Data mining over biological datasets: an integrated approach based on computational intelligence

Georgina Stegmayer, *Member, IEEE,* Matias Gerard and Diego H. Milone, *Member, IEEE*

Abstract

Biology is in the middle of a data explosion. The technical advances achieved by the genomics, metabolomics, transcriptomics and proteomics technologies in recent years have significantly increased the amount of data that are available for biologists to analyze different aspects of an organism. However, $*$omics data sets have several additional problems: they have inherent biological complexity and may have significant amounts of noise as well as measurement artifacts. The need to extract information from such databases has once again become a challenge. This requires novel computational techniques and models to automatically perform data mining tasks such as integration of different data types, clustering and knowledge discovery, among others. In this article, we will present a novel integrated computational intelligence approach for biological data mining that involves neural networks and evolutionary computation. We propose the use of self-organizing maps for the identification of coordinated patterns variations; a new training algorithm that can include a priori biological information to obtain more biological meaningful clusters; a validation measure that can assess the biological significance of the clusters found; and finally, an evolutionary algorithm for the inference of unknown metabolic pathways involving the selected clusters.

Index Terms

Data mining, systems biology, clustering, validation, biological assessment, pathways analysis, evolutionary search.

1 Introduction

Modern biology studies generate a large amount of data, that require dedicated computational tools for their analysis. Data integration is also gaining importance given the need for extracting knowledge from multiple data types and

- *G. Stegmayer is with Center for Research & Development of Information Systems (CIDISI), National Scientific and Technical Research Council (CONICET), Argentina (email: georgina.stegmayer@ieee.org).*
- *D. H. Milone and M. Gerard are with Research Center for Signals, Systems and Computational Intelligence (sinc(i)), FICH-UNL, National Scientific and Technical Research Council (CONICET), Argentina (email: d.milone@ieee.org).*

sources, with the aim of infering insights from the genetic processes underlying them [1], [2], [3]. In fact, since the completion of genome sequences, functional identification of unknown genes has become a principal challenge in systems biology. Bioinformatics plays an important role here, allowing biologists to make full use of the advances in computer science in analyzing large and complex datasets.

At the beginning of the genomics revolution, bioinformatics referred only to the creation and management of large databases to store biological data. However, the discipline has evolved over time, mainly from the application and adaptation of classical statistical methods and standard clustering algorithms, such as hierarchical clustering (HC) and k-means (KM) [4], [5], [6], towards more recent approaches based on computational intelligence [7], [8], [9], with promising results. Yet their application to bioinformatics problems has gained popularity only recently [10].

From an application point of view, a current trend is to achieve integration of different types of biological data to reveal hidden correlations between them, allowing the inference of new knowledge regarding the biological processes that affect them. However, the discovery of hidden patterns in such data is currently a challenge because the use of any type of algorithm for pattern recognition is hampered by a limited number of samples and a very high number of dimensions. Besides, biological data sets may have significant amounts of noise as well as measurement artifacts. This highlights the need to develop new techniques aimed at overcoming the limitations of existing ones. New computational models to perform several data mining tasks, such as integration of different data types, unsupervised clustering and knowledge discovery, are required.

In this article we will present a novel integrated computational intelligence approach for biological data mining (Figure 1). It involves the use and application of two of the most important and well-tested techniques in the computational intelligence field: neural networks and evolutionary algorithms. The different models and techniques involved in the proposed approach could be used separately, since they tackle different data mining aspects that can be treated as separated problems: data pre-processing and integration, clustering, clusters validation and selection, and pathway search. We will show the integration among them for the purpose of data mining and knowledge discovery in biological data. We will present and explain each step of the proposed approach in detail, using as a case study for its application a real biological data set of *Arabidopsis thaliana*, which is the model species of current plant genomics research.

The first step involves the obtention and selection of the biological data, the kind and number of data types and sources, such as microarray experiments, metabolic profiles and pathways information, among others; the number of experiments and repetitions for each dataset, as well as the structure and type of datafiles that contain them. It also requires cleaning and artifacts elimination from data, as well as the application of appropriate selection criteria with the objective of including only sufficiently expressed data [11]. This step also needs a treatment of the

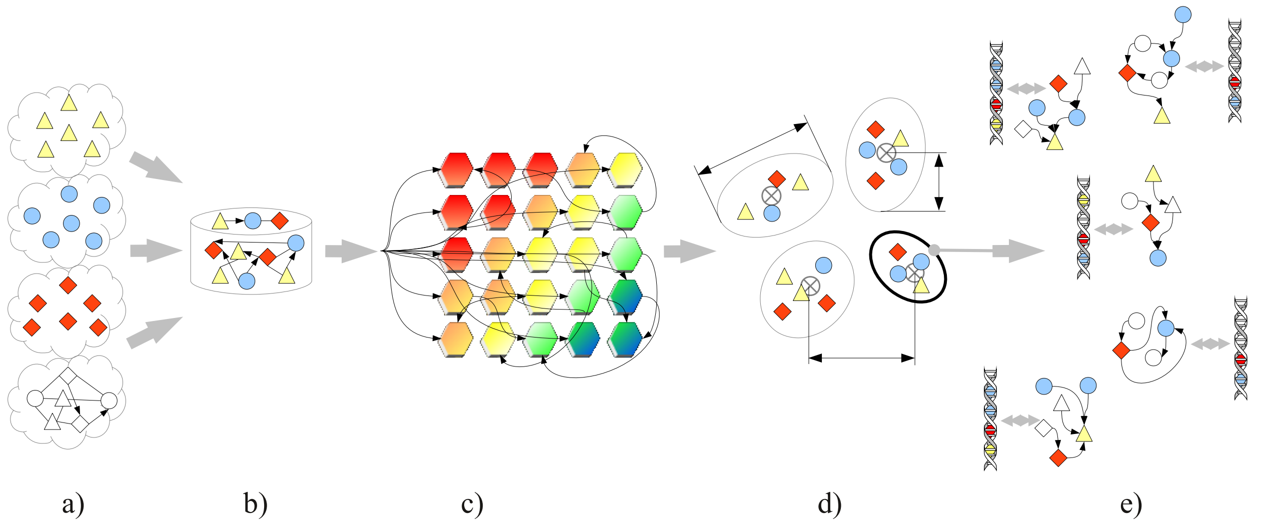

Fig. 1. Data mining of biological data as an integrated computational intelligence approach: a) different biological data sources; b) data pre-processing, normalization and integration; c) self-organizing map clustering; d) validation measures for cluster selection; e) evolutionary algorithm for metabolic pathway inference.

expression intensity values over the control sample in the case of data coming from several experimental sources [12] (Figure 1.a). The next stage requires the integration of the different data sources (Figure 1.b). For example, with an appropriate normalization, metabolome and transcriptome data obtained from the same plant material, can be integrated into a single multivariate dataset suitable for further analysis with clustering tools [13].

After the integration of heterogeneous data sources, clustering can be used for finding hidden relationships among different kinds of patterns. A software called *omeSOM [14], which implements a neural model for biological data integration, clustering and visualization through simple interfaces for the identification of coordinated variations in the data, will be shown. This visual information is then linked to the most widely used biological annotation databases, such as Arabidopsis Annotations [15] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [16]. Moreover, instead of using classical algorithms that calculate distance among patterns according to a metric such as Euclidean distance or correlation, the incorporation of a biological similarity measure into, for example, a self-organizing map (SOM) (Figure 1.c), could significantly improve the biological meaning of the clusters obtained, which are later subjected to computational analysis or scrutiny by biologists. Thus, we will describe a novel training algorithm that integrates biological similarities (derived from metabolic pathways information) into a SOM and will demonstrate that doing so, improves the quality of the clustering results. This new algorithm weighs the biological significance of the patterns during the training of the clustering method, while the clusters are being formed.

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 and select the best partition, especially from a biological perspective. In this domain, it is a common practice to validate each group returned by a clustering algorithm according to *a priori* biological knowledge (Figure 1.d). For each pattern, its annotations and memberships to well-known metabolic pathways are assessed, since they can indicate functionally related patterns. For this stage, we will show a measure that allows the comparison of clustering methods over metabolic datasets [17]. Such measure compactly summarizes the objective analysis of clustering methods: coherence and clusters distribution. Furthermore, it also evaluates the biological internal connections of such clusters considering common pathways, allowing the selection of the best clusters by effectively measuring the biological significance of each solution.

Although the clusters found reveal the presence of relations, they do not make them explicit. After the application of a clustering technique and once meaningful biological clusters are found, the identification of the relations among the data is a common problem in bioinformatics. Thus, the last step of the proposed approach is an evolutionary algorithm for the identification of novel metabolic pathways (Figure 1.e). Inside a cluster, the identification of biochemical links between its elements (genes, proteins, reactions, etc.) is not a trivial task, and it is of particular interest for the reconstruction of a metabolic network. Finding novel or non-standard metabolic pathways has important applications in metabolic engineering, metabolic network analysis and construction, as well as in the elimination of gaps in metabolic models [18]. Traditionally, this has been a manual and time-consuming process. We will present here a novel evolutionary algorithm for finding metabolic pathways, which, when given the desired beginning and target compounds, can identify pathways that link them and that are biologically meaningful.

This paper is organized as follows. Section 2 shows the use of self-organizing maps for data clustering and identification of coordinated patterns variations. Section 3 presents a new training algorithm that can include a priori biological information to obtain more biologically meaningful clusters. A validation measure that can assess the biological significance of the clusters found is explained in Section 4. Section 5 introduces an evolutionary algorithm for the inference of unknown metabolic pathways from data in clusters. Finally, the conclusions can be found in Section 6.

2 *omeSOM: transcript/metabol-ome Self Organizing Map

In this section the focus is primarily on class discovery or clustering, where data are explored from the perspective that previously unknown relations can be identified and could lead to the formulation of novel hypotheses [19]. For the analysis of biological data, clustering is implemented under the assumption that behaviorally similar samples may be related to common pathways. According to this principle, a set of genes involved in a biological process is co-expressed under the control of the same regulatory network [20].

Most of the clustering algorithms assume, at least indirectly, that the cluster structure of the data under consideration exhibits particular characteristics. For instance, HC assumes that the clusters are well separated and KM supposes that the shape of clusters is spherical [21]. When the number of samples and features involved is large, neural networks such as self-organizing maps (SOMs) [22] may be considered as a guide for an exploratory analysis of the data. Such models 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, SOM can be appropriate for cluster analysis when looking for underlying hidden patterns in data.

SOMs have been used for unsupervised clustering of transcriptome profiles increasingly over the past decade [23], [24], [25]. Recently, a method for automatically clustering SOM ensembles of high-dimensional data, such as those from whole genome microarrays, was proposed [26]. Regarding metabolites, in [27] a correlation network analysis has revealed gene-metabolite relationships in *Arabidopsis thaliana*. In [28], [29] SOM clustering is used for the analysis of *Arabidopsis thaliana* metabolome and transcriptome datasets, helping in the hypothesis validation of a metabolic mechanism responding to sulfur deficiency. In [13] a SOM model is proposed for finding relationships among introgression lines compared to a wild type control at a given developmental stage in contrast to genotype-specific data representing a time-course.

The *omeSOM software implements a neural model for biological data clustering [14]. It trains a two-dimensional SOM for the analysis and interpretation of large amounts of different types of data, such as gene expression and metabolite profiling. The software is focused on the easy identification of groups including different molecular entities, independently of the number of clusters formed. The *omeSOM software provides easy-to-visualize interfaces for the identification of coordinated variations, offering several visualization features, which are easy to understand by non-expert users. Additionally, this information is linked to the most widely used gene annotation and metabolic pathway databases. It is a software designed to give support to the data mining task on biological datasets derived from different databases.

2.1 *omeSOM main features

The *omeSOM software builds a SOM model oriented towards discovering unknown relationships among biological data, showing groups of coordinated up-regulated and down-regulated patterns in each genotype. The initial vectors are set by a principal component analysis, obtaining a learning process independent of the order of vector input, and hence reproducible. The learning method is the standard batch training algorithm [22], where the whole training set is gone through at once, and only afterwards is the map updated with the net effect of all the samples.

The *omeSOM software provides the following main options:

- *Training *omeSOM model*: creating an *omeSOM model requires an input file with comma separated values, for example *datasetname.csv*. The map size ($n \times n$ neurons) should be typed by the user in the command line. Several model topologies, map sizes, number of training epochs and initialization strategies are possible.

- *Neurons map*: several views of a trained map are possible, showing transcript (red), metabolite (blue) and both molecular entities (black) grouped into neurons. The marker size indicates the number of patterns grouped. Detailed plots of normalized and un-normalized data are shown. Additionally, in the case of genes, their corresponding Arabidopsis [15] and Solanaceae Unigene [30] annotations can be retrieved. Also, a list of KEGG metabolic pathways [16] associated with each metabolite is shown.
- *Search*: any input data point can be located on *omeSOM by name. This function returns the neuron number where a given compound has been grouped.
- *Neurons error measure*: a typical measure of clustering quality (cohesion) is calculated for each neuron and graphically shown over the feature map with different marker sizes.

The features described above constitute the fundamental functions of the software, which are constantly extended according to the users' feedback.

2.2 *omeSOM visualizations

In a standard SOM, clusters are recognized as a group of nodes rather than considering each node as a cluster. The identification of neuron clusters is mainly achieved through visualization methods such as the U-matrix [31]. It computes the average distance between the codebook vectors of adjacent nodes, yielding a landscape surface where light-colors stands for short distance (a valley) and dark-colors for larger distance (a hill). Then, the number of underlying clusters must be determined by visual inspection.

The visualizations provided by *omeSOM model, instead, present a simple interface for quick identification of co-expressed and co-accumulated genes and metabolites through a simple color code. An appropriate visualization of the resulting characteristics map, painting the neurons according to the type of data grouped, is proposed for helping the rapid identification of combined data types. The focus is on the easy identification of groups, independent of the number of neurons in a cluster. The setting of several possible visualization neighborhoods of a neuron is also helpful for the easy detection of groups of combined data types. When a visualization neighborhood is defined, all the neurons in the neighborhood radius are considered as a group and treated altogether accordingly.

An example is presented in Figure 2 for a real biological data set that comprises primary metabolites and transcripts measured in *Arabidopsis thaliana* leaves. The purpose of the integrated analysis of these data is to study the effects of cold temperatures on circadian-regulated genes in this plant [32]. The data set includes metabolites and transcripts under light-dark cycles at two control temperatures (20°C and 4°C), involved in diurnal cycle and cold-stress responses. More details on how these data were processed, filtered and normalized can be found in [32]. A total of 1549 genes and 51 metabolites were used in the integrated analysis. The plus/minus sign of each transcript/metabolite was reversed as suggested in [33], [34] to obtain items inversely correlated to each other. The

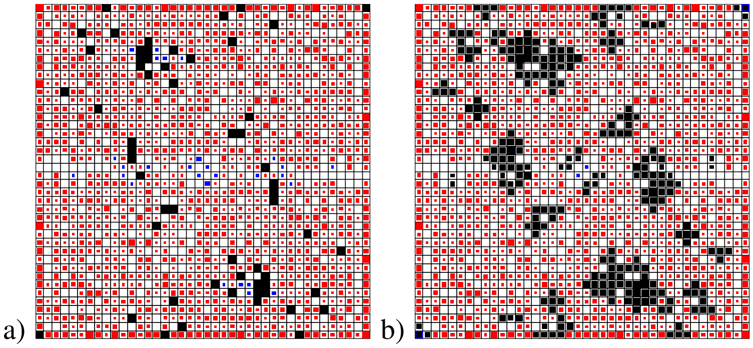

Fig. 2. *omeSOM neuron map visualizations. Color code for data type grouped into neurons: transcript (red), metabolite (blue), both (black). a) Visualization neighborhood of radius 0. b) Visualization neighborhood of radius 1.

inverted-patterns were added to the training set, resulting in a total of 3200 data points. Figure 2 shows a SOM map obtained with *omeSOM when trained with this dataset and the painting of the neurons according to the color codes mentioned before. The influence of two visualization neighborhoods on the neurons painting is shown as well.

For the special case of *omeSOM, many interesting representations of clusters can be obtained from the projection of patterns in the lattice of neurons. If the dataset includes the original data and all the data with inverted sign, the resulting map shows a symmetrical triangular configuration. This means that the top-right and bottom-left zones of the map group exactly the same data but have opposite sign. This way, it can be directly seen from the data visualisation the genes and metabolites that are up-regulated and down-regulated together or inversely (down regulated genes grouped together with up-regulated metabolites).

*omeSOM provides support for data mining tasks and it is applicable to basic research as well as applied breeding programs. This software could be used to analyze many different types of omics data. The source code and sample datasets are available free of charge at http://sourcesinc.sourceforge.net/omesom/.

3 Biological self-organizing maps

As previously introduced, systems biology clustering is implemented under the guilt-by-association principle [20], that is to say, the assumption that compounds involved in a biological process behave similarly under the control of the same regulatory networks [35]. It is presumed that if a metabolic compound with unknown function varies in a similar fashion with a known metabolite of a defined metabolic pathway, it can be inferred that the unknown element is also likely to be involved in the same pathway [36]. Therefore, those clusters that group metabolites provide evidence about the metabolic pathways associated with them. This pathway-based approach to identify metabolic traits results in more biological information or hypotheses that have to be tested through the design of biological experiments (wet experiments) to confirm the results [37].

In this context, when evaluating a clustering solution, it is a common and necessary practice to validate each group returned by a clustering algorithm through manual analysis and visual inspection, according to *a priori* biological knowledge. For each pattern, its annotations and memberships to well-known metabolic pathways are generally assessed. Traditionally, the known annotations have been used only as a second step, after data have been clustered according to their expression pattern. Only those clusters in which many data points are annotated with the same information (for example, the same biologic process), are then selected for further analysis [38], [39]. The results obtained after the examination of each cluster by hand may indicate functionally related patterns [40], [41].

Therefore, there is a growing interest in improving the clustering of biological data by incorporating prior knowledge into it, such as the Gene Ontology (GO) [42] annotations of genes, in order to improve the biological meaning of clusters that are subjected to later scrutiny [43]. In the last few years, several methods have been introduced with that aim, since integrating a biological similarity measure into a clustering method can lead to potential enhancement in the performance of the clustering [44], [45], as a result of a good correlation between biological similarity and gene co-expression [46]. In [47] the proposal is to shrink the distances between pairs of genes that share a common annotation. In fact, the similarity measure between genes can combine expression profiles and functional similarity [48][49]. Most of these proposals utilize only the annotations provided by the GO ontology or its hierarchical structure, that can be taken advantage of through the use of similarity measures between terms. However, genes that are currently unannotated are either excluded or handled as exceptional cases in those methods.

In this section, we describe a novel training algorithm that integrates biological similarities, derived from metabolic pathway information, into the procedure of obtaining clusters with a SOM model when used over biological datasets. We will demonstrate that doing so improves the quality of the clustering. This new approach named biological SOM (bSOM) weighs the biological significance of the patterns during the training of the clustering method, through the use of a new term for the biological assessment of clusters while they are being formed.

3.1 Biological-SOM training algorithm

In bSOM, we propose the use of a combination of a classical metric to measure distance between patterns and neuron centroids, plus an additional term that measures the internal biological connectivity of the patterns grouped in a cluster. Thus, when forming the clusters, the distances among patterns are computed using the weighted sum

$$d_{ij} = (1-\alpha)\,\epsilon_{ij} + \alpha(\pi_{i\notin j} - \pi_{i\in j}), \tag{1}$$

where ϵ_{ij} is a standard metric distance between a pattern i and a neuron prototype j, (for example, $\epsilon_{ij} = \|\mathbf{x}_i - \mathbf{w}_j\|_2$); $\pi_{i \notin j}$ is the average number of biological connections between all the patterns clustered in neuron j *not including* pattern i; $\pi_{i \in j}$ is the average number of biological connections between all the patterns clustered in neuron j *including* pattern i; and α is a regularization parameter that can be varied between 0 and 1 and controls the importance that is given to the biological distance during training. The biological connections are calculated as the average number of metabolic pathways in common among the patterns clustered in a neuron.

The biological term ($b_{ij} = \pi_{i \notin j} - \pi_{i \in j}$) measures how close a pattern i is to a neuron j, in terms of the improvement of the common number of known pathways in that cluster j. If the pattern has already been correctly assigned to a neuron, $b_{ij} = 0$. If a pattern has $b_{ij} > 0$ with respect to neuron j, it means that if the pattern i is assigned to neuron j, the average number of common pathways among all the data patterns clustered in that neuron would decrease. Instead, if $b_{ij} < 0$, the assignment of the pattern i to neuron j would certainly increase the number of average common pathways connections, clearly enhansing the biological value of that cluster. The parameter α is used to balance the two goals: when $\alpha = 0$, d_{ij} becomes a classical metric distance and the algorithm becomes the standard SOM clustering (sSOM); and when $\alpha = 1$ the algorithm completely disregards the expression measures and groups data only according to biological connections.

3.2 Analysis of bSOM results

This subsection presents the results obtained from the application of the new biologically-inspired training algorithm, in comparison to standard training, to the direct metabolites of the *Arabidopsis thaliana* data set presented in Section 2.2. Table 1 reports the results of the comparison of bSOM vs. sSOM training algorithms for three standard validation measures: compactness ($\overline{C}$), separation ($\overline{S}$) and the combined index of Davies-Bouldin [50], [51], [17].

For the evaluation of clusters from the viewpoint of their biological meaning, we will use the number of common pathways among patterns grouped in a cluster normalized by the number of all the possible shared pathways among patterns grouped in this cluster and any other pattern in the dataset. The average biological connectivity $\overline{P}$ is defined as the average of this relation over all the clusters. For the calculation of this biological connectivity index we used the KEGG[1] pathway database [16]. SOM was trained during 10 epochs with both training methods, weighing the Euclidean distance and the biological terms during clusters formation in bSOM using a regularization term $\alpha = 0.75$.

As shown in Table 1, improved cohesion and average separation of the clusters are achieved when using bSOM in comparison to sSOM. The DB measure does not improve when using bSOM. However, this measure is designed to qualify well-separated clusters better, and in the biological data set the average separation and clusters cohesion

[1] http://www.genome.jp/kegg/pathway.html/

TABLE 1
Validation measure comparison for sSOM and bSOM. The best value for each measure is underlined.

training →	sSOM	bSOM
$\overline{C}$	6.92	$\underline{6.81}$
$\overline{S}$	4.25	$\underline{4.19}$
DB	$\underline{8.38}$	9.55
$-\log(\overline{P})$	1.22	$\underline{0.89}$

are large. Therefore, DB does not provide good results in any case since it should be closer to zero, and this does not happen independently of the training algorithm used and data distribution. Finally, considering the average biological connectivity of the clusters found, $\overline{P}$ has clearly improved in the new proposed algorithm bSOM when compared to sSOM. It can be stated that, in general, the new algorithm has a better performance than the standard one. That is to say, it effectively improves the clusters obtained, from the viewpoint of their biological meaning, which is measured in terms of belonging to known metabolic pathways.

We have performed a detailed analysis of one of the clusters obtained after training using sSOM and bSOM. The metabolites grouped together when using sSOM were: *Valine, Aspartate, Xylose, Raffinose and Citrate*. The number of common pathways among these metabolites was only one: ko2010[2]. The equivalent cluster obtained by bSOM has grouped: *Valine, Aspartate, Xylose, Raffinose, Citrate, Glutamine and Arginine*. In this case, the pathways in common are five: ko2010, ko0970, ko1060, ko2010 and ko4974. This example shows how the patterns grouped in the cluster obtained by the biological training have significantly increased the number of pathways in common in the cluster, increasing the biological meaning of the clusters found.

4 Global measure for linked clustering

As shown in the previous section, after the application of an unsupervised mining technique, it is rather difficult to validate the results obtained. A good clustering solution should perform reasonably well under multiple measures. Although a set of objective measures can be used to quantify the quality of the clusters obtained by the different methods available, 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. As stated before, it is common practice to validate the groupings returned by a clustering algorithm through manual analysis and visual inspection, according to *a priori* biological knowledge.

Most existing validation measures [50] evaluate different aspects of a clustering solution separately, which are based only on the raw data. None of them uses explicit information from the application domain to evaluate the clusters found nor do they evaluate the differences among the solutions found from a biological point of view.

[2] KEGG pathway code.

Therefore, a way of measuring the biological relevance of the results might be useful, since unsupervised clustering may produce useless clusters.

The clustering results are evaluated by examining the relationship between the clusters produced and the known attributes (annotations) of the genes in those clusters. Existing proposals [52], however, mainly concentrate on enriching or evaluating clusters according to well-known gene functions only. 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. These aspects are important when qualifying a cluster.

A Global Measure for Linked Clustering (GMLC) [17], which compactly assesses the kind of cluster analysis often made by biologists, is explained here. In our work, the use of pathway information for assessing clusters allows the integration of both transcriptional data and metabolic profiles for a more general evaluation. 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 GMLC compactly reflects the objective analysis of clustering methods regarding coherence and cluster distribution. Moreover, it also evaluates their biological internal connections considering common pathways.

4.1 Combined biological assessment validation measure

In this subsection, the following notation is used: X is the dataset formed by $\mathbf{x}_i$ data samples; Ω 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 Ω. To define the GMLC, we need the following three factors:

1) Clustering homogeneity:

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

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

2) Grouping coherence:

$$\overline{\Gamma} = \frac{1}{|X|} \sum_i \epsilon(-\mathbf{x}_i) \left(\frac{|\Omega_{(i)}| - \mathit{\Theta}_{(-i)}}{|\Omega_{(i)}|} \right),$$

where $\Omega_{(i)}$ is the node in which pattern $\mathbf{x}_i$ is grouped; $\mathit{\Theta}_{(-i)}$ 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, $\mathit{\Theta}_{(-i)} = 0$. If not, $\mathit{\Theta}_{(-i)}$ counts how many of the original data points that were grouped together with $\mathbf{x}_i$ are not grouped now with $-\mathbf{x}_i$.

3) Internal biological connectivity:

$$\overline{P} = \frac{1}{k} \sum_{m} \frac{p_m}{p_{m*}},$$

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 ρ_{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. This measure is, conceptually, the same one used in Section 3.2 for bSOM result evaluation, but it is formally defined here.

The Global Measure for Linked Clustering is defined as a weighted combination

$$G = \gamma_H \log(\check{H}) + \; \gamma_\Gamma \log(\overline{\Gamma}) + \; \gamma_P \log(\overline{P}), \tag{2}$$

where the γ 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 the expected values

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

and

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

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

4.2 Application to clustering methods

The application of standard measures to the clusters obtained on the *Arabidopsis thaliana* data are presented in Table 2. We included the clustering methods most widely used in bioinformatics research nowadays [5], [9], [53]:

TABLE 2
Comparison of validation measures for the clustering methods. The best value of each k for each measure is underlined.

	HC			KM			SOM		
	k_1	k_2	k_3	k_1	k_2	k_3	k_1	k_2	k_3
$\overline{C}$	7.06	7.21	8.66	6.85	7.11	7.19	$\underline{6.55}$	$\underline{6.33}$	$\underline{6.29}$
$\overline{S}$	7.09	7.36	7.06	6.35	6.69	6.99	$\underline{4.58}$	$\underline{4.55}$	$\underline{5.40}$
DB	$\underline{2.67}$	$\underline{2.89}$	$\underline{3.63}$	5.12	4.93	4.52	14.9	15.7	13.0
G	10.08	10.32	9.55	17.1	21.7	19.8	$\underline{8.72}$	$\underline{8.27}$	$\underline{6.49}$

HC, KM (100 repetitions) and SOM. For all the methods, we used the Euclidean norm to measure the distance between patterns. The Gap Statistic [54] was used to select an appropriate number of clusters for the comparisons among methods to show the application and use of the GMLC. These three top gap scores were selected for the comparisons: $k_1 = 50$, $k_2 = 200$ and $k_3 = 450$. For further details on this study and additional comparisons see [17].

Additionally, we tested the significance of all the results obtained by performing 100 resamplings of 80% of the transcripts and metabolites in the dataset, for all the methods in each k. We have statistically analyzed the significance of such results using an ANOVA to test the null hypothesis in which, for each partition k, the difference among the clustering methods, in a given measure, is not significant. The analysis revealed that all the measures, including the new GMLC, show significant differences ($p < 0.01$) among the methods in each partition, validating the conclusions obtained from the detailed analysis of each of the measures in the presented table.

It can be seen from compactness and separation that SOM is the best clustering method in any of the tested configurations. Since these two types of measures evaluate opposite aspects of a clustering solution, a combined measure such as DB was calculated because it combines compactness and separation into one single index. Due to this fact, one would expect a better discerning capacity. Interestingly, DB has favored HC for all the configurations tested. However, being a measure that summarizes compactness and separation, it contradicts those indications in the table. In particular, SOM always obtains the worst 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. As patterns that are at large distance to a centroid (possible *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.

These contradictory results can be very confusing. As stated, a good clustering solution should perform well under multiple evaluation points. However, as it can be seen, it is very difficult to indicate one of the evaluated methods as the clear winner according to all the measures applied; or to indicate a solution and configuration as the most adequate. 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 explicitly select one method as the one providing more interesting clusters to be analyzed by further computational analysis or by biologists in order to discover new relationships among data. It can be observed that the GMLC is consistent with the results obtained from the application of compactness and separation measures. In fact, the GMLC could aid in deciding which clustering solution to use in order to obtain a coherent and biologically significant solution for a particular biological experiment. In future works, a fuzzy-GMLC measure could be defined using fuzzy memberships for the pathways.

Nevertheless, the analysis performed does not provide a clear clue regarding the connection or the reason why the data were clustered, from the viewpoint of their involvement in a common metabolic pathway. This point is addressed in the next section.

5 Evolutionary metabolic pathway search

Although SOM can be applied to group biochemical entities into clusters, the relationships among them remain hidden. In fact, searching metabolic pathways is a relevant task in bioinformatics, particularly when working with different types of data. In many cases, employing classical strategies for sequential state space exploration allows solutions to be found more rapidily [55]. However, it is a well-known fact that there are several problems where a very high number of solutions must be explored, making classical methods practically inapplicable [56]. In the last few years, nature inspired approaches have been proposed to tackle these problems. Among them, the ones with higher impact and relevance were evolutionary algorithms, providing outstanding results in several disciplines [57], [58], [59]. Their success is related to their ability to perform an effective and efficient global search in complex solution spaces. Some interesting aspects about them are the simplicity of the operators used, the possibility of using fitness functions with very few formal requirements, and the ability to explore multiple points of the search space in each iteration [60].

Different search strategies to find metabolic pathways that relate compounds have been recently proposed. PathComp [61] uses an algorithm based on a classical search algorithm to build paths that connect the compounds, taking them in pairs and combining them through allowed relations (metabolic reactions). Linked Metabolites [62] builds an integrated graph first and performs the pathway search by specifying the maximum number of reactions between source and target compounds. Metabolic PathFinding Tool [63] assigns to each operator a cost that is equal to the number of reactions where the compound participates. PathMiner [64] uses the A* search algorithm in wich its heuristic function employs the structural information of the compounds to generate characteristic descriptors, and explores the search space using a cost function based on the Manhattan distance. Most of these methods require detailed and specific information about the molecular structure of metabolic compounds, which in many cases is not available. The proposals based on classical methods suffer from limitations in the computational resources associated with the path length and degree of branching of the tree search. Furthermore, the order in which the nodes of the tree are visited can bias the search for particular solutions.

In literature, there are many different applications of evolutionary algorithms to bioinformatics [18], and in particular, to metabolic pathway analysis [65] and optimization [66], [67]. A genetic algorithm can be used for specific optimal metabolic network design of energetically favorable pathways for production of a determined compound of interest [68]. An optimum metabolic pathway can be found by studying a large number of alternative pathways. This could show if very efficient pathways share common structural properties, which can be used, for example, for the optimization of the flux and stoichiometry of a determined biological system [69]. An evolutionary algorithm can be also used for the determination of the kinetic parameters of a time-varying model of a metabolic system [70]. Another recent field of application is metabolic engineering within biotechnology –the targeted manipulation of cells and enhancement of a desired product. Evolutionary algorithms can be used for the understanding of how regulatory elements interact with each other, to control such processes [71]. In line with this last application field, this section will present an evolutionary algorithm to find metabolic pathways (EAMP) that relates two compounds, and compare its performance with solutions based on classical search algorithms. To achieve this the *omeSOM data mining tool was used to generate clusters from a real biological dataset, and pairs of compounds within the clusters were used for metabolic pathways search. Afterwards, objective measures are defined to quantify the performance of the algorithms.

5.1 Evolutionary algorithm for the search of metabolic pathways

The state space is defined as the set C of all metabolic compounds in the KEGG database [16]. This database contains information of genes, proteins and metabolic compounds of hundreds of different organisms, and the allowed binary relations between compounds are described by r transformations. The compound on which the transformation is applied will be called substrate s, and p will be the product or new resulting state. Transformations will be represented as ordered pairs $r_i = (s_i, p_i)$, with $s_i, p_i \in C \wedge s_i \neq p_i$, being $\widehat{s}$ and $\widehat{p}$ the initial and final compounds of the metabolic pathway. In this way, a metabolic pathway is built as a sequence of transformations that produce $\widehat{p}$ starting from $\widehat{s}$. Finally, the sequence of possible states $\mathbf{q} = [\widehat{s}, p_1, p_2, \ldots, \widehat{p}]$ is defined as the sequence of compounds that take part in the transformation. Thus, the sequence of r_i transformations leading to the production of $\widehat{p}$ from $\widehat{s}$ can be coded in a chromosome $\mathbf{c} = [r_1, r_2, \ldots, r_i, \ldots, r_N]$, where N indicates the number

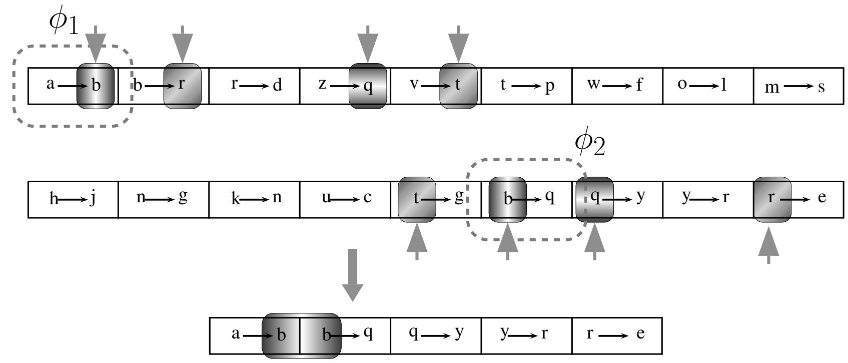

Fig. 3. Crossover operator. Each block corresponds to a gene encoding a transformation. In each gene, substrate and product are represented by letters on the left and right of each arrow, respectively. Shaded elements indicate pairs of positions (ϕ_1, ϕ_2) where a valid crossover can be made.

of genes and the sequence is read from left to right [3].

Due to the requirements of the application under study, it has been necessary to make various changes to classical genetic operators, which, if directly applied, would limit the convergence of the algorithm. In order to facilitate their explanation, four sets of transformations are defined: R^* contains the complete set of allowed transformations; $R^1 = \{r_i / r_i = (\widehat{s}, p_i)\} \wedge R^1 \subset R^*$ contains only those transformations that use $\widehat{s}$; $R^N = \{r_i / r_i = (s_i, \widehat{p})\} \wedge R^N \subset R^*$, contains all transformations that produce $\widehat{p}$; and $R^+ = R^1 \cup R^N$.

The crossover point ϕ_k for each parent $\mathbf{c}_k$ is randomly selected from a set containing pairs of positions (ϕ_1, ϕ_2) that satisfy $s_i = p_j$ for $s_i \in \mathbf{c}_1 \wedge p_j \in \mathbf{c}_2$. Figure 3 shows a diagram of this crossover operator in the case of two parents having reactions that are not completely valid. Each gene codes a chemical reaction in which letters represent the substrates and products. It can be noticed that if a simple crossover method is applied without considering the sequence of reactions, the validity of the generated offspring will probably diminish. However, if the crossover is carried out in one of the highlighted pairs of positions (ϕ_1, ϕ_2), the validity of the offspring will increase or, at least, remain constant.

The mutation operator replaces a gene with another according to

$$mut(r_i) = \begin{cases} \tau_i \in R^+ & \text{if } N = 1, \\ \tau_i \in R^1 & \text{if } N > 1 \wedge i = 1, \\ \tau_i \in R^N & \text{if } N > 1 \wedge i = N, \\ \tau_i \in R^* / \varrho_i = s_{i+1} & \text{if } N > 1 \wedge 1 < i < N \wedge u \leq 0.5, \\ \tau_i \in R^* / \varsigma_i = p_{i-1} & \text{if } N > 1 \wedge 1 < i < N \wedge u > 0.5, \end{cases} \tag{3}$$

[3] In the context of evolutionary computation, in spite of the name, chromosomes and genes are not biological entities but data structures that model a problem.

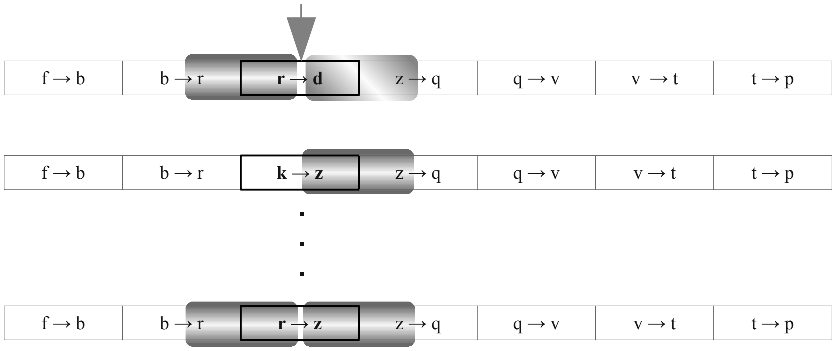

Fig. 4. Mutation operator. The selected gene to mutate is marked with a black rectangle. This mutation may lead to maintaining or increasing the validity of the chromosome.

where s_{i+1} is the substrate of the gene that is located in the position next to the mutated gene r_i, and p_{i-1} is the product of the gene that is in the position previous to the mutated gene. Value u is random with uniform distribution in [0,1]. Figure 4 presents a diagram of the mutation operator. It is observed that in the chromosome placed at the top of the figure the gene selected to mutate has a valid relation with the previous gene. If a classical mutation operator is applied, there is high probability that the validity of the chromosome will diminish. If, on the contrary, a valid mutation strategy is applied like in (3), the new chromosome increases its validity, as it can be seen in the chromosome at the bottom of the figure.

To evaluate the solutions, a fitness function was built taking into account features of biochemical reactions, and it was employed to guide the search. The fitness function for the chromosome $\mathbf{c}$ is defined as $f(\mathbf{c}) = \Delta\left[V(\mathbf{c}) + \beta E(\mathbf{c}) + Q(\mathbf{c}) + I(\mathbf{c})\right]$, where $\Delta = 1/(3+\beta)$ is a normalization constant and β determines the relative contribution of E measure. This fitness function takes value 1 when a valid and loop-free metabolic pathway that transforms $\widehat{s}$ in $\widehat{p}$ is found. In case of having information about the relative abundance of compounds, this function could be modified to weight the reactions according to the probability of occurrence. The four measurements of the fitness function are:

- Validity (V): it quantifies the number of valid concatenations present in the chromosome, defining them as those consecutive pairs of transformations where the product p_i of r_i is the substrate s_{i+1} of the transformation r_{i+1}. V varies in the range [0,1], being 1 when all operators are well concatenated.
- Valid extremes (E): this term evaluates transformations r_1 and r_N to verify they contain the desired $\widehat{s}$ and $\widehat{p}$ compounds. The calculation is done according to $E(\mathbf{c}) = \frac{1}{2}\left[\delta(\widehat{s}, s_1) + \delta(p_N, \widehat{p})\right]$, where δ is the Kronecker delta. This term varies in the range [0,1] and reaches its maximum value when compounds s_1 and p_N are the desired ones.
- Unique reactions rate (Q): it penalizes the repetition of transformations in the chromosome. The rate is

TABLE 3
Comparison between BFS and EAMP. Time $\bar{t}$ is expressed in seconds and L in number of transformations. $|\Psi|$ indicates the number of compounds in each cluster.

$\|\Psi\|$		6			12		
Ends		62 - 47	258 - 77	47 - 258	37 - 82	135 - 65	135 - 82
$\bar{t}$	EAMP	17.6	8.6	8.3	3.4	15.2	8.4
	BFS	170.4	613.5	533.2	33.5	14.2	84.7
L_M	EAMP	13	11	17	9	19	18
	BFS	6	6	6	4	6	7
$\widehat{L}$	EAMP	6.5	7	8	5	9	6
	BFS	5	6	6	4	6	6
L_m	EAMP	4	5	5	3	5	5
	BFS	4	5	5	3	5	5
ψ_M	EAMP	3	3	2	4	3	4
	BFS	3	2	2	3	2	2
$\overline{\psi}$	EAMP	2.4	2.1	2.0	2.3	2.1	2.1
	BFS	2.3	2.1	2.0	2.6	2.1	2.0

calculated as $Q(\mathbf{c}) = (\varphi(\mathbf{c}) - 1)/(N - 1)$, where φ counts the number of unique elements present in $\mathbf{c}$, and $Q(\mathbf{c}) = 0$ when $N = 1$. Q varies in the range [0,1] and reaches its minimum value when the sequence contains a unique element repeated N times.

- Unique compound rate (I): this term penalizes the repetition of compounds in the pathway. The rate is calculated as $I(\mathbf{c}) = (\varphi(\mathbf{q}) - 2)/(N - 1)$ and it is defined $I(\mathbf{c}) = 0$ when $N = 1$. I varies in the range [0,1] and reaches its minimum value when the chromosome contains transformations that lead only to s_1 or p_1. For example, for the metabolic pathway $\mathbf{c} = [a \to b], [b \to a], [a \to b], [b \to d]$ the number of reactions is $N = 4$ and the sequence of states associated to it is $\mathbf{q} = [\underline{a}, \underline{b}, a, b, \underline{d}]$ where only three compounds are unique ($\varphi(\mathbf{q}) = 3$). In consequence, $I(\mathbf{c}) = 1/3$.

5.2 Evolutionary algorithm results and discussion

The proposed EAMP has been compared with two classical search algorithms that do not require specific information from the problem: breadth-first search (BFS), the classical method most widely used by related work, and deep-first search (DFS) [55]. However, since the length of the paths found by DFS tended to be equal to the maximum allowed value (100 reactions in our experiments) and metabolic pathways containing such number of transformations are of no biological interest, the results obtained with this algorithm are not presented.

Several performance measures obtained for the search of metabolic pathways on different number of clusters were used. Table 3 shows measures obtained with EAMP and BFS for each pair of compounds searched[4]. The rows in the table correspond to: the mean search time ($\bar{t}$); the maximum, median, and minimum number of transformations (L_M, $\widehat{L}$ and L_m); and the maximum and mean number of cluster compounds incorporated into the pathway (ψ_M and $\overline{\psi}$). An analysis of this table reveals that BFS employed times which are 10 times higher than EAMP to

[4] Named using KEGG codes.

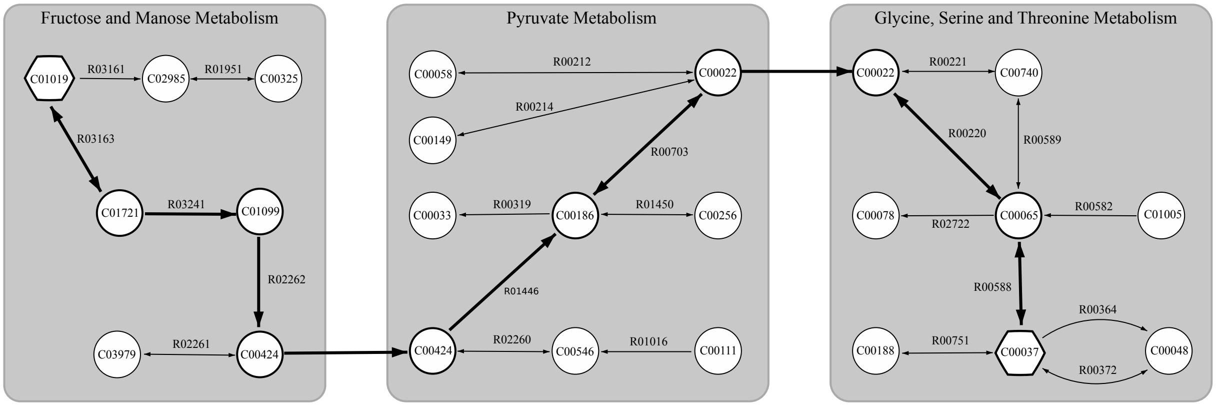

Fig. 5. Metabolic pathway found by using EAMP (indicated in bold lines) for linking compounds C01019 and C00037. Initial and final compounds are drawn as bold hexagons. Large gray rectangles indicate related known pathways and their compounds (circles) and reactions (arrows), shown with the KEGG codification.

perform every search ($p < 0.003$). With regard to the length of the paths, it can be noticed that the median for each algorithm was similar, whereas the diversity in the number of transformations was higher for the EAMP, as reflected in the maximum length for each algorithm. On the other hand, it was observed that measures related to the presence of cluster compounds in the paths were similar for both algorithms. It must be highlighted that the EAMP generated a higher dispersion of lengths in the paths, which translates into an increase in the variety of pathways found and the richness of possibilities for further analysis from a biological point of view.

As an example of a biological evaluation, a search for a pathway linking two metabolic compounds, C01019 (Fucose) and C0037 (Glycine), was performed. The linked metabolites have been clustered in neighborhood neurons in a SOM map and they do not have any common metabolic pathway inside KEGG. Figure 5 shows the obtained results. Initial and final compounds are drawn as bold hexagons. Large gray rectangles indicate, as a reference, parts of well-known pathways. Their compounds (circles) and reactions (arrows) are shown with the KEGG codification. The pathway found by the EAMP is indicated in bold lines. It presents a novel mechanism for the conversion of L-Fucose in Glycine, not reported so far in literature. This mechanism could be of interest since it corresponds to an alternative route for the production of Glycine, which is an amino acid of great importance that acts as precursor of many other metabolites [72]. All metabolites in this pathway have been reported in *Arabidopsis thaliana* [73] with the exception of the compounds C01721 (L-Fuculose) and C01099 (L-Fuculose 1-Phosphate). This might be due to the fact that they have not been determined experimentally yet. With respect to the enzymes necessary for the reactions that produce these metabolites, a gene encoding the enzyme that may catalyze the first reaction was found in *Oryza sativa* [74], while a gene encoding an enzyme associated to the third reaction was found in *Medicago truncatula* (legume) [74]. Although a gene encoding the enzyme needed for the second reaction has not been found in these plants yet, it should not be excluded that it could be found in the future. In fact, the functionality of a large variety of *Arabidopsis thaliana* genes is currently unknown [75]. The verification of inferred non-standard pathways should be performed through the corresponding biological "wet" experiments.

6 Conclusions

In this article we have presented a novel integrated computational intelligence approach for biological data mining. It is an approach encompassing several steps, which involves the application of two of the most important and well-tested techniques in the computational intelligence field: neural networks and evolutionary algorithms. Each step has been explained in detail, through a real case study involving genes from microarray measurements and metabolite profiles from *Arabidopsis thaliana*.

First of all, we have proposed the use of self-organizing maps for the integration and clustering of data from heterogeneous sources, since SOM has proved to be a useful tool for the identification of coordinated variations in data patterns. We have also presented a novel algorithm for training a SOM model, in such a way that a priori biological information could be used during clusters formation to obtain more biologically meaningful results. Several quality measures have been applied to the results, showing improved clusters formation in comparison to standard methods. Due to the need for further validation of clustering results, not only from an objective point of view but also taking into account knowledge from the application domain, a validation measure that can assess the biological significance of the clusters has been presented. It explained how such measure could aid in deciding which clustering configuration to use in order to obtain a coherent and a biologically significant solution. Finally, an evolutionary algorithm for the search of novel metabolic pathways among data grouped inside clusters was presented. Such an algorithm was able to find a biological pathway between two metabolic compounds, in spite of the fact that they have no known pathway that relates them. The sequence of reactions found by the proposed evolutionary method could provide clues for hypothesis formulation and further investigation of the biological processes involving those compounds.

Acknowledgements

The authors would like to thank Dr. Fernando Carrari, Dra. Laura Kamenetzky and Dra. Mariana Lopez, from National Institute for Agricultural Technology (INTA-ARG) for the fruitful discussion of some of the topics presented in this work. This work was supported by the National Scientific and Technical Research Council [PIP 2072008] and the National Agency for the Promotion of Science and Technology [PICT 2008 00140, PAE 37122, PICT 2009 00052].

REFERENCES

[1] R. Bino, R. Hall, O. Fiehn, J. Kopka, K. Saito, J. Draper, B. Nikolau, P. Mendes, U. Roessner-Tunali, M. Beale, R. Trethewey, B. Lange, E. Wurtele, and L. Sumner, "Potential of metabolomics as a functional genomics tool." *Trends Plant Sci*, vol. 9, no. 9, pp. 418–425, September 2004.

[2] F. Carrari, C. Baxter, B. Usadel, E. Urbanczyk-Wochniak, M.-I. Zanor, A. Nunes-Nesi, V. Nikiforova, D. Centero, A. Ratzka, M. Pauly, L. J. Sweetlove, and A. R. Fernie, "Integrated analysis of metabolite and transcript levels reveals the metabolic shifts that underlie tomato fruit development and highlight regulatory aspects of metabolic network behavior," *Plant Physiol.*, vol. 142, pp. 1380–1396, Dec. 2006.

[3] M. Bylesjo, D. Eriksson, M. Kusano, T. Moritz, and J. Trygg, "Data integration in plant biology: the o2pls method for combined modeling of transcript and metabolite data," *Plant Journal*, vol. 52, no. 6, pp. 1181–1191, 2007.

[4] 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, pp. 54–77, 2008.

[5] 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.

[6] B. Andreopoulos, A. An, X. Wang, and M. Schroeder, "A roadmap of clustering algorithms: finding a match for a biomedical application," *Briefings in Bioinformatics*, pp. 297–314, 2009.

[7] E. Keedwell and A. Narayanan, *Intelligent Bioinformatics: The Application of Artificial Intelligence Techniques to Bioinformatics Problems*. Wiley, 2005.

[8] 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.

[9] 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 Transactions on Computational Biology and Bioinformatics*, vol. 7, pp. 64–79, 2010.

[10] G. Fogel and D. Corne, *Computational Intelligence in Bioinformatics*. Morgan Kaufmann, 2008.

[11] C. J. Baxter, M. Sabar, W. P. Quick, and L. J. Sweetlove, "Comparison of changes in fruit gene expression in tomato introgression lines provides evidence of genome-wide transcriptional changes and reveals links to mapped qtls and described traits," *J. Exp. Bot.*, vol. 56, pp. 1591–1604, 2005.

[12] C. Causton, J. Quackenbush, and A. Brazma, *Microarray Gene Expression Data Analysis: A Beginner's Guide*. Blackwell Publishers, 2003.

[13] 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.

[14] 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.

[15] Arabidopsis annotations. [Online]. Available: http://www.arabidopsis.org

[16] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, pp. 27–30, 2000.

[17] G. Stegmayer, D. H. Milone, L. Kamenetzky, M. G. Lpez, and F. Carrari, "A biologically-inspired validity measure for comparison of clustering methods over metabolic datasets," *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, 2012, in press.

[18] G. Fogel, D. Corne, and Y. Pan, *Evolutionary Computation in Bioinformatics*. Wiley-IEEE Press, 2003.

[19] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science (New York, N.Y.)*, vol. 286, pp. 531–7, Oct. 1999.

[20] 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.

[21] F. Azuaje and N. Bolshakova, *Clustering Genome Expression Data: Design and Evaluation Principles*. Springer, 2002.

[22] T. Kohonen, M. R. Schroeder, and T. S. Huang, *Self-Organizing Maps*. Springer-Verlag New York, Inc., 2005.

[23] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. Lander, and T. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and applications to hematopoietic differentiation," *Proc Natl Acad Sci USA*, vol. 96, no. 1, pp. 2907–2912, 1999.

[24] J. Wang, J. Delabie, H. Aasheim, E. Smeland, and O. Myklebost, "Clustering of the som easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study," *BMC Bioinformatics*, vol. 3, no. 1, pp. 36–46, 2002.

[25] M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, and J. Mesirov, "Genepattern 2.0," *Nature Genetics*, vol. 38, no. 5, pp. 500–501, 2006.

[26] A. Newman and J. Cooper, "Autosome: A clustering method for identifying gene expression modules without prior knowledge of cluster number," *BMC Bioinformatics*, vol. 11, no. 1, p. 117, 2010.

[27] E. Allen, A. Moing, T. Ebbels, M. Maucourt, A. Tomos, D. Rolin, and M. Hooks, "Correlation network analysis reveals a sequential reorganization of metabolic and transcriptional states during germination and gene-metabolite relationships in developing seedlings of arabidopsis," *BMC Systems Biology*, vol. 4, no. 1, pp. 62–72, 2010.

[28] M. Y. Hirai, M. Yano, D. B. Goodenowe, S. Kanaya, T. Kimura, M. Awazuhara, M. Arita, T. Fujiwara, and K. Saito, "Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in arabidopsis thaliana," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 10 205–10, Jul. 2004.

[29] M. Hirai, M. Klein, Y. Fujikawa, M. Yano, D. Goodenowe, Y. Yamazaki, S. Kanaya, Y. Nakamura, M. Kitayama, H. Suzuki, N. Sakurai, D. Shibata, J. Tokuhisa, M. Reichelt, J. Gershenzon, and K. Saito, "Elucidation of gene-to-gene and metabolite-to-gene networks in arabidopsis by integration of metabolomics and transcriptomics," *J Biological Chemistry*, vol. 280, no. 27, pp. 25 590–25 595, 2005.

[30] Solanaceae unigene annotations. [Online]. Available: http://www.sgn.cornell.edu

[31] A. Ultsch, *Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series in Kohonen Maps*. Elsevier, 1999.

[32] C. Espinoza, T. Degenkolbe, C. Caldana, E. Zuther, A. Leisse, L. Willmitzer, D. Hincha, 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.

[33] 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.

[34] K. Saito, M. Y. Hirai, and K. Yonekura-Sakakibara, "Decoding genes with coexpression networks and metabolomics - majority report by precogs," *Trends in Plant Science*, vol. 13, pp. 36–43, 2008.

[35] V. Lacroix, L. Cottret, P. Thebault, and M.-F. Sagot, "An introduction to metabolic networks and their structural analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, no. 4, pp. 594–617, 2008.

[36] B. Usadel, T. Obayashi, M. Mutwil, F. Giorgi, G. Bassel, M. Tanimoto, A. Chow, D. Steinhauser, S. Persson, and N. Provart, "Co-expression tools for plant biology: opportunities for hypothesis generation and caveats," *Plant, Cell & Environment*, vol. 32, no. 12, pp. 1633–1651, 2009.

[37] 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.

[38] E. Buehler, J. Sachs, K. Shao, A. Bagchi, and L. Ungar, “The crasss plug-in for integrating annotation data with hierarchical clustering results,” *Bioinformatics*, vol. 20, no. 17, pp. 3266–3269, 2004.

[39] J. Doherty, L. Carmichael, and J. Mills, “GOurmet: a tool for quantitative comparison and visualization of gene expression profiles based on gene ontology (GO) distributions.” *BMC bioinformatics*, vol. 7, 2006.

[40] R. K. Curtis, M. Oresic, and A. Vidal-Puig, “Pathways to the analysis of microarray data,” *Trends in Biotechnology*, vol. 23, no. 8, pp. 429 – 435, 2005.

[41] P. Toronen, “Selection of informative clusters from hierarchical cluster tree with gene classes,” *BMC Bioinformatics*, vol. 5, no. 1, p. 32, 2004.

[42] M. Ashburner, “Gene ontology: tool for the unification of biology,” *Nat. Genet.*, vol. 25, no. 1, pp. 25–9, 2000.

[43] D. Dotan-Cohen, S. Kasif, and A. A. Melkman, “Seeing the forest for the trees: using the gene ontology to restructure hierarchical clustering,” *Bioinformatics*, pp. 1789–1795, 2009.

[44] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer, “Co-clustering of biological networks and gene expression data,” in *ISMB (Supplement of Bioinformatics)*, 2002, pp. 145–154.

[45] J. Cheng, M. Cline, J. Martin, D. Finkelstein, T. Awad, D. Kulp, and M. A. Siani-Rose, “A knowledge-based clustering algorithm driven by gene ontology,” *Journal of Biopharmaceutical Statistics*, vol. 14, no. 3, pp. 687–700, 2004.

[46] H. Wang, F. Azuaje, O. Bodenreider, and J. Dopazo, “Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships,” in *CIBCB ’04. Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2004, pp. 25–31.

[47] 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.

[48] N. Speer, C. Spieth, and A. Zell, “A memetic co-clustering algorithm for gene expression profiles and biological annotation,” in *In: Proc. of Congress on Evolutionary Computation (CEC)*, vol. 2, 2004, pp. 1631–8.

[49] R. Kustra and A. Zagdanski, “Data-fusion in clustering microarray data: Balancing discovery and interpretability,” *IEEE/ACM Trans. Comput. Biology Bioinform.*, vol. 7, no. 1, pp. 50–63, 2010.

[50] 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.

[51] D. Davies and D. Bouldin, “A cluster separation measure,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 1, no. 4, pp. 224–227, 1979.

[52] 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.

[53] U. S. Murty, M. S. Rao, K. Sriram, and K. M. Rao, “Applications of self-organising map som for prioritisation of endemic zones of filariasis in andhra pradesh, india,” *International Journal of Data Mining and Bioinformatics*, vol. 5, no. 4, pp. 417–427, 2011.

[54] 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.

[55] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach (3 Ed.)*. Prentice Hall, 2009.

[56] S. Sivanandam, *Introduction to Genetic Algorithms*. Springer, 2008.

[57] C.-K. Ting, W.-M. Zeng, and T.-C. Lin, “Linkage discovery through data mining [research frontier],” *IEEE Computational Intelligence Magazine*, vol. 5, pp. 10–13, 2010.

[58] O. C. S. Damas and J. Santamaria, "Medical image registration using evolutionary computation: An experimental survey," *IEEE Computational Intelligence Magazine*, vol. 6, pp. 26–42, 2011.

[59] J. Zhang and et al., "Evolutionary computation meets machine learning: A survey," *IEEE Computational Intelligence Magazine*, vol. 6, pp. 68–75, 2011.

[60] G. Fogel, "Computational intelligence approaches for pattern discovery in biological systems," *Briefings in Bioinformatics*, vol. 9, no. 4, pp. 307–316, 2008.

[61] H. Ogata, S. Goto, W. Fujibuchi, and M. Kanehisa, "Computation with the KEGG pathway database," *BioSystems*, vol. 47, pp. 119–128, 1998.

[62] J. Easton, L. Harris, M. Viant, A. Peet, and T. Arvanitis, "Linked metabolites: A tool for the construction of directed metabolic graphs," *Computers in Biology and Medicine*, vol. 40, pp. 340–349, 2010.

[63] D. Croes, F. Couche, S. Wodak, and J. van Helden, "Metabolic Pathfinding: inferring relevant pathways in biochemical networks," *Nucleic Acids Research*, vol. 33, pp. W326–W330, 2005.

[64] D. McShan, S. Rao, and I. Shah, "PathMiner: predicting metabolic pathways by heuristic search," *Bioinformatics*, vol. 19, pp. 1692–1698, 2003.

[65] J. Blazeck and H. Alper, "Systems metabolic engineering: Genome-scale models and beyond," *Biotechnology Journal*, vol. 5, pp. 647–659, 2010.

[66] O. Ebenho and R. Heinrich, "Evolutionary optimization of metabolic pathways. theoretical reconstruction of the stoichiometry of ATP and NADH producing systems," *Bulletin of Mathematical Biology*, vol. 63, pp. 21–55, 2001.

[67] D. Na, T. Kim, and S. Lee, "Construction and optimization of synthetic pathways in metabolic engineering," *Current Opinion in Microbiology*, vol. 13, pp. 363–370, 2010.

[68] B. A. Boghigian, H. Shi, K. Lee, and B. A. Pfeifer, "Utilizing elementary mode analysis, pathway thermodynamics, and a genetic algorithm for metabolic flux determination and optimal metabolic network design," *BMC Systems Biology*, vol. 4, pp. 49–66, 2010.

[69] A. Stephani, J. C. Nuez, and R. Heinrich, "Optimal stoichiometric designs of atp-producing systems as determined by an evolutionary algorithm," *Journal of Theoretical Biology*, vol. 199, pp. 45–61, 1999.

[70] T. Geyer, X. Mol, S. Blass, and V. Helms, "Bridging the gap: Linking molecular simulations and systemic descriptions of cellular compartments," *PLOS One*, vol. 5, p. 14070, 2010.

[71] J. Yang, S. Wongsa, V. Kadirkamanathan, S. A. Billings, and P. C. Wright, "Metabolic flux estimation - a self-adaptive evolutionary algorithm with singular value decomposition," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, pp. 126–138, 2007.

[72] R. Li, M. Moore, and J. King, "Investigating the regulation of one-carbon metabolism in arabidopsis thaliana," *Plant Cell Physiology*, vol. 44, pp. 233–241, 2003.

[73] L. A. Mueller, P. Zhang, and S. Y. Rhee, "Aracyc: A biochemical pathway database for Arabidopsis," *Plant Physiology*, vol. 132, pp. 453–460, 2003.

[74] C. Liang and et al., "Gramene: a growing plant comparative genomics resource," *Nucleic Acids Research*, vol. 36, pp. 947–953, 2008.

[75] H. Lan, R. Carson, N. J. Provart, and A. J. Bonner, "Combining classifiers to predict gene function in Arabidopsis Thaliana using large-scale gene expression measurements," *BMC Bioinformatics*, vol. 8, p. 358, 2007.