compare the results against biological based validation measures.

In this work, we propose a new measure for biological datasets that combines biological semantic-based and expression-based distances for clusters formation, in order to be used as an input to the clustering algorithm. This approach incorporates biological information into the training process, in order to improve the quality of the resulting clusters. We have evaluated the proposal using three semantic-based and two expression-based measures as an input to the PAM clustering method [20] on two real datasets of the species *Saccharomyces cerevisiae* [21] and *Arabidopsis thaliana* [22]. We have validated the results using a homogeneity measure that takes into account biological knowledge [23], obtaining meaningful clusters which outperform those obtained only by gene expression data when validated in a biological sense.

This work is organized as follows. Section 2 presents the distance measures used and the clustering algorithm in detail. Section 3 describes the input data, the preprocessing operations and the validation measures. Section 4 depicts the obtained results. Section 5 gives conclusions on the evaluation and introduces ideas for future work.

2 Distance measure for biological clustering: a new approach

2.1 Classical and semantic distance measures

Several distance measures are currently used for biological data. Firstly, let us consider measures which are calculated upon gene expression data only. Common choices are the classical Euclidean distance and the Pearson distance, defined as

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; "A novel clustering approach for biological data using a new distance based on Gene Ontology"
Argentine Symposium on Artificial Intelligence (ASAI 2013) - 42º JAIIO, 2013.

$d_P(\mathbf{g}_1, \mathbf{g}_2) = 1 - r(\mathbf{g}_1, \mathbf{g}_2)$, where $r(\mathbf{g}_1, \mathbf{g}_2)$ is the correlation coefficient between genes $\mathbf{g}_1$ and $\mathbf{g}_2$. Genes positively correlated are considered similar to each other, and not similar for lower and negative correlations [24]. Both measures are usually normalized in the range $[0, 1]$ [19][24]. Thus, the maximum distance between two genes is 1 and the minimum is 0.

Semantic similarities can be calculated upon objective knowledge representations, which can be found in ontologies such as GO. This ontology can be thought of as a structured, controlled vocabulary, which is used to associate biological knowledge to a pre-defined set of descriptions or *terms*. In particular, GO is a specific set of organized classifications or *taxonomies* for annotating genes to terms [13]. GO is presented as a Directed Acyclic Graph, in which each term can have one or more ancestors. Its structure consists of a root node with no practical importance, which has three children nodes: Molecular Function, Biological Process and Cellular Component. These nodes correspond to orthogonal categories which represent different aspects of gene function. GO terms are related to each other by "is-a" and "part-of" relationships, meaning a class-subclass and a part-whole relationship respectively [17]. A gene can be annotated to one or more terms.

The adoption of ontologies for annotation provides means to compare entities on aspects that otherwise would not be comparable. Semantic similarity measures can be defined as functions that, given two ontology terms or two sets of terms annotating two entities, return a numerical value reflecting the closeness in meaning between them [17]. Several distance measures have been developed following this approach [14]. We will consider three *information content* based measures, considering their relatively ease of use and wide applicability in gene data. These measures are defined as follows.

Let $p(t)$ be the probability of finding an instance of a term t in GO. This can be computed as the number of genes annotated to t or one of its descendants, divided by the total number of genes in the ontology. Following the standard argumentation of information theory [15], the IC of a term t can be quantified as the negative log likelihood of $p(t)$. Let $S(t_i, t_j)$ as the set of common ancestors of t_i and t_j. Based on these definitions, the Resnik similarity measure [16] is defined as

$$Sim_R(t_i, t_j) = -\log\left(\min_{t \in S(t_i,t_j)} p(t)\right) = \max_{t \in S(t_i,t_j)} I(t), \tag{1}$$

where $I(t)$ is the value of the IC.The term t that maximizes the IC is called the *minimum subsumer* (ms). It is the common ancestor between t_i and t_j with the higher information content, and therefore the closest one to both t_i and t_j. We will call for simplicity $\max_{t \in S(t_i,t_j)} I(t) = I_{ms_{(t_i,t_j)}}$. This measure has a minimum value of 0, and has not a maximum value.

Note that Resnik's measure is insensitive to the location of the comparing nodes and their minimum subsumer in the ontology [16]. Let us consider the following case. The nodes t_{i_1} and t_{j_1} are located near the top of the ontology, close to their minimum subsumer $ms_{(t_{i_1},t_{j_1})}$. Another pair of nodes t_{i_2} and t_{j_2} are located deep in the ontology, with their minimum subsumer $ms_{(t_{i_2},t_{j_2})} = ms_{(t_{i_1},t_{j_1})}$. Resnik's similarity evaluated on the two cases will provide the same value, but it is intuitively clear that in the first case there is a very general (and distantly located) concept that subsumes them both whereas in the second case the minimum subsumer is closer to both terms and therefore their meaning should be more similar. This issue is addressed by the Lin similarity measure

[25], which is defined as

$$Sim_L(t_i, t_j) = \max_{t \in S(t_i,t_j)} \left\{ \frac{2 \log p(t)}{\log p(t_i) + \log p(t_j)} \right\} = \frac{2 I_{ms_{(t_i,t_j)}}}{I(t_i) + I(t_j)}. \quad (2)$$

This measure compares the information of two terms with their minimum subsumer. If two specific terms are too far from their minimum subsumer, and hence deeply located in the ontology, their IC will be high and the IC associated with the minimum subsumer will be low, and therefore the similarity measure will give a small value. On the other hand, two terms that are close to their minimum subsumer will have similar IC and therefore the value of their Lin similarity measure will be closer to 1.

There is an issue which remains unsolved for this measure. Consider the nodes t_{i_1} and t_{j_1} located near the top of the ontology, close to their minimum subsumer $ms_{(t_{i_1},t_{j_1})}$. Another two nodes t_{i_2} and t_{j_2} are located deep in the ontology, with their minimum subsumer $ms_{(t_{i_2},t_{j_2})}$ being at the same relative location, that is, equally closer from t_{i_2} and t_{j_2} than $ms_{(t_{i_1},t_{j_1})}$ from t_{i_1} and t_{j_1}. In this case, it can be seen that despite their equal closeness, which might provide similar results for the Lin similarity, two terms that are very specific are more likely to be similar than two abstract terms. This fact can be easily intuited and therefore the location of the ms within the ontology should also be taken into account by a similarity measure. A new measure was proposed by Schlicker [26] to take into account that fact. It is called relevance similarity, and is defined as

$$\begin{aligned} Sim_r(t_i, t_j) &= \max_{t \in S(t_i,t_j)} \left\{ \frac{2 \log p(t)}{\log p(t_i) + \log p(t_j)} \left(1 - p(t)\right) \right\} \\ &= \frac{2 I_{ms_{(t_i,t_j)}}}{I(t_i) + I(t_j)} \left(1 - e^{-I_{ms_{(t_i,t_j)}}}\right). \end{aligned} \quad (3)$$

The relevance measure uses the level of detail of the minimum subsumer, that is, its location within the ontology $1 - p(t)$, as a weight to Lin's measure. Minimum subsumers which are very specific will provide higher similarity between the terms subsumed than those located near the root of the ontology. Both Lin and Relevance measures vary in the range $[0, 1]$.

It is important to consider that all of these similarity measures compare ontologic *terms*, not genes. Since we want to compare genes based on a distance measure, we will derive semantic distances. Firstly, we will calculate distances for each respective similarity: $d_R(\mathbf{g}_i, \mathbf{g}_j) = 1 - Sim_R(\mathbf{g}_i, \mathbf{g}_j)$, $d_L(\mathbf{g}_i, \mathbf{g}_j) = 1 - Sim_L(\mathbf{g}_i, \mathbf{g}_j)$ and $d_r(\mathbf{g}_i, \mathbf{g}_j) = 1 - Sim_r(\mathbf{g}_i, \mathbf{g}_j)$. To develop a gene measure from a term measure, let $GO_{\mathbf{g}_i}$ and $GO_{\mathbf{g}_j}$ be the sets of terms annotating the genes $\mathbf{g}_i$ and $\mathbf{g}_j$, respectively. Then a simple gene measure can be defined as the minimum distance (or the maximum similarity) between any pair of terms $t_{i_1} \in GO_{\mathbf{g}_i}$ and $t_{j_1} \in GO_{\mathbf{g}_j}$ [16]. We adopt the approach named *best-match average* (BMA) for its calculation [26,17], which is the average of the maximum similarities between $GO_{\mathbf{g}_i}$ and $GO_{\mathbf{g}_j}$. Let us consider a similarity matrix between the sets of terms $GO_{\mathbf{g}_i}$ and $GO_{\mathbf{g}_j}$. This matrix is not necessarily symmetric or square since the number of terms in each set can be different. Let also Sim_{GO} be any of the semantic GO-based similarities described above. From the following column and row *scores*

$$\overline{S_c} = \frac{1}{\left|GO_{\mathbf{g}_j}\right|} \sum_{t_j \in GO_{\mathbf{g}_j}} \max_{t_i \in GO_{\mathbf{g}_i}} Sim_{GO}(t_i, t_j), \quad (4)$$

$$\overline{S_r} = \frac{1}{\left|GO_{\mathbf{g}_i}\right|} \sum_{t_i \in GO_{\mathbf{g}_i}} \max_{t_j \in GO_{\mathbf{g}_j}} Sim_{GO}(t_i, t_j), \quad (5)$$

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; "A novel clustering approach for biological data using a new distance based on Gene Ontology"
Argentine Symposium on Artificial Intelligence (ASAI 2013) - 42º JAIIO, 2013.

BMA is calculated as the average

$$Sim_{BMA}(\mathbf{g}_i, \mathbf{g}_j) = \frac{\overline{S_c} + \overline{S_r}}{|GO_{\mathbf{g}_i}| + \left|GO_{\mathbf{g}_j}\right|}, \tag{6}$$

and following our distance-based approach, $d_{GO}(\mathbf{g}_i, \mathbf{g}_j) = 1 - Sim_{BMA}(\mathbf{g}_i, \mathbf{g}_j)$.

2.2 New distance measure for clustering of biological data

Based on the expression and semantic distances described above, a new distance that takes into account both will be defined here. Let $d_e(\mathbf{g}_i, \mathbf{g}_j)$ be one expression distance between genes such as Euclidean or Pearson. Then, given a set of genes X, we propose the *gamma distance*

$$d_\gamma(\mathbf{g}_i, \mathbf{g}_j) = \gamma d_{GO}(\mathbf{g}_i, \mathbf{g}_j) + (1-\gamma) d_e(\mathbf{g}_i, \mathbf{g}_j); \quad 0 \leq \gamma \leq 1; \quad \mathbf{g}_i, \mathbf{g}_j \in X, \tag{7}$$

where the value of γ indicates the importance given to the semantic information in the measure. A value of 0 corresponds to a pure expression-based distance, and a value of 1 corresponds to a pure semantic-based distance between genes.

The proposed distance can be used as an input for a clustering algorithm in order to obtain biologically relevant groups of genes. As our gamma distance works with both expression and semantic distances, it is not possible to directly use a classic centroid-based method as k-means [27]. Its standard form requires random initial points which are not necessarily objects in the original dataset and cannot be defined when, as in our case, data consists only of a set of distances [20]. Furthermore, it is unfeasible to calculate the semantic distance between an artificial centroid and a gene defined in terms of GO. Thus, the clustering algorithm proposed in our work is based on the PAM method [20]. PAM provides the ability to work only with a collection of dissimilarities or distances. The aim of this method, shown on Algorithm 1, is to find k representative objects. The representative object of a cluster is the object for which the average dissimilarity to all the objects of the cluster is minimal. This object is called the *medoid* of the cluster. The method consists of two phases, called *build* and *swap*. In the build phase, an initial clustering is obtained by the successive selection of k representative objects (lines 2 to 8). The first object is the one which minimizes the sum of dissimilarities to all other objects. Subsequently, at each iteration, another object is selected which decreases the sum (over all objects) of the dissimilarities to the most similar selected object, as much as possible. The process is continued until k objects have been found (line 7). The dissimilarities are calculated using the d_γ distance matrix provided as input. In the swap phase, it is attempted to improve the set of representative objects and therefore also improving the clustering. This is done by calculating the effect on swapping the representative objects with non-representative ones (lines 9 to 19). This procedure is iterated until no further reduction is possible.

3 Materials and performance measures

In this section we present the datasets with which the experiments have been performed. We also introduce the validation measures used to analyze the obtained results

3.1 Datasets

ARA dataset. This biological dataset comprises genes measured in *Arabidopsis thaliana* leaves. The original work was aimed to study the effects of cold temperatures on circadian-regulated genes in this plant [22]. Genes under light-dark cycles at two control temperatures (20°C and 4°C) and also involved in diurnal cycle and cold-stress responses were selected for the study. From a total of 1549 genes only those annotated to the *Biological Process* category of the Gene Ontology were considered. Genes annotated into GO but marked as "ND"

Algorithm 1: PAM algorithm with γ distance

Data:
X: set of genes
D_γ: distance matrix where each cell has
$d_\gamma(\mathbf{g}_i, \mathbf{g}_j) = \gamma d_{GO}(\mathbf{g}_i, \mathbf{g}_j) + (1-\gamma) d_e(\mathbf{g}_i, \mathbf{g}_j); \quad \mathbf{g}_i, \mathbf{g}_j \in X$
k: number of clusters
Result:
Ω: clusters
M: medoids

begin
  build **begin**
    Assign to the first medoid $\mathbf{m}_1$ the object which minimizes the sum of all d_γ distances.
    repeat
      Consider the previously selected medoid $\mathbf{m}_i$
      Select a previously nonselected object $\mathbf{g}_h$ so that
      $$\sum_j d_\gamma(\mathbf{g}_h, \mathbf{m}_i) - d_\gamma(\mathbf{g}_h, \mathbf{g}_j) < \sum_j d_\gamma(\mathbf{g}_l, \mathbf{m}_i) - d_\gamma(\mathbf{g}_l, \mathbf{g}_j) \quad \forall \mathbf{g}_l \neq \mathbf{g}_h \ .$$
    until *k medoids have been found*
  end
  swap **begin**
    repeat
      for *each* $\mathbf{m}_i \in M$ **do**
        for *each* $\mathbf{g}_h \notin M$ **do**
          Swap $\mathbf{m}_i$ and $\mathbf{g}_h$.
          Calculate the value of this clustering configuration as the
          $$\sum_j \min_i \{d_\gamma(\mathbf{g}_j, \mathbf{m}_i)\}$$
        Select the swappings that provide a lower value than the current one.
      Select the swapping that minimizes the value of the clustering configuration.
      Perform the swap.
    until *no lower value of the clustering configuration can be reached by swapping*
  end
  for *each* $\mathbf{g}_i \in X$ **do**
    Assign $\mathbf{g}_i$ to $\Omega_j / d^*_{\gamma_{ij}}(\mathbf{g}_i, \mathbf{m}_j) = \arg\min_{\forall j} d_\gamma(\mathbf{g}_i, \mathbf{m}_j)$
end

(no biological data available) were removed[4]. The final dataset used in our work has 1042 genes.

YEAST dataset. This biological dataset consists of gene expression data from the budding yeast *Saccharomyces cerevisiae*. Several characteristics such as diauxic shift, mitotic cell division cycle and sporulation, were collected in order to study cluster analysis of expression patterns. The activities of collecting and preprocessing the dataset are thoroughly explained in [21]. From an original dataset of 2467 genes, only those with complete attributes were considered. A filtering process to include only those genes annotated to the GO *Biological Process* category and to exclude those marked as "ND" was also applied. The final dataset has 587 genes.

3.2 Performance measures

In this subsection, the following notation is used: X is the dataset formed by $\mathbf{g}_i$ data samples; Ω is the set of samples that have been grouped in a cluster; and M is the set of $\mathbf{m}_j$ medoids of each cluster Ω_j. The following validation measures have been used in this study:

[4] For more information visit the website `http://www.geneontology.org/GO.evidence.shtml#nd` [Online; accessed Apr-2013]

Compactness. It is a validation measure which assesses cluster compactness or homogeneity. Intracluster variance or within-cluster scatter is the most popular representative [28]:

$$\overline{C}_j = \frac{1}{|\Omega_j|} \sum_{\forall \mathbf{g}_i \in \Omega_j} \|\mathbf{g}_i - \mathbf{m}_j\|_2, \tag{8}$$

where $|.|$ stands for set cardinality. As a global measure of compactness, the average of all clusters is calculated $\overline{C} = \frac{1}{k}\sum_j \overline{C}_j$, where k is the number of clusters.

Values of $\overline{C}$ close to 0 indicate more compact clusters.

Silhouette. It measures clustering quality comparing the "within" similarity against the "between" similarity [29]. It is defined as

$$Sil(\mathbf{g}_i) = \frac{b(\mathbf{g}_i) - a(\mathbf{g}_i)}{\max\left(a(\mathbf{g}_i), b(\mathbf{g}_i)\right)}, \tag{9}$$

where $a(\mathbf{g}_i)$ is the average dissimilarity of $\mathbf{g}_i$ to its assigned cluster Ω_a and $b(\mathbf{g}_i)$ is the minimum average distance of $\mathbf{g}_i$ to all the other clusters. The cluster for which $b(\mathbf{g}_i)$ is calculated, Ω_b, is called the *neighbor* of object $\mathbf{g}_i$. The silhouette indicates if $\mathbf{g}_i$ is better classified in Ω_a or Ω_b. Values of silhouette near 1 imply that the "within" dissimilarity $a(\mathbf{g}_i)$ is much smaller than the "between" dissimilarity $b(\mathbf{g}_i)$, and thus $\mathbf{g}_i$ is better clustered in Ω_a than in its neighbor Ω_b. On the other hand, values of silhouette near -1 imply that it would have been more natural to assign $\mathbf{g}_i$ to cluster Ω_b than to Ω_a. Values of silhouette near 0 mean that $\mathbf{g}_i$ lies equally far away from both, and it is considered an intermediate case. To provide a global measure, the overall average $\overline{Sil}(\Omega) = \frac{1}{k}\sum_k \frac{1}{|\Omega_j|} \sum_{\forall \mathbf{g}_i \in \Omega_j} Sil(\mathbf{g}_i)$ is calculated.

Davies-Bouldin index. It is a very popular metric for evaluating clustering algorithms [30],

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left\{ \frac{\overline{C}_i + \overline{C}_j}{\|\mathbf{m}_i - \mathbf{m}_j\|_2} \right\}. \tag{10}$$

This index is a function of the ratio of the sum of within-cluster scatter to between-cluster separation. This is an indication of clusters overlap, therefore DB close to 0 indicates that the clusters are compact and far from each other.

Biological homogeneity index. The Biological Homogeneity Index (BHI) measures the quality of the clusters on a biological basis. It can be thought of as an average proportion of gene pairs with matched GO terms[5] clustered together [23]. Let $F\left(GO_{\mathbf{g}_i}, GO_{\mathbf{g}_j}\right)$ be an indicator function that has the value 1 if at least one of the GO terms with which $\mathbf{g}_i$ and $\mathbf{g}_j$ are annotated match, and 0 in any other case. Then

$$BHI = \frac{1}{k}\sum_j \frac{1}{|\Omega_j|\left(|\Omega_j| - 1\right)} \sum_{\mathbf{g}_i \neq \mathbf{g}_j \in \Omega_j} F\left(GO_{\mathbf{g}_i}, GO_{\mathbf{g}_j}\right). \tag{11}$$

BHI can be interpreted as the proportion of common GO annotations within the obtained clusters and it varies in the range $[0, 1]$. A value of BHI close to 1 indicates that the clusters are homogeneous in terms of biological meaning.

Biological compactness. We propose a new measure based on the aforementioned compactness measure. In this case the average of the pairwise distances within each cluster is calculated, considering the semantic-based measures only.

[5] This measure is presented in a generalized way as a proportion of gene pairs with matched *functional classes*, which can be either GO terms, or GO ancestor terms at a higher, pre-defined level within the ontology. In our study, we consider only GO terms.

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; "A novel clustering approach for biological data using a new distance based on Gene Ontology"
Argentine Symposium on Artificial Intelligence (ASAI 2013) - 42º JAIIO, 2013.

Thus we define biological compactness for the cluster Ω_j as

$$BC_j = \frac{1}{|\Omega_j|} \sum_{\mathbf{g}_i \in \Omega_j} \sum_{\mathbf{g}_j \in \Omega_j} d_{GO}(\mathbf{g}_i, \mathbf{g}_j). \tag{12}$$

A low value of BC means a cluster with close elements in terms of semantic distances, which can be interpreted as a higher amount of GO-based information in common within the cluster. The overall biological compactness can be calculated as $\overline{BC} = \frac{1}{k} \sum_j BC_j$. Values of BC closer to 0 indicate that the clusters are more compact from a semantic point of view.

Global measure for linked clustering. For evaluating both coherence and biological significance of clusters found over biological datasets, we have used the G measure, which is a biologically-inspired validity measure for comparison of clustering methods over biological datasets [11]. It is defined as

$$G = \log(\check{H}) + \log(\overline{I}) + \log(\overline{P}), \tag{13}$$

where $\check{H}$ measures clustering homogeneity or the flatness of the distribution of patterns along clusters; $\overline{I}$ indicates if the data samples have been coherently grouped when having a sign-inverted value and $\overline{P}$ evaluates biological internal connectivity in terms of the number of common metabolic pathways among patterns grouped in a cluster. Lower values of G indicate better clusters.

z-score. It is a figure of merit based on the mutual information (MI) jointly held by the GO annotations and the cluster membership of all the genes clustered [24]. This measure is calculated as follows: the mutual information for the clustered data (MI_c) is calculated using the attribute database derived from GO; MI is calculated for a clustering obtained by randomly assigning genes to clusters of uniform size (MI_{rand}); finally z for MI_c and the distribution of MI_{rand} values (with mean MI_{rand} and standard deviation σ_{rand}) is calculated according to

$$z = \frac{MI_c - MI_{rand}}{\sigma_{rand}}. \tag{14}$$

The z-score can be interpreted as a standardized distance between the MI obtained by clustering and those MI obtained by random assignment of genes to clusters. The larger the z-score, the greater the distance to the random clustering. Thus higher scores indicate clustering results more significantly related to gene function. It is available online and implemented only to be used for the YEAST dataset[6].

4 Results and Discussion

The experiments were run with both datasets, YEAST and ARA, considering a range of values for k based on the relation of the total number of objects in each dataset. A very high value of k would assign each object to a cluster, whereas a very low value would cause the clusters to be excessively large. The following values for k were used: 50, 100 and 150 for the YEAST dataset; 100, 200 and 300 for the ARA dataset[7].

Table 1 comprises the results of the validation measures applied to the YEAST (A) and the ARA (B) dataset. The table is divided in three parts, one for each k. Each row shows the results for the validation measures presented on Section 3.2. Additionally, in order to have a validation index which measures the quality of the obtained clusters in terms of their expression data exclusively, a silhouette measure, denoted as Sil_e, was calculated considering the Euclidean

[6] ClusterJudge software. `http://llama.mshri.on.ca/cgi/ClusterJudge/cluster_judge.pl` (2011) [Online; accessed Apr-2013]

[7] For the implementation of the algorithm, we used the R language. The semantic distances were calculated using the *geneSim* function, included in the *GOSemSim* package [31]. The PAM algorithm was adapted from the *pam* function, included in the *cluster* package [32].

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; "A novel clustering approach for biological data using a new distance based on Gene Ontology"
Argentine Symposium on Artificial Intelligence (ASAI 2013) - 42º JAIIO, 2013.

Table 1. Validation measures comparison for the YEAST dataset (A) and the ARA dataset (B). The best values for each measure is underlined.

A. YEAST dataset.

$k = 50$					
$\gamma \rightarrow$	0.00	0.25	0.50	0.75	1.00
$\overline{C}$	4.75	4.88	5.30	5.65	6.12
Sil	0.14	0.12	0.16	0.22	0.26
Sil_e	-0.03	-0.07	-0.12	-0.15	-0.15
DB	2.26	2.78	2.94	2.83	2.74
BHI	0.09	0.22	0.29	0.32	0.31
BC_{Resnik}	0.76	0.69	0.64	0.63	0.63
BC_{Lin}	0.53	0.37	0.29	0.26	0.26
BC_{Rel}	0.58	0.42	0.33	0.31	0.31
z	18.60	15.10	18.80	20.90	18.80
$k = 100$					
$\gamma \rightarrow$	0.00	0.25	0.50	0.75	1.00
$\overline{C}$	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_{Resnik}	0.68	0.61	0.58	0.57	0.57
BC_{Lin}	0.39	0.24	0.20	0.18	0.18
BC_{Rel}	0.44	0.29	0.25	0.22	0.22
z	11.60	11.60	13.50	10.00	11.80
$k = 150$					
$\gamma \rightarrow$	0.00	0.25	0.50	0.75	1.00
$\overline{C}$	3.44	3.69	4.18	4.67	4.98
Sil	0.12	0.13	0.18	0.27	0.32
Sil_e	-0.02	-0.08	-0.19	-0.24	-0.25
DB	2.85	3.18	3.31	3.64	3.63
BHI	0.09	0.28	0.36	0.38	0.38
BC_{Resnik}	0.62	0.57	0.55	0.53	0.53
BC_{Lin}	0.28	0.17	0.14	0.12	0.12
BC_{Rel}	0.34	0.22	0.18	0.16	0.16
z	6.78	7.86	9.55	7.49	8.01

B. ARA dataset.

$k = 100$					
$\gamma \rightarrow$	0.00	0.25	0.50	0.75	1.00
$\overline{C}$	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_{Resnik}	0.84	0.78	0.72	0.69	0.68
BC_{Lin}	0.59	0.40	0.30	0.26	0.23
BC_{Rel}	0.66	0.47	0.37	0.33	0.29
G	5.06	4.85	3.43	4.50	4.67
$k = 200$					
$\gamma \rightarrow$	0.00	0.25	0.50	0.75	1.00
$\overline{C}$	2.39	2.71	3.23	3.80	5.65
Sil	0.14	0.12	0.16	0.26	0.41
Sil_e	0.50	0.46	0.39	0.35	0.26
DB	5.13	4.99	5.21	5.43	5.33
BHI	0.07	0.20	0.27	0.28	0.31
BC_{Resnik}	0.77	0.72	0.66	0.64	0.62
BC_{Lin}	0.45	0.28	0.20	0.17	0.14
BC_{Rel}	0.51	0.35	0.27	0.24	0.20
G	1.50	1.67	2.47	2.41	2.77
$k = 300$					
$\gamma \rightarrow$	0.00	0.25	0.50	0.75	1.00
$\overline{C}$	2.12	2.43	3.04	3.57	5.43
Sil	0.14	0.12	0.16	0.27	0.44
Sil_e	0.51	0.45	0.38	0.33	0.25
DB	5.04	5.39	5.17	5.27	5.29
BHI	0.07	0.21	0.28	0.31	0.34
BC_{Resnik}	0.73	0.67	0.61	0.59	0.57
BC_{Lin}	0.34	0.19	0.14	0.10	0.08
BC_{Rel}	0.41	0.27	0.20	0.16	0.14
G	1.32	1.39	1.33	1.63	2.54

distance only. The Biological Compactness, BC, was calculated based on the Resnik, Lin and Relevance distances, respectively, as semantic-based measures. The other 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 the table. In Table 1.A, for the YEAST dataset, it can be noticed that for $k = 50$, 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 $\gamma = 0.50$, and then decreasing at a very low rate. 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 values maintain the raising 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, behaving in a similar way for the three semantic measures (Resnik, Lin and Relevance), although at a lower rate for the first measure, achieving their

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; "A novel clustering approach for biological data using a new distance based on Gene Ontology"
Argentine Symposium on Artificial Intelligence (ASAI 2013) - 42º JAIIO, 2013.

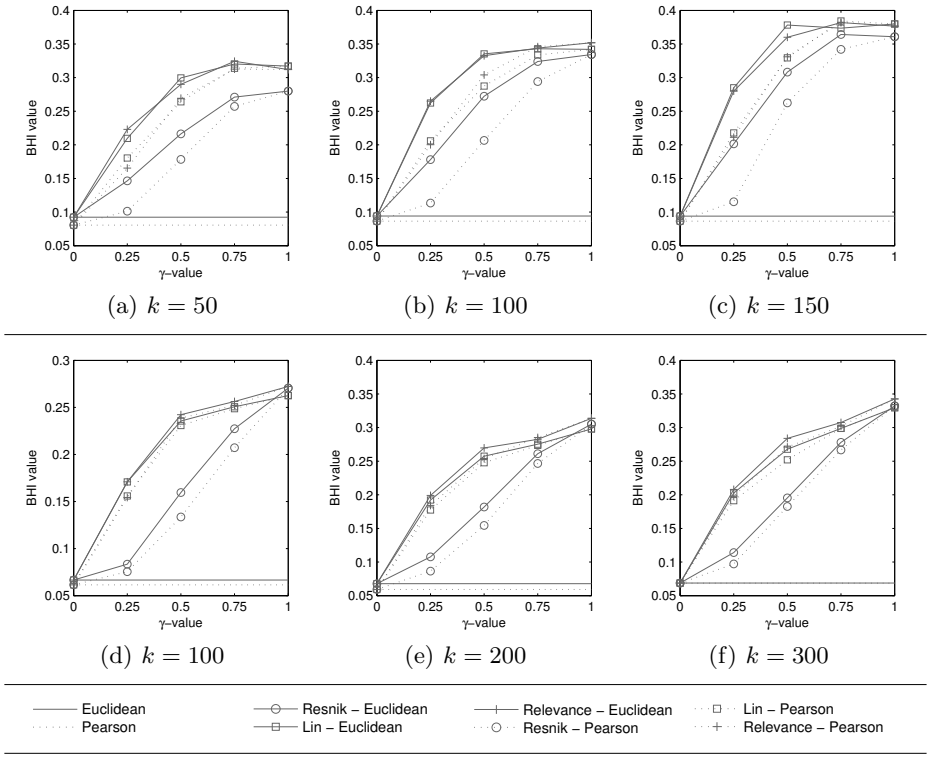

Fig. 1. BHI values for YEAST Dataset (a, b and c) with $k = 50, 100, 150$, and for ARA Dataset (d, e and f) with $k = 100, 200, 300$.

best value at $\gamma = 1$. 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 $\gamma = 0.75$, and for $\gamma = 0.50$ as k increases to 100 and 150. It can be noticed that for the remaining values of k, the validation measures behave in a similar way. In Table 1.B, the results of all the validation measures for the ARA dataset are shown. It can be clearly seen that very similar trends are achieved with respect to Table 1.A 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 the G index measures biological connectivity in terms of metabolic pathways, it is not directly improved by adjusting γ, except for $k = 50$, when it obtains the best score for $\gamma = 0.50$

Figure 1 show the values for BHI calculated with all the combinations of the semantic and expression based measures covered in this work, for the YEAST and ARA dataset respectively. Three graphics are shown for each dataset, corresponding to the different values of k. The five values of γ used for the evaluation in Table 1 have been considered for all graphics. The Euclidean and Pearson measures, which correspond to $\gamma = 0$ have also been included, in order to compare their performance against the gamma distance.

In Figure 1.(a), BHI values calculated for the YEAST dataset show an increasing trend for higher values of γ for all the measures, which indicates better

biological quality when considering a higher amount of GO based information as γ moves closer to 1 ($p < 0.001$). When used as d_{GO}, a similar trend is shown between Lin and Relevance measures. The Resnik measure increases at a lower rate than the former ones. Gamma distances using the Pearson distance as d_e behave in a similar way when compared with the measures using the Euclidean distance, but with consistently lower values. Figures 1.(b) and 1.(c) show a very similar behavior, with higher increasing rates and reaching higher values for values of $k = 100$ and $k = 150$. In Figure 1.(d), the values calculated for the ARA dataset show an analogous behavior to the YEAST dataset in terms of similar trends between Lin and Relevance measures when used as d_{GO}, and increasing values at a lower rate for the Resnik measure. The latter measure also shows a low increase for $\gamma = 0.25$ and then raises at a higher rate for $\gamma > 0.25$. Figures 1.(e) and 1.(f) show similar results, reaching maximum values for $\gamma = 1$.

5 Conclusions and Future Work

In this work we addressed the issue of incorporating biological information into clustering during training. We combined expression and semantic based measures into a new distance measure, which was evaluated on two real datasets. The validation measures compared in this work have shown better semantic quality of the clusters as the biological information was given more importance, consistently with lower cluster quality for validation measures which are calculated upon expression data only. However, this decrease in classical clustering measures was not so significant. The obtained results showed that the gamma parameter can be effectively used to control the biological quality of the partitions obtained by a clustering algorithm, by taking into account related biological information during training. Our approach appears promising for the development of new biological clustering algorithms.

As future work, we will extend this approach to other algorithms, such as k-means, 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. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. Wiley-interscience (2012)
2. Lacroix, V., Cottret, L., Thébault, P., Sagot, M.F.: An introduction to metabolic networks and their structural analysis. Computational Biology and Bioinformatics, IEEE/ACM Transactions on **5**(4) (2008) 594–617
3. Wolfe, C.J., Kohane, I.S., Butte, A.J.: Systematic survey reveals general applicability of. BMC bioinformatics **6**(1) (2005) 227
4. Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F.M., Bassel, G.W., Tanimoto, M., Chow, A., Steinhauser, D., Persson, S., Provart, N.J.: Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant, cell & environment **32**(12) (2009) 1633–1651
5. de Souto, M.C., Costa, I.G., de Araujo, D.S., Ludermir, T.B., Schliep, A.: Clustering cancer gene expression data: a comparative study. BMC bioinformatics **9**(1) (2008) 497
6. Fischer, B., Buhmann, J.M.: Path-based clustering for grouping of smooth curves and texture segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on **25**(4) (2003) 513–518
7. Nguyen, U.T., Park, L.A., Wang, L., Ramamohanarao, K.: A novel path-based clustering algorithm using multi-dimensional scaling. In: AI 2009: Advances in Artificial Intelligence. Springer (2009) 280–290
8. Bayá, A.E., Granitto, P.M.: Clustering gene expression data with a penalized graph-based metric. BMC bioinformatics **12**(1) (2011) 2
9. Xu, R., Wunsch, D.: Clustering. Volume 10. Wiley-IEEE Press (2008)
10. Milone, D., Stegmayer, G., Kamenetzky, L., López, M., Lee, J., Giovannoni, J., Carrari, F.: * omesom: a software for clustering and visualization of transcrip-

tional and metabolite data mined from interspecific crosses of crop plants. BMC bioinformatics **11**(1) (2010) 438
11. Stegmayer, G., Milone, D.H., Kamenetzky, L., López, M.G., Carrari, F.: A biologically inspired validity measure for comparison of clustering methods over metabolic data sets. Computational Biology and Bioinformatics, IEEE/ACM Transactions on **9**(3) (2012) 706–716
12. Tohge, T., Fernie, A.R.: Combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function. Nature protocols **5**(6) (2010) 1210–1227
13. Consortium, G.O.: The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research **32**(Database issue) (January 2004) 258D–261
14. Lord, P.W., Stevens, R.D., Brass, A., Goble, C.A.C.: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics **19**(10) (July 2003) 1275–1283
15. Sheldon, R., et al.: A First Course In Probability, 6/E. Pearson Education India (2002)
16. Resnik, P.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. **11** (1999) 95–130
17. Pesquita, C., Faria, D., Falcão, A.O., Lord, P., Couto, F.M.: Semantic similarity in biomedical ontologies. PLoS computational biology **5**(7) (July 2009) e1000443
18. Kustra, R., Zagdański, A.: Data-fusion in clustering microarray data: balancing discovery and interpretability. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM **7**(1) (2010) 50–63
19. Dotan-Cohen, D., Kasif, S., Melkman, A.a.: Seeing the forest for the trees: using the Gene Ontology to restructure hierarchical clustering. Bioinformatics (Oxford, England) **25**(14) (July 2009) 1789–95
20. Rousseeuw, L., Kaufman, L.: Clustering by means of medoids. In Dodge, Y., ed.: Statistical data analysis based on the L1-norm and related methods. North Holland/Elsevier (1987) 405–416
21. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences **95**(25) (1998) 14863–14868
22. Espinoza, C., Degenkolbe, T., Caldana, C., Zuther, E., Leisse, A., Willmitzer, L., Hincha, D.K., Hannah, M.a.: Interaction with diurnal and circadian regulation results in dynamic metabolic and transcriptional changes during cold acclimation in Arabidopsis. PloS one **5**(11) (January 2010) e14101
23. Datta, S., Datta, S.: Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. BMC bioinformatics **7**(1) (2006) 397
24. Gibbons, F., Roth, F.: Judging the quality of gene expression-based clustering methods using gene annotation. Genome research (617) (2002) 1574–1581
25. Lin, D.: An information-theoretic definition of similarity. In: ICML. Volume 98. (1998) 296–304
26. Schlicker, A., Domingues, F.S., Rahnenführer, J., Lengauer, T.: A new measure for functional similarity of gene products based on Gene Ontology. BMC bioinformatics **7** (January 2006) 302
27. Jain, A.K.: Data clustering: 50 years beyond K-means. Pattern Recognition Letters **31**(8) (June 2010) 651–666
28. Handl, J., Knowles, J., Kell, D.B.: Computational cluster validation in post-genomic data analysis. Bioinformatics **21**(15) (2005) 3201–3212
29. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20** (November 1987) 53–65
30. Davies, D.L., Bouldin, D.W.: A cluster separation measure. Pattern Analysis and Machine Intelligence, IEEE Transactions on (2) (1979) 224–227
31. Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., Wang, S.: Gosemsim: an R package for measuring semantic similarity among go terms and gene products. Bioinformatics **26**(7) (2010) 976–978
32. Reynolds, A., Richards, G., De La Iglesia, B., Rayward-Smith, V.: Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. Journal of Mathematical Modelling and Algorithms **5**(4) (2006) 475–504