

# A biologically-inspired validity measure for comparison of clustering methods over metabolic datasets

Georgina Stegmayer, Diego H. Milone, Laura Kamenetzky,  
Mariana G. López and Fernando Carrari

January 3, 2012

## Abstract

In the biological domain, clustering is based on the assumption that genes or metabolites involved in a common biological process are co-expressed / co-accumulated under the control of the same regulatory network. Thus, a detailed inspection of the grouped patterns to verify their memberships to well-known metabolic pathways could be very useful for the evaluation of clusters from a biological perspective. The aim of this work is to propose a novel approach for the comparison of clustering methods over metabolic datasets, including prior biological knowledge about the relation among elements that constitute the clusters. A way of measuring the biological significance of clustering solutions is proposed. This is addressed from the perspective of the usefulness of the clusters to identify those patterns that change in coordination and belong to common pathways of metabolic regulation. The measure summarizes in a compact way the objective analysis of clustering methods, which respects coherence and clusters distribution. It also evaluates the biological internal connections of such clusters considering common pathways. The proposed measure was tested in two biological databases using three clustering methods.

## 1 Introduction

A systems biology approach needs the integration and mining of large biological datasets to discover hidden relationships in such data. Starting from the analysis of post-genomic data, the discovery of novel biological knowledge relies mainly on the use of unsupervised data mining methods, in particular clustering techniques. In fact, much of the recent research in bioinformatics has been focused on the adaptation and application of classical clustering algorithms, such as hierarchical clustering and  $k$ -means [1, 2, 3], as well as on more recent approaches based on computational intelligence models [4, 5, 6, 7].

To avoid inconsistencies in the results, any clustering solution should be validated. However, after the application of an unsupervised mining technique, it is rather difficult

to validate the results obtained. A set of objective measures can be used to quantify the quality of the clusters obtained by the different available methods [8]. Nevertheless, it is very difficult to clearly indicate one as providing interesting clusters to be analyzed by biologists in order to discover new relationships among data. It is common practice to validate the returned groupings by a clustering algorithm through manual analysis and visual inspection, according to *a-priori* biological knowledge.

In the biological domain, clustering is implemented under the guilt-by-association principle [9], that is to say, the assumption that genes involved in a biological process are co-expressed ( behave similarly) under the control of the same regulatory network [10]. If an unknown gene is co-expressed with well-known genes in a biological process, it is assumed that this unknown gene might be involved in the same metabolic pathway. Similar reasoning can be applied to metabolites. Thus, from this perspective, it could be useful to perform a detailed inspection of the patterns inside a cluster to determine their membership to well-known metabolic pathways. Those clusters that group metabolites and transcripts provide evidence about the metabolic pathways associated with certain transcripts. That is to say, in the same analysis, the genes of interest, as well as their effect on the metabolites, are determined. This pathway-based approach to identify metabolic traits results in more biological information (hypothesis) that has to be tested through the design of biological experiments to confirm the results [11].

Looking at the analysis made by biologists when they evaluate the elements that are part of a cluster, coherent groupings are verified, as well as their belonging to well-known metabolic pathways. Those elements are important to qualify a cluster. Therefore, this work presents a new approach for evaluating both coherence and biological significance of the clusters found by clustering methods over a biological dataset. A new measure, which assesses, in a compact way, the kind of cluster analysis often made by biologists, is here proposed.

Several computational cluster validation techniques are available in the general data mining literature, but they have not been widely used in bioinformatics [12]. In order to choose an adequate clustering technique for a biological dataset, it is common practice to compare several methods through objective clustering measurements, which indicate the quality of the clusters found. A good clustering solution should tend to perform reasonably well under multiple measures. Also, functional coherence of clusters can be used to identify biologically meaningful clusters, compare clustering algorithms and identify biological pathways associated with the biological process under investigation [13].

In fact, there are several recent proposals in the literature for clustering validation obtained from biological datasets [14]. The work reported in [15], for example, proposes a statistical method for assessing a clustering solution according to prior biological knowledge. The method is based on projecting vectors of biological attributes of the clustered elements, such that the ratio of between-groups and within-group variance estimators is maximum. The work reported in [13] presents a framework for integrating knowledge-based functional categories into the cluster analysis of gene expression data. Given a hierarchical clustering of genes based on their expression profiles and a set of functional categories (for example, gene ontologies), a score for a gene is calculated by correlating the clustering structure with functional categories of interest. The proposal in [16] consists of two validation measures: one measuring the statistical sta-

bility of the clusters produced and the other one representing their biological functional consistency.

The incorporation of well-known gene functions into a new distance metric to form the clusters with the aim of reducing the distance between genes if both share a common function is proposed in [17]. This is based on the fact that co-expressed genes are likely to share the same biological function. Similarly, a very interesting figure of merit (z-score), based on the information jointly held by the functional annotation and cluster membership of genes, is proposed in [18]. This score is based on the premise that a good clustering algorithm tends to bring genes of similar function together, where the function is known. The clustering results are evaluated by examining the relationship between the clusters produced and the known attributes (annotations) of the genes in those clusters.

All of these proposals, however, mostly concentrate on enriching or evaluating clusters according to well-known gene functions only. Since the datasets evaluated in this study also have metabolites, they will not be included in the evaluation because it would not be a fair comparison. Instead, in our work, the use of pathway information for assessing the clusters allows the integration of both transcriptional data and metabolic profiles for a more general evaluation. Furthermore, our approach takes into account an analysis of the clustering methods regarding coherence and cluster distribution.

This paper is organized as follows. Section 2 provides a description of the dataset used, the clustering methods and the standard measures used for unsupervised clustering comparison. Section 3 explains the new biologically inspired measure proposed for the assessment of clusters. Section 4 presents a comparison of clustering methods through validity and biological analysis. Experimental results regarding the new biologically inspired validity measure and their discussion can be found in Section 5. Finally, the conclusions are presented in Section 6.

## 2 Material and methods

### 2.1 Datasets

In this work, the measures for the comparison of clustering methods were calculated over two metabolic datasets described in this subsection. For the calculation of the biological connectivity in the proposed new measure, we used the Kyoto Encyclopedia of Genes and Genomes (KEGG)<sup>1</sup> pathway database [19].

#### 2.1.1 *Solanum lycopersicum* dataset

The first case study presented in this paper involves the analysis of metabolic and transcriptional profiles from Introgression<sup>2</sup> Lines (ILs) of *Solanum lycopersicum*. The ILs harbor, in certain chromosome segments, introgressed portions of the wild *Solanum*

<sup>1</sup><http://www.genome.jp/kegg/pathway.html/>

<sup>2</sup>Introgression: Infiltration of the genes of one species into the gene pool of another through the repeated backcrossing of an interspecific hybrid with one of its parents.

species (*Solanum pennelli*). The use of ILs allows the study and creation of new varieties of such species by introducing exotic traits and thus constitutes a useful tool in crop domestication and breeding [20, 21]. The interest in comparing the cultivated *Solanum lycopersicum* with the different ILs lies on the fact that some wild *Solanum lycopersicum* fruits have proven to be the source of several specific agronomic traits, which could be used for the improvement of *Solanum lycopersicum* commercial lines. After log-transforming the green/red ratio values over the entire dataset, genes whose expression did not change significantly were discarded from further analysis. As a result of the pre-processing and selection steps, 70 metabolites and 1159 genes were selected [22]. Before the integration of these two datasets, the plus/minus sign of each transcript/metabolite was reversed to obtain items negatively correlated to each other, as suggested in [23, 24]. To find all possible relations<sup>3</sup>, the training set considered here includes the original one as well as the inverted-sign versions of all the data samples. Therefore, the final number of data patterns used to feed the clustering methods was 2458.

### 2.1.2 *Arabidopsis thaliana* dataset

The second biological dataset comprises primary metabolites and transcripts measured in *Arabidopsis thaliana* leaves. The integrated analysis of these data is aimed to study the effects of cold temperatures on circadian-regulated genes in this plant [25]. In this study, we included metabolites and transcripts under light-dark cycles at two control temperatures (20°C and 4°C). Genes involved in diurnal cycle and cold-stress responses were selected for further studies. More details on how the data were processed, filtered and normalized can be found in [25]. A total of 1549 genes and 51 metabolites were used in the integrated analysis. As in the previous dataset, the plus/minus sign of each data point was reversed and the inverted patterns were added to the training set, resulting in a total of 3200 data patterns.

## 2.2 Clustering methods

This subsection presents the clustering methods used for the comparison. In this study, we included a comprehensive evaluation of the clustering methods most widely and most commonly used in computational biology research nowadays [3, 7, 26]. For all the methods, we used the Euclidean distance to measure the distance between patterns because it is the most available and widely used metric in most bioinformatics studies [12, 16, 7].

The use of another distance/similarity could be possible; in fact, there are several available measures [26]. For example, the correlation coefficient has been established to be a suitable quantifier of pairwise gene coexpression. However, the correlation coefficient is good for linear dependence but not for nonlinear one. In fact, alternatives such as that in [26] are being proposed nowadays to better capture the degree of coexpression between a pair of genes with a similar temporal pattern. Nevertheless, it

<sup>3</sup>Direct relations between transcripts ( $t$ ) and metabolites ( $m$ ):  $\uparrow t \leftrightarrow \uparrow m$  (inverted-sign  $\downarrow t \leftrightarrow \downarrow m$ ),  $\uparrow t \leftrightarrow \uparrow t$  (inverted-sign  $\downarrow t \leftrightarrow \downarrow t$ ) and  $\uparrow m \leftrightarrow \uparrow m$  (inverted-sign  $\downarrow m \leftrightarrow \downarrow m$ ). Cross relations:  $\uparrow t \leftrightarrow \downarrow m$  (inverted-sign  $\downarrow t \leftrightarrow \uparrow m$ ),  $\uparrow t \leftrightarrow \downarrow t$  (inverted-sign  $\downarrow t \leftrightarrow \uparrow t$ ) and  $\uparrow m \leftrightarrow \downarrow m$  (inverted-sign  $\downarrow m \leftrightarrow \uparrow m$ ).

has been shown that no method can outperform the Euclidean distance for ratio-based measurements (such as the ones involved in the datasets evaluated in this study) since these types of data are log-transformed before clustering, which compresses the scale of variation, and Euclidean distance has proven to be more robust than other metrics to such processing [18].

The Gap Statistic [27] was used to select an appropriate number of clusters for the comparisons among methods to show the application and use of the proposed measure. As stated in literature [7], this statistic is intended to estimate adequate cluster numbers from a dataset. The Gap Statistic was calculated for  $k$ -means in a range that varied between 2 and 600 clusters. These three top gap scores were selected for the comparisons:  $k_1 = 50$ ,  $k_2 = 200$  and  $k_3 = 450$ .

### 2.2.1 Hierarchical clustering (HC)

Hierarchical clustering (HC) is one of the simplest and most popular unsupervised methods in post-genomic data analyses. It clusters data by forming a tree diagram or dendrogram, which shows the relationships between samples according to proximity matrices. The root node of the dendrogram represents the whole dataset, and each leaf is regarded as a data point. The clusters are obtained by cutting the dendrogram at different levels [8]. HC has three main variants: i) single linkage, also called nearest neighbor, which uses the smallest distance between objects in two clusters to further group them; ii) complete linkage, which uses the largest distance between objects; and iii) average linkage, called un-weighted average distance, which uses the average distance between all pairs of objects in a pair of clusters. The HC algorithm was used in this study in its popular average distance variant (named here HCa), since average linkage is generally considered to be better than both single and complete linkage [18]. In fact, the complete linkage version of HC has been tested as well, but due to the fact that results are very similar to those obtained with average linkage, complete linkage was not included in the comparisons. The single linkage version of HC was not included in the study either, because 95% of the data were grouped in only one cluster. In fact, the shortcomings and poor results of this method have been well-known for long [18].

### 2.2.2 $K$ -means (KM)

$K$ -means (KM) is one of the best-known and most popular clustering algorithms [28, 29, 30]. KM begins by selecting the desired number of  $k$  clusters and assigning centroids to data points randomly chosen from the training set. At every iteration, data points are classified by assigning them to the cluster with the closest centroid and then, new cluster centroids are computed as the average of all the points belonging to each cluster. This process continues until both the cluster centroids and the class assignments no longer change. This technique inherently looks for compact and spherical clusters. Due to the random initialization, 100 iterations of this algorithm were performed and average results are reported.

### 2.2.3 Self-organizing map (SOM)

Self-organizing maps (SOMs) were introduced by Kohonen [31]. They represent a special class of neural networks that use competitive learning. They can represent complex high-dimensional input patterns into a simpler low-dimensional discrete map with prototype vectors that can be visualized in a two-dimensional lattice structure, while preserving the proximity relationships of the original data as much as possible. Thus, SOMs can be appropriate for cluster analysis when looking for underlying hidden patterns in data. They have been recently proposed for the integration and visualization of coordinated variations in transcriptomics and metabolomics data [22]. In this study, for all configurations tested, the SOM map has been trained 100 epochs with the standard batch training algorithm by using gaussian neighborhood functions, and it has been initialized using the Principal Component Analysis [32].

## 2.3 Validation measures

In this subsection, the following notation is used:  $X$  is the dataset formed by  $\mathbf{x}_i$  data samples;  $\Omega$  is the set of samples that have been grouped in a cluster; and  $W$  is the set of  $\mathbf{w}_j$  centroids of the clusters in  $\Omega$ . We will call node to each of the  $k$  elements of the clustering method. The term integration node will be used to refer to a node containing different kinds of patterns: metabolites and transcripts.

A survey of computational cluster validation techniques and measures for biological data is provided in [12], in which the following classification of internal<sup>4</sup> objective measures is proposed:

### Type I: Compactness

It comprises validation measures assessing cluster compactness or homogeneity. Intra-cluster variance is their most popular representative:

$$\bar{C}_j = \frac{1}{|\Omega_j|} \sum_{\forall \mathbf{x}_i \in \Omega_j} \|\mathbf{x}_i - \mathbf{w}_j\|_2, \quad (1)$$

where  $|\cdot|$  stands for set cardinality. As a global measure of compactness, the average of all clusters is calculated  $\bar{C} = \frac{1}{k} \sum_j \bar{C}_j$ , where  $k$  is the number of clusters. Values of  $\bar{C}$  close to 0 indicate more compact clusters.

### Type II: Separation

This group includes all those measures that quantify the degree of separation between individual clusters. Such separation can be evaluated by measuring mean, minimum

<sup>4</sup>Internal measures evaluate the clustering solutions based on clustering results and the classified dataset.

and maximum Euclidean distance among cluster centroids

$$\bar{S} = \frac{2}{k^2 - k} \sum_{i=1}^k \sum_{j=i+1}^k \|\mathbf{w}_i - \mathbf{w}_j\|_2, \quad (2)$$

$$S_m = \min_{0 < i \neq j \leq k} \{\|\mathbf{w}_i - \mathbf{w}_j\|_2\}, \quad (3)$$

$$S_M = \max_{0 < i \neq j \leq k} \{\|\mathbf{w}_i - \mathbf{w}_j\|_2\}, \quad (4)$$

where  $\bar{S}$  close to 0 indicates closer clusters.

### Type III: Combined

They are a combination of the two previous types of measures. They combine the effects of opposing trends. For example, while compactness improves according to the number of clusters, the distance between them tends to deteriorate.

The first combined measurement used in this work is the internal cluster dispersion rate of the final partition, defined as

$$\Upsilon = 1 - \left( \frac{\sum_{j=1}^k \left\| \mathbf{w}_j - \left( \frac{1}{N} \sum_{\ell=1}^N \mathbf{x}_\ell \right) \right\|_2}{\sum_{i=1}^N \left\| \mathbf{x}_i - \left( \frac{1}{N} \sum_{\ell=1}^N \mathbf{x}_\ell \right) \right\|_2} \right). \quad (5)$$

In this equation, the numerator corresponds to the sum of the distances among the centroids and the overall sample mean vector; the denominator is the sum of the distances between each pattern and the overall sample mean vector. The smaller the value of  $\Upsilon$ , the smaller the intracluster dispersion [33].

The Davies-Bouldin index [34] is a popular metric for evaluating clustering algorithms. It is defined as

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left( \frac{\bar{C}_i + \bar{C}_j}{\|\mathbf{w}_i - \mathbf{w}_j\|_2} \right). \quad (6)$$

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

The Dunn Validity Index [35] combines dissimilarity between clusters and their diameters. It is based on the idea of identifying cluster sets that are compact and well separated. It measures inter-cluster distances (separation) over intra-cluster distances (compactness). This index is defined as:

$$D = \frac{\min_{0 < m \neq n \leq k} \left\{ \min_{\substack{\forall \mathbf{x}_i \in \Omega_m \\ \forall \mathbf{x}_j \in \Omega_n}} \{\|\mathbf{x}_i - \mathbf{x}_j\|_2\} \right\}}{\max_{0 < m \leq k} \max_{\forall \mathbf{x}_i, \mathbf{x}_j \in \Omega_m} \{\|\mathbf{x}_i - \mathbf{x}_j\|_2\}}. \quad (7)$$

If a dataset contains well-separated clusters, the distances among them are usually large and their diameter is expected to be small. Therefore, a larger  $D$  value means better cluster configuration.

It should be noticed that due to the fact that the last combined measures are not linear, they are very sensitive to outliers. If a data point is located farther away from the dataset center due to the non-linear max/min selection, it will have a large impact on the measurement result.

### 3 Global measure for linked clustering

Each of the validation measures presented above evaluates different aspects of a clustering solution separately, and based only on the raw data. None of them uses explicit information from the domain of application to evaluate the clusters found. For example, for the data used in this study, the internal measures presented do not evaluate the differences among the solutions found from a biological point of view. Therefore, a way of measuring the biological relevance of the clusters found might be useful, since unsupervised clustering may produce clusters that are useless.

On the one hand, once clusters are found on a dataset, biologists perform a manual examination of the clusters obtained searching for groups that would be useful for inferring new biological knowledge [36]. If the number and size of the clusters is high, this task might be very difficult and time-consuming. Thus, a solution that provides small and cohesive groupings should be more desirable. On the other hand, a cluster solution should be robust and stable to, for example, data perturbation or the number of clusters used [8]. A particular kind of data perturbation is the change of sign of the data samples, as explained at the end of Section 2.1.1, which could give biologists more clues for the identification of new functions from inverse correlations. If an object is clustered together with other similar objects, when the sign of all of these objects is changed they should continue being grouped together. In other words, the normal and inverted-sign versions of a data point should have the same behavior. A coherent grouping of these original and inverted-sign data is to be expected for a good clustering solution.

Moreover, and as stated before, it is common practice in biology to validate a clustering solution obtained by a clustering algorithm according to *a-priori* biological knowledge. For example, it is an established fact that two genes expressed simultaneously have a high chance of sharing common biological pathways. Therefore, under the assumption that genes involved in a biological process behave in a similar way under the same biological process, each of the objects that are part of a cluster have to be analyzed in detail to determine their membership to well-known metabolic pathways. In fact, the study of such similarity patterns is a significant problem to be dealt with nowadays [26]. This way, this pathway-based approach may help in the elucidation of the functions of new genes.

All of these elements are important to qualify a cluster. This work presents a new measure that assesses, in a compact way, the elements aforementioned. Furthermore, it includes an evaluation of clusters from the viewpoint of their biological meaning. The new type of measure is defined in the following way:



#### Type IV: Combined biological assessment

To define it, we need the following three factors:

1. Clustering homogeneity:

$$\check{H} = \frac{1 + \text{med}_m \{|\Omega_m|\}}{\max_m \{|\Omega_m|\}},$$

where the numerator counts the median of the number of elements in the clusters and the denominator is the maximum number of elements in the clusters. The constant value 1 is added to the numerator because this factor will be added along with the others in logarithmic scale, and this avoids having  $\log(0)$  when there is an empty cluster.

$\check{H}$  is a measure of the flatness of the distribution of patterns along clusters. For the analysis of their possible biological relations, it is preferable to have many small clusters than few large ones (with many data points). See Section 4.1 for a discussion of this point.

2. Grouping coherence:

$$\bar{\Gamma} = \frac{1}{|X|} \sum_i \epsilon(-\mathbf{x}_i) \left( \frac{|\Omega_{(i)}| - \Theta_{(-i)}}{|\Omega_{(i)}|} \right),$$

where  $\Omega_{(i)}$  is the node in which the pattern  $i$  is grouped;  $\Theta_{(-i)}$  is the number of misplaced inverted-patterns grouped in the node where  $-\mathbf{x}_i$  is grouped; and the indicator  $\epsilon(-\mathbf{x}_i)$  is 1 only when  $-\mathbf{x}_i \notin \Omega_{(i)}$ .

This factor indicates if the data sample  $\mathbf{x}_i$  has been coherently grouped when having an inverted value. That is to say, the normal and inverted-sign versions of a data point should have the same behavior. For example, let us suppose  $\mathbf{x}_i$  has been grouped together with  $\mathbf{x}_j$  and  $\mathbf{x}_k$ . If the sign of each data point in the dataset is changed, one should expect that  $-\mathbf{x}_i$  would be grouped together with  $-\mathbf{x}_j$  and  $-\mathbf{x}_k$ . If this is the case,  $\Theta_{(-i)} = 0$ . If not,  $\Theta_{(-i)}$  counts how many of the original data points that were grouped together with  $\mathbf{x}_i$  are not grouped now with  $-\mathbf{x}_i$ . See Section 4.2.1 for an example of a detailed analysis of coherence in a real dataset.

3. Internal connectivity:

$$\bar{P} = \frac{1}{k} \sum_m \frac{p_m}{p_{m*}},$$

where

$$p_m = 1 + \sum_{i \in \Omega_m} \sum_{j \in \Omega_m, j \neq i} \rho_{ij}$$

is the number of common pathways among patterns grouped in cluster  $m$ , with  $\rho_{ij}$  the number of pathways that contain patterns  $i$  and  $j$ , and

$$p_{m*} = 1 + \sum_{i \in \Omega_m} \sum_{j \neq i} \rho_{ij}$$

is the number of all the possible shared pathways among patterns grouped in cluster  $m$  and any other pattern in the dataset.

The new Global Measure for Linked Clustering (GMLC) is defined as a weighted combination

$$G = \gamma_H \log(\check{H}) + \gamma_\Gamma \log(\bar{\Gamma}) + \gamma_P \log(\bar{P}), \quad (8)$$

where the  $\gamma$  parameters are empirically determined.

A simple criterion to set them, according to their distribution in a given dataset, could be to define  $\gamma_P = -1$  and calculate the other weights as the ratio of averages

$$\gamma_H = -\mathcal{E}_\ell [\log \bar{P}_\ell] / \mathcal{E}_\ell [\log \bar{H}_\ell],$$

and

$$\gamma_\Gamma = -\mathcal{E}_\ell [\log \bar{P}_\ell] / \mathcal{E}_\ell [\log \bar{\Gamma}_\ell].$$

Here,  $\mathcal{E}_\ell$  is the average over all the cluster methods and configurations evaluated with the dataset. Thus, in order to equate the influence of all three terms on the final GMLC score, one of the  $\gamma$  parameters can be taken as a reference (any of them) and the others scaled accordingly.

## 4 Analysis of clustering methods

This Section presents, first of all, a comparison of the clustering solutions with respect to cluster validity and distribution. After that, the methods are analyzed from a biological perspective.

### 4.1 Validity comparison and cluster distribution

The application of standard measures to the clusters obtained from both datasets *Solanum lycopersicum* and *Arabidopsis thaliana* are presented in Table 1 and 2, respectively. The best value for each measure in each of the three  $k$  columns is underlined. As an additional reference for the comparison, we calculated the Gap Statistic for  $k_1$ ,  $k_2$  and  $k_3$  in both datasets: 0.697, 0.688 and 0.707 (*Solanum lycopersicum*); 0.797, 0.784 and 0.793 (*Arabidopsis thaliana*). It can be seen that the values are high and extremely close. In fact, as stated at the end of Section 2.2., the three best gap scores were chosen for the comparisons among clustering methods.

To disregard the influence of the distance measure used for clustering, we performed a one-way analysis of variance (ANOVA). The null hypothesis states that the difference between Euclidean distance and correlation coefficient is not significant. The p-values obtained for the means of the validation scores were  $p > 0.01$ . Thus the null hypothesis was accepted and only the results with Euclidean distance were reported in the tables. This is shown graphically for  $k_1$  and KM with the *Solanum lycopersicum* dataset in the boxplot of Figure 1. For each measure in the axis, the results obtained using the Euclidean distance (left box) vs correlation coefficient (right box) are presented.

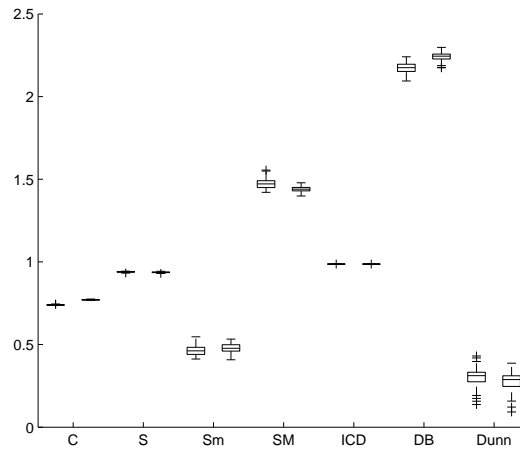


Figure 1: Boxplot showing, for each validation measure, the results obtained using Euclidean distance (left box) vs correlation coefficient (right box), for  $k_1$  and KM in the *Solanum lycopersicum* dataset.

Considering only Type I measure (cluster compactness) from the *Solanum lycopersicum* dataset, it could be stated that HCa is the best clustering method in any of its tested configurations. However, for the *Arabidopsis thaliana* dataset the best results are rather distributed along the HCa and KM methods, depending on the number of clusters tested. According to Type II measures, contradictory results with respect to the compactness measure were obtained. In the first dataset, while HCa was clearly indicated as a good method according to the compactness of its clusters, now separation indicated SOM as the best candidate in all its variants. The same happened in the case of the second dataset. However, a close look at the values obtained by Type II measures showed that they are actually quite similar to all the methods and configurations evaluated. All clusters are almost equally separated from each other in all cases.

Since these two types of measures evaluate opposite aspects of a clustering solution, Type III measures were calculated because they combine those two types into one single index. Due to this fact one would expect a better discerning capacity. In the case of  $\Upsilon$ , the internal cluster dispersion is always close to the optimum for all methods. This measure is not useful for choosing one method among all the possibilities available. Interestingly, while the Dunn measure qualifies clusters taking into account the same general criteria as DB (high compactness and separation), it has clearly favored HCa over KM in the *Solanum lycopersicum* dataset for all the configurations tested. This result is different from the DB index indication for two configurations over the *Solanum lycopersicum* dataset. In the *Arabidopsis thaliana* dataset, a similar situation occurred since there was a lack of concordance. Moreover, both DB and Dunn indexes reflect dissimilar results with respect to Type I and Type II measures. These contradictory results can be very confusing. As it can be seen, it is also difficult here to indicate a solution and configuration as the most adequate.

Another way of evaluating clustering methods in a dataset is by using histograms

Table 1: *Solanum lycopersicum* dataset. Comparison of validation measures for clustering methods. The best value of each  $k$  for each measure is underlined.

	HCa			KM			SOM		
	$k_1$	$k_2$	$k_3$	$k_1$	$k_2$	$k_3$	$k_1$	$k_2$	$k_3$
$C$	<u>0.69</u>	<u>0.55</u>	<u>0.47</u>	0.74	0.63	0.53	0.81	0.71	0.64
$\bar{S}$	1.00	1.20	1.31	0.94	1.08	1.18	<u>0.59</u>	<u>0.77</u>	<u>0.87</u>
$S_m$	0.63	0.47	0.52	0.46	0.47	0.48	<u>0.13</u>	<u>0.11</u>	<u>0.10</u>
$S_M$	1.37	1.68	1.79	1.47	1.73	1.80	<u>1.16</u>	<u>1.34</u>	<u>1.44</u>
$\Upsilon$	<u>0.99</u>	0.93	0.83	<u>0.99</u>	0.94	<u>0.93</u>	<u>0.99</u>	<u>0.95</u>	0.88
$DB$	3.09	3.18	<u>3.13</u>	<u>2.17</u>	<u>1.87</u>	3.32	8.65	10.5	11.7
$D$	<u>0.37</u>	<u>0.46</u>	<u>0.56</u>	0.31	0.41	0.55	0.32	0.41	0.47

Table 2: *Arabidopsis thaliana* dataset. Comparison of validation measures for clustering methods. The best value of each  $k$  for each measure is underlined.

	HCa			KM			SOM		
	$k_1$	$k_2$	$k_3$	$k_1$	$k_2$	$k_3$	$k_1$	$k_2$	$k_3$
$C$	2.87	<u>2.31</u>	<u>1.90</u>	<u>2.80</u>	2.43	2.10	2.94	2.52	2.31
$\bar{S}$	7.03	7.34	7.47	6.35	6.70	6.99	<u>4.95</u>	<u>5.50</u>	<u>5.75</u>
$S_m$	2.91	2.16	2.01	1.78	1.31	1.34	<u>0.59</u>	<u>0.40</u>	<u>0.27</u>
$S_M$	9.80	10.3	10.6	10.3	10.7	10.94	<u>9.68</u>	<u>10.1</u>	<u>10.4</u>
$\Upsilon$	<u>0.99</u>	0.94	0.86	<u>0.99</u>	0.94	<u>0.94</u>	<u>0.99</u>	<u>0.95</u>	0.88
$DB$	3.16	<u>3.91</u>	<u>4.37</u>	<u>2.08</u>	5.24	5.45	10.05	16.5	19.7
$D$	<u>0.50</u>	<u>0.48</u>	<u>0.52</u>	0.19	0.29	0.44	0.17	0.18	0.24

of pattern distribution. The histograms of patterns for each cluster, for each method, and with  $k_2$ , were also calculated for the *Solanum lycopersicum* dataset (Figure 2). As it can be seen, HCa comprises the vast majority of the patterns in a few branches. When comparing the histograms of KM and SOM, it can be seen that their distribution is quite similar and uniform. On the one hand, a recent study of clustering methods in biological datasets has shown that there is a relationship between reduced coverage<sup>5</sup> and an increase in the ability of a clustering algorithm to group the samples correctly [37]. On the other hand, since unsupervised clustering groups data without any biological constraints, it may not produce clusters that are useful to discover new biological knowledge. Thus, biologists have to manually examine the clustering solutions to interpret the results and draw conclusions, which might be very difficult when a cluster is very large. Therefore, it can be stated that a uniform distribution of patterns may give biologists more confidence in finding interesting relationships among data.

Besides, we statistically tested the significance of all the results obtained by performing 100 resamplings of 80% of the transcripts and metabolites in each dataset, for all the methods in each  $k$ . The ANOVA was performed to test the null hypothesis in which, for each partition  $k$ , the difference among the clustering methods, in a given combined measure, is not significant. The analysis revealed that all the combined measures show significant differences ( $p < 0.01$ ) among the methods in each

<sup>5</sup>Reduced coverage: in the partition obtained there are more clusters than actually exist in the underlying data.

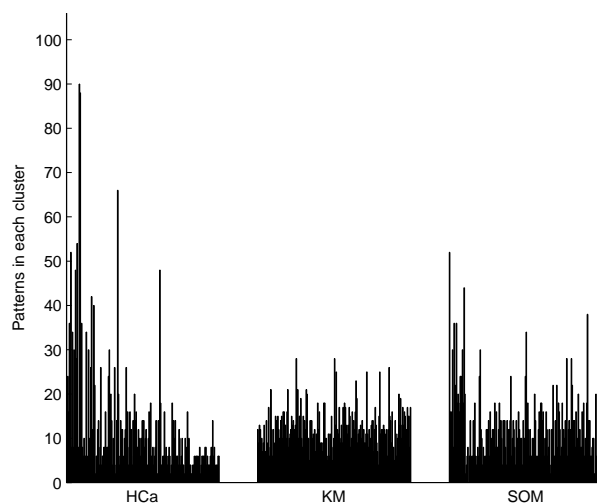


Figure 2: Pattern distribution histograms by clustering method for size  $k_2$  for the *Solanum lycopersicum* dataset.

partition, which validates the conclusions obtained from the detailed analysis of each of the measures in the tables.

A good clustering solution should perform well under multiple evaluation points. However, as it can be seen from this subsection, it is very difficult to indicate one of the evaluated methods as the clear winner according to all the measures applied.

## 4.2 Biological clusters analysis

In this subsection, the clustering methods are analyzed from a biological perspective. Some examples are presented to show the two typical tasks that biologist generally perform over the clusters found by a clustering algorithm. These illustrative examples will be developed in two sections by using the same *Solanum lycopersicum* dataset in different ways. In Section 4.2.1 coherence will be checked. Then, in Section 4.2.2, a well-known pathway for validation will be used with the aim to figure out if compounds clustered together are connected under the guilt-by-association principle. The ultimate objective is to show how these biological analyses are not reflected in the objective measures presented in the previous subsection.

### 4.2.1 Coherence

Table 3 shows an example of a coherent grouping analysis that can be performed, for instance, by looking at some metabolites of the *Solanum lycopersicum* training set. These metabolites are the main intermediates in the well-known metabolic pathway of Citrate cycle (TCA or Krebs cycle<sup>6</sup>) and closely related pathways such as alanine

<sup>6</sup><http://www.genome.jp/kegg/pathway/map/map00020.html>

metabolism<sup>7</sup>.

The left part of the table shows the direct values of each metabolite. The right part has the same metabolites with inverted values (changed sign) and their corresponding clustering locations. In order to better highlight the incoherence of the clusters obtained, some numbers are colored. When a metabolite is grouped in a node together with other elements, but the inverted version of this metabolite does not group with those inverted elements as well, the numbers are painted black. On the contrary, when an inverted metabolite is clustered with other inverted metabolites, but their direct versions are not, dark-gray color is used.

In the analyzed example, HCa and SOM consistently grouped all the compounds for all cluster numbers and, therefore, they were not included in the table. For example, for HCa with  $k_1$ , the direct compounds *alanine* and *GABA* clustered in node 41 and were also grouped together in node 42 when their sign was inverted. However, in the case of KM with  $k_2$ , the metabolites *citrate*, *glutamate* and *2-oxoglutarate* fitted into node 109 and their inverted patterns *citrate(inv)*, *glutamate(inv)* and *2-oxoglutarate(inv)* were split into three different clusters 115, 191 and 16, respectively. This case is marked in Table 3 with black background. An example of the opposite kind of incoherence happens in KM with  $k_1$ , where *alanine(inv)*, *asparagine(inv)* and *GABA(inv)* grouped together in cluster 50, but *alanine* was assigned to cluster 42 while *asparagine* and *GABA* were put together in cluster 17. In this case, the cluster numbers were painted dark-gray. These inconsistencies (found in all the KM repetitions for all the numbers of clusters tested) are certainly a limitation of KM and throw doubts on its applicability to these kinds of biological data. However, as shown in the previous subsection, this aspect has no impact on any of the standard measures when they are applied to qualify a clustering solution.

#### 4.2.2 Assessment regarding a metabolic pathway

The choice of biological processes common to the vast majority of organisms is an important starting point for a comparison between clustering algorithms, because it is assumed that any method used to analyze biological data should be able to find such relations in a few nodes. For this purpose, the aforementioned metabolic pathways were used. In this part of the analysis, a detailed inspection was performed to verify the membership of the patterns in each cluster to a well-known metabolic pathway.

Figure 3 shows the glycolytic and TCA cycle intermediates and enzymes encoding genes connecting these pathways with amino acid metabolism. Those measured compounds that belong to the training set are marked with a rectangle. For each marked metabolite, the three numbers above the rectangle indicate the cluster number where the metabolite was grouped in HCa, KM and SOM, respectively, for size  $k_2$ . It can be noticed that most of the marked compounds grouped together in a few SOM nodes. Meanwhile, in the case of HCa and KM, the data points are spread over a large number of different nodes.

<sup>7</sup><http://www.genome.jp/kegg/pathway/map/map00250.html>

Table 3: Biological analysis of cluster coherence for KM over the *Solanum lycopersicum* dataset. Black background: clusters with metabolites whose inverted versions were not grouped coherently. Dark-gray background: clusters with inverted metabolites whose direct versions were not grouped coherently.

	direct			inverted		
	$k_1$	$k_2$	$k_3$	$k_1$	$k_2$	$k_3$
alanine	42	78	78	50	127	72
asparagine	17	105	159	50	99	161
aspartate	9	140	438	42	138	423
citrate	32	109	186	49	115	424
fumarate	3	92	15	24	194	64
GABA	17	105	261	50	43	130
glutamate	1	109	349	20	191	424
malate	5	83	167	13	156	199
2-oxoglutarate	4	109	113	21	16	379
succinate	38	95	178	22	177	270

## 5 Biologically inspired measure: results and discussion

To be able to show the new Type IV measure together with Type I, II and III, Table 4 shows the results obtained in the comparison of the measures to discern among several clustering methods, considering in this case only integration nodes in the *Solanum lycopersicum* dataset. A similar analysis for the *Arabidopsis thaliana* data is shown in Table 5.

The interest in this particular analysis lies on the fact that the enzymes and metabolites grouped together into integration nodes may be part of the same metabolic pathway according to the guilt-by-association principle. Five rows were added to these tables with respect to Table 1 and 2: the second row shows now the percentage of integration nodes found by each technique in relation to the total number of clusters (*int%*), whereas the last four rows show the new GMLC and the three factors involved in its computation. We have statistically analyzed the significance of the obtained results using an ANOVA. The null hypothesis states that for each  $k$  the difference among the clustering methods is not significant. The test showed that are significant differences among the methods in each partition ( $p < 0.001$ ) for all the combined measures, including the new GMLC.

Although Table 4 shows that similar results to Table 1 for the *Solanum lycopersicum* dataset were obtained with respect to Type II measures, SOM showed now the best cohesion rate. A similar reasoning can be applied to Table 5 for the *Arabidopsis thaliana* case. The internal cluster dispersion results do not vary with respect to the previous tables. In the case of the DB and Dunn indexes, they also highlight HCa and KM here, since they favor compact clusters well separated from each other. However, for being measures that summarize compactness and separation, they contradict Type I

Table 4: *Solanum lycopersicum* dataset. Comparison of validation measures for the clustering methods, considering only integration nodes. The best value of each  $k$  for each measure is underlined.

int.%→	HCa			KM			SOM		
	$k_1$	$k_2$	$k_3$	$k_1$	$k_2$	$k_3$	$k_1$	$k_2$	$k_3$
	50	18	6	58	32	17	56	20	8
$\overline{C}$	1.20	1.26	1.32	1.17	1.25	1.30	<u>1.01</u>	<u>1.10</u>	<u>1.12</u>
$\overline{S}$	1.00	1.19	1.28	0.94	1.08	1.18	<u>0.54</u>	<u>0.67</u>	<u>0.79</u>
$S_m$	0.65	0.65	0.56	0.50	0.53	0.55	<u>0.13</u>	<u>0.12</u>	<u>0.11</u>
$S_M$	1.34	1.61	1.71	1.43	1.63	1.73	<u>0.91</u>	<u>1.08</u>	<u>1.19</u>
$\Upsilon$	<u>0.99</u>	0.93	0.83	<u>0.99</u>	0.94	<u>0.93</u>	<u>0.99</u>	<u>0.95</u>	0.88
$DB$	<u>2.97</u>	<u>2.82</u>	<u>2.62</u>	3.38	3.36	3.21	10.5	12.8	11.2
$D$	0.38	<u>0.55</u>	0.54	0.32	0.47	0.61	<u>0.47</u>	0.50	<u>0.66</u>
$\gamma_H \log(\overline{H})$	6.78	6.47	3.74	1.42	5.51	4.32	2.39	4.60	3.92
$\gamma_\Gamma \log(\overline{\Gamma})$	0.00	0.00	0.00	12.8	14.9	11.3	0.00	0.00	0.00
$\gamma_P \log(\overline{P})$	3.89	4.23	5.36	3.54	4.15	4.71	4.10	4.56	4.61
$G$	10.67	10.70	9.10	17.76	24.56	20.33	<u>6.49</u>	<u>9.16</u>	<u>8.53</u>

and II indications in both tables. Moreover, the results are contradictory with respect to all the previous analyses. For example, in the case of the *Solanum lycopersicum* data, the Dunn index in Table 1 indicated HCa as the most suitable method for all  $k$  partitions. However, when looking at the integration nodes only in Table 4, SOM is better than the others in two cases. There are also opposite and inconsistent indications for the *Arabidopsis thaliana* data. In general, SOM always obtains the highest DB scores because the distances between centroids are always the smallest since these centroids are better distributed and are not associated with remote and isolated patterns. Since farther away patterns (probably *outliers*) have to be associated with any centroid, cluster compactness also decreases. However, from a biological point of view, it would be useful to have clusters with a high DB index because there are patterns that should be close to many other patterns, if we think that the groupings reflect components of common metabolic pathways and that there are patterns that certainly participate in several pathways simultaneously.

This detailed analysis of the measures over integration clusters shows contradictory results with respect to the all-cluster analysis. Again, although objective measures should give an indication of which clustering technique would be more appropriate for the dataset under study, it is very difficult to clearly select a method as the one providing more interesting clusters to be analyzed by biologists in order to discover new relationships among data. Furthermore, this analysis does not provide a clear clue regarding the connection or the reason why the integrated data clusters have been found, from the viewpoint of their involvement in a common metabolic pathway.



Table 5: *Arabidopsis thaliana* dataset. Comparison of validation measures for the clustering methods, considering only integration nodes. The best value of each  $k$  for each measure is underlined.

int.%→	HCa			KM			SOM		
	$k_1$	$k_2$	$k_3$	$k_1$	$k_2$	$k_3$	$k_1$	$k_2$	$k_3$
	16	6	2	62	23	10	48	15	6
$\overline{C}$	7.06	7.21	8.66	6.85	7.11	7.19	<u>6.55</u>	<u>6.33</u>	<u>6.29</u>
$\overline{S}$	7.09	7.36	7.06	6.35	6.69	6.99	<u>4.58</u>	<u>4.55</u>	<u>5.40</u>
$S_m$	3.08	3.09	3.47	1.97	1.86	1.87	<u>0.59</u>	<u>0.59</u>	<u>0.72</u>
$S_M$	9.08	9.61	10.2	10.1	10.3	10.5	<u>9.02</u>	<u>8.52</u>	<u>9.22</u>
$\Upsilon$	<u>0.99</u>	0.94	0.86	<u>0.99</u>	0.94	<u>0.94</u>	<u>0.99</u>	<u>0.95</u>	0.88
$DB$	<u>2.67</u>	<u>2.89</u>	<u>3.63</u>	5.12	4.93	4.52	14.9	15.7	13.0
$D$	0.18	<u>0.43</u>	<u>0.70</u>	<u>0.21</u>	0.31	0.47	0.17	0.25	0.33
$\gamma_H \log(\overline{H})$	6.97	6.33	5.32	1.38	4.11	3.29	4.68	3.22	2.09
$\gamma_\Gamma \log(\overline{\Gamma})$	0.00	0.00	0.00	12.1	13.4	11.9	0.00	0.00	0.00
$\gamma_P \log(\overline{P})$	3.11	3.99	4.23	3.63	4.26	4.63	4.04	5.05	4.40
$G$	10.08	10.32	9.55	17.1	21.7	19.8	<u>8.72</u>	<u>8.27</u>	<u>6.49</u>

A Gene Ontology (GO) based analysis [38] was performed for further validation of the results using the widely available biological annotations for *Arabidopsis thaliana*. The GO-based score measures how biologically homogeneous are the clusters by checking whether grouped genes also belong to the same functional classes. In the case of  $k_1$ , the score was 0.10, 0.12 and 0.12 for HCa, KM and SOM, respectively. In particular, for the KM method, the values obtained for  $k_1$ ,  $k_2$  and  $k_3$  were 0.12, 0.14 and 0.13, respectively. It should be noticed that this GO-based measure evaluates clusters according only to transcripts (derived from genes) and their functions. However, the datasets evaluated in this study involve not only transcripts but also metabolites, and the GMLC uses information (metabolic pathways) integrating both molecular entities.

It can be observed that the points addressed in Sections 4.1, 4.2.1 and 4.2.2 are contemplated in the values of the proposed GMLC. Thus, the three parts of this measure assess the three parts of the analysis of clusters that biologists often make, which is summarized into a single and compact measure. The first and second parts evaluate the homogeneity and coherence of a solution, while the third part is a pathway-related metric that evaluates internal connectivity. This is clearly reflected in the results obtained when the new Type IV measure is applied to the integration clusters of the methods here evaluated, as shown in the last rows of Table 4 and 5. Moreover, the proposed GMLC is consistent with the results obtained from the application of Type I and II measures for the integration nodes as well as for most of the all-node analyses in both databases. Furthermore, it includes a biological point of view in the evaluation of clustering solutions.

The  $\overline{H}$  values are a measure of the homogeneity of the solutions. For both datasets

analyzed, KM and SOM have similar  $\tilde{H}$  scores. The distribution patterns of HCa are clearly non-uniform, which is reflected by its high  $\tilde{H}$  values. In fact, this was shown above in the histogram of Figure 2 for the first dataset with  $k_2$  and is now reflected in this part of the GMLC measure with the highest  $\tilde{H}$  value for this partition with respect to KM and SOM.

The  $\bar{\Gamma}$  value highlights whether there are patterns that were grouped coherently or not. Its value equal to 0 for HCa and SOM in all cases and for both datasets is a clear sign that these methods coherently group the data points. KM, instead, obtained a high value for this score. In fact, a detailed analysis of the clusters provided by this method in one dataset indicates that patterns were not grouped coherently in all cases. It was shown above that there are clusters that group non-inverted versions of patterns, but their inverted versions were dispersed along several nodes, even mixing non-inverted and inverted versions of data in some cases (which is also inconsistent). This fact is shown by a  $\bar{\Gamma}$  value higher than 0.

The  $\bar{P}$  score evaluates elements grouped together and belonging to the same metabolic pathway. It shows that there are differences among the methods tested, according to each partition, for both datasets. In fact, the detailed analysis of the clusters found in each method, counting the elements that were grouped together and that belong to the same metabolic pathway, showed several differences among them. It can be noticed, for example, that for the *Solanum lycopersicum* dataset and the  $k_3$  case, SOM is better than HCa and KM for obtaining nodes with interrelated patterns by pathways. While for the *Arabidopsis thaliana* dataset and  $k_3$ , the most suited method under this point of view would be HCa.

## 6 Conclusions

This work proposes a new approach for evaluating both coherence and biological significance of clusters found by unsupervised clustering methods over biological datasets. Looking at the analysis made by biologists when they evaluate the elements that are part of a cluster, coherent grouping of the measured components, as well as their relative belonging to well-known metabolic pathways, were evaluated.

A novel validity measure for the biological significance of clustering solutions is proposed. This is addressed from the perspective of the usefulness of clusters to identify those patterns that change in coordination and belong to common pathways of metabolic regulation. The proposed Global Measure for Linked Clustering reflects, in a compact way, the objective analysis of the clustering methods regarding coherence and cluster distribution. Moreover, it also evaluates their biological internal connections considering common pathways.

The application of the GMLC to two biological datasets presented consistent results for several unsupervised clustering methods and cluster numbers tested. The GMLC could aid in deciding which clustering solution to use in order to obtain a coherent and a biologically significant solution for a particular biological experiment. Furthermore, it could be used as an optimization criterion during cluster formation in order to guide the clustering process towards better and biologically interpretable meaningful solutions.

## Acknowledgements

This work was supported by the National Scientific and Technical Research Council [PIP 1122008, PIP 2072008, PIP 1142009]; the National Agency for the Promotion of Science and Technology [PICT 2008 00100, PICT 2008 00140, PAE 37122, PICT 2009 00152] and the National Institute for Agricultural Technology [PE 243542] of Argentina.

## References

- [1] E. Keedwell and A. Narayanan, *Intelligent Bioinformatics: The Application of Artificial Intelligence Techniques to Bioinformatics Problems*. Wiley, 2005.
- [2] P. V. Gopalacharyulu, E. Lindfors, J. Miettinen, C. K. Bounsaythip, and M. Oresic, "An integrative approach for biological data mining and visualisation," *International journal of data mining and bioinformatics*, vol. 2, no. 1, pp. 54–77, 2008.
- [3] S. Datta and S. Datta, "Evaluation of clustering algorithms for gene expression data," *BMC Bioinformatics*, vol. 7, pp. S17+, 2006.
- [4] G. B. Fogel, "Computational intelligence approaches for pattern discovery in biological systems," *Briefings in Bioinformatics*, vol. 9, no. 4, pp. 307–316, 2008.
- [5] B. Andreopoulos, A. An, X. Wang, and M. Schroeder, "A roadmap of clustering algorithms: finding a match for a biomedical application," *Briefings in Bioinformatics*, vol. 10, no. 3, pp. 297–314, 2009.
- [6] M. Vignes and F. Forbes, "Gene clustering via integrated markov models combining individual and pairwise features," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 6, pp. 260–270, April 2009.
- [7] O. Rubel, G. Weber, M.-Y. Huang, E. W. Bethel, M. Biggin, C. Fowlkes, C. L. Hendriks, S. Keranen, M. Eisen, D. Knowles, J. Malik, H. Hagen, and B. Hamann, "Integrating data clustering and visualization for the analysis of 3d gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 7, pp. 64–79, 2010.
- [8] R. Xu and D. C. Wunsch, *Clustering*, D. B. Fogel, Ed. Wiley and IEEE Press, 2009.
- [9] C. J. Wolfe, I. S. Kohane, and A. J. Butte, "Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks." *BMC Bioinformatics*, vol. 6, pp. 227–237, 2005.
- [10] V. Lacroix, L. Cottret, P. Thebault, and M.-F. Sagot, "An introduction to metabolic networks and their structural analysis," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 5, no. 4, pp. 594–617, 2008.

- [11] T. Tohge and A. Fernie, "Combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function," *Nature Protocols*, vol. 5, no. 6, pp. 1210–1227, June 2010.
- [12] J. Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis," *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, 2005.
- [13] J. Freudenberg, V. Joshi, Z. Hu, and M. Medvedovic, "Clean: Clustering enrichment analysis," *BMC Bioinformatics*, vol. 10, pp. 234–244, 2009.
- [14] S. Datta and S. Datta, "Validation measures for clustering algorithms incorporating biological information," in *Proceedings of the First International Multi-Symposiums on Computer and Computational Sciences - Volume 1 (IMSCCS'06)*. IEEE Computer Society, 2006, pp. 131–135.
- [15] I. Gat-Viks, R. Sharan, and R. Shamir, "Scoring clustering solutions by their biological relevance," *Bioinformatics*, vol. 19, no. 18, pp. 2381–2389, 2003.
- [16] V. Pihur, S. Datta, and S. Datta, "Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach," *Bioinformatics*, vol. 23, no. 13, pp. 1607–1615, 2007.
- [17] D. Huang and W. Pan, "Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data," *Bioinformatics*, vol. 22, no. 10, pp. 1259–1268, 2006.
- [18] F. Gibbons and F. Roth, "Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation," *Genome Research*, vol. 12, pp. 1574–1581, 2002.
- [19] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, pp. 27–30, 2000.
- [20] L. Rieseberg and J. Wendel, *Introgression and its consequences in plants*, R. G. Harrison, Ed. Oxford University Press, 1993, vol. 1.
- [21] Z. Lippman, Y. Semel, and D. Zamir, "An integrated view of quantitative trait variation using tomato interspecific introgression lines," *Current Opinion in Genetics and Development*, vol. 17, pp. 1–8, 2007.
- [22] G. Stegmayer, D. Milone, L. Kamenetzky, M. Lopez, and F. Carrari, "Neural network model for integration and visualization of introgressed genome and metabolite data," in *IEEE International Joint Conference on Neural Networks*. IEEE Computational Intelligence Society, 2009, pp. 3177–3183.
- [23] M. Yano, S. Kanaya, M. Altaf-Ul-Amin, K. Kurokawa, M. Y. Hirai, and K. Saito, "Integrated data mining of transcriptome and metabolome based on bl-som," *Journal of Computer Aided Chemistry*, vol. 7, pp. 125–136, 2006.

- [24] K. Saito, M. Y. Hirai, and K. Yonekura-Sakakibara, "Decoding genes with co-expression networks and metabolomics - majority report by precogs," *Trends in Plant Science*, vol. 13, pp. 36–43, 2008.
- [25] C. Espinoza, T. Degenkolbe, C. Caldana, E. Zuther, A. Leisse, L. Willmitzer, D. Hinch, and M. Hannah, "Interaction with Diurnal and Circadian Regulation Results in Dynamic Metabolic and Transcriptional Changes during Cold Acclimation in Arabidopsis." *PLoS one*, vol. 5, no. 11, 2010.
- [26] S. Bandyopadhyay and M. Bhattacharyya, "A biologically inspired measure for coexpression analysis," *IEEE/ACM Trans. Comput. Biology Bioinform.*, vol. 8, no. 4, pp. 929–942, 2011.
- [27] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a dataset via the gap statistic," *J. R. Statist. Soc. B.*, vol. 63, pp. 411–423, 2001.
- [28] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. Wiley, 2003.
- [29] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.
- [30] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, pp. 651–666, 2010.
- [31] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, pp. 59–69, 1982.
- [32] D. Milone, G. Stegmayer, L. Kamenetzky, M. Lopez, J. Giovannoni, J. M. Lee, and F. Carrari, "omeSOM: a software for integration, clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants," *BMC Bioinformatics*, vol. 11, pp. 438–448, 2010.
- [33] S. A. Mingoti and J. O. Lima, "Comparing som neural network with fuzzy c-means, k-means and traditional hierarchical clustering algorithms," *European Journal of Operational Research*, vol. 174, no. 3, pp. 1742–1759, November 2006.
- [34] D. Davies and D. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 1, no. 4, pp. 224–227, 1979.
- [35] J. Dunn, "Well separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, pp. 95–104, 1974.
- [36] D. Dotan-Cohen, S. Kasif, and A. A. Melkman, "Seeing the forest for the trees: using the gene ontology to restructure hierarchical clustering," *Bioinformatics*, vol. 25, no. 14, pp. 1789–1795, 2009.
- [37] M. de Souto, I. Costa, D. de Araujo, T. Ludermiter, and A. Schliep, "Clustering cancer gene expression data: a comparative study," *BMC Bioinformatics*, vol. 9, pp. 497–507, 2008.

- [38] G. Brock, V. Pihur, S. Datta, and S. Datta, "clvalid: An r package for cluster validation," *Journal of Statistical Software*, vol. 25, no. 4, pp. 1–22, 2008.

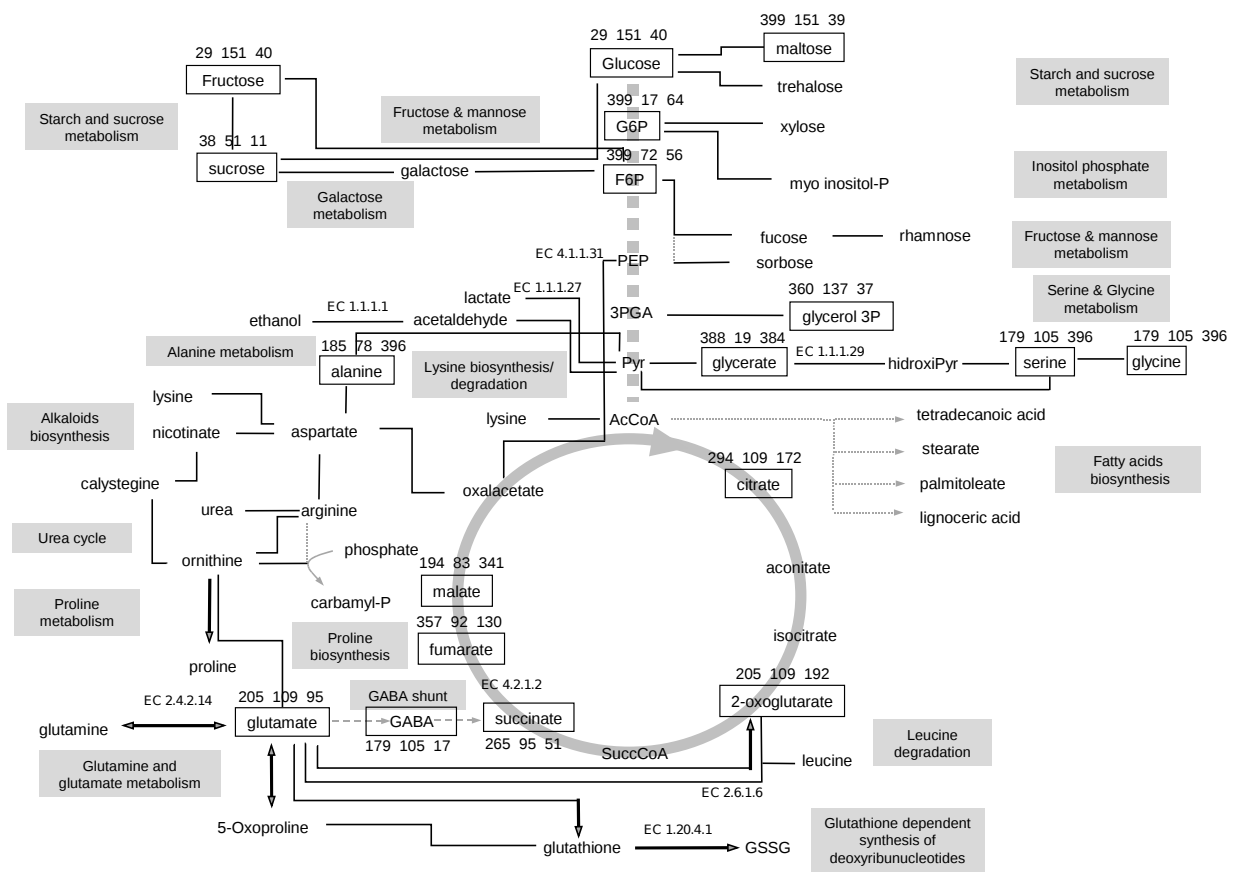


Figure 3: Simplified representation of the primary carbon metabolic pathways (mainly glycolysis and TCA) in *Solanum lycopersicum* fruits. Pathways are represented. Full arrows indicate single reactions and dotted arrows represent multiple chemical reactions. Associated biological pathways are highlighted in gray boxes. Compounds measured and belonging to the training set are marked with a rectangle. The three numbers above the rectangle indicate the cluster number for size  $k_2$  in HCa, KM and SOM, respectively.