sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone, G. Stegmayer, L. Kamenetzky, M. López, J. Giovannoni, Je Min Lee & 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. 4, pp. 438-448, 2010.

*omeSOM: a software for integration, clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants

Diego Milone[1], Georgina Stegmayer*[2], Laura Kamenetzky[3], Mariana López[3], James J. Giovannoni[4], Fernando Carrari[3]

[1]Sinc(i), FICH-UNL, CONICET, Argentina
[2]CIDISI, UTN-FRSF, CONICET, Argentina
[3]INTA-Castelar, CONICET, Argentina
[4]Boyce Thompson Institute for Plant Research, Cornell University, USA

Email: Diego Milone - d.milone@ieee.org; Georgina Stegmayer - gstegmayer@santafe-conicet.gov.ar; Laura Kamenetzky - lkamenetzky@cnia.inta.gov.ar; Mariana López - mglopez@cnia.inta.gov.ar; James J. Giovannoni - jjg33@cornell.edu; Fernando Carrari - fcarrari@cnia.inta.gov.ar;

*Corresponding author

Abstract

Background: Modern biology uses experimental systems that involve the exploration of phenotypic variation as a result of the recombination of several genomes. Such systems encompass millions of recombination events between the alleles of domesticated and wild species and are useful to investigate the functional evolution of metabolic networks. This kind of studies involve the generation of a large amount of data, which require dedicated computational tools for their analysis.

Results: This paper presents a novel software named *omeSOM (transcript/metabol-ome SOM) that implements a neural clustering model for biological data integration. This model allows the discovery of relationships between changes in transcripts and metabolites of crop plants harboring introgressed exotic alleles. The model is focused on the easy identification of groups including different molecular entities, independently of the number of clusters formed. The *omeSOM software provides interfaces that are easy to visualize for the identification of coordinated variations in the co-expressed and co-accumulated genes and metabolites, respectively. Additionally, this information is linked to the most used gene annotation and metabolic pathway databases.

Conclusions: *omeSOM is a software tool designed to give support to the data mining task of metabolic and transcriptional datasets derived from different databases. It provides a user-friendly interface and offers several visualization tools easy to understand by non-expert users. Therefore, *omeSOM can be proposed as a software tool designed to give support to the data mining task applied to basic research as well as breeding programs. Application including an example dataset is available free of charge at
http://sourceforge.net/projects/sourcesinc/files/omesom.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone, G. Stegmayer, L. Kamenetzky, M. López, J. Giovannoni, Je Min Lee & 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. 4, pp. 438-448, 2010.

Background

Nowadays, the biology field is in the middle of a data explosion. A series of technical advances in the last years have led to an increase in the amount of data that biologists can record about different aspects of an organism, both at genomic and post-genomic levels. Bioinformatics is playing an important role at these levels, allowing biologists to make full use of the advances in computer science for analyzing these kinds of data. The discipline has evolved mainly from the development of data mining techniques and their application to automatic prediction and discovery of classes [1]. The prediction of classes uses the information available on the expression profiles and the known features of the sets of data or experiments to build classifiers for further data. However, in this work we focus on class discovery, where data are explored from the viewpoint of the existence of unknown relations that could lead to the formulation of novel explaining hypotheses [2]. Two distinct types of class discovery methods exist: supervised ones, which are guided by a few hypotheses to be tested; and unsupervised ones, where no target variable is identified a priori and the mining algorithm searches for structures among all variables. The most common unsupervised data mining method is clustering [3]. Clustering refers to the grouping of observations or samples into classes of similar objects (named clusters) [4]. These algorithms segment the entire dataset trying to maximize the similarity of the samples within a cluster, minimizing their similarity to outside records at the same time [5]. Among the current proposals in the area, artificial intelligence tools have been compiled by Keedwell and Narayanan [6], and, in particular, artificial neural networks have been recently stressed [7, 8].

Discovery of hidden patterns of gene expression in plants of economic importance to agro-biotechnology

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone, G. Stegmayer, L. Kamenetzky, M. López, J. Giovannoni, Je Min Lee & 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. 4, pp. 438-448, 2010.

may aid improving the quality of crop products. In addition, transcript and metabolite integration is also gaining attention given the need for extracting knowledge from signals of different sources, with the aim of finding hidden relations to infer new knowledge about the biological processes underlying them [9, 10].

In plant experimental biology and crop breeding, two of the most used systems are introgression lines and recombinant inbred lines (ILs and RILs), defined as genotypes carrying exotic alleles from related species. Although the use of ILs and RILs have constituted useful tools in crop domestication and breeding since time immemorial, their applicability as experimental systems exposing thousands of quantitative trait loci has become popular only in the past decade [11, 12, 13]. The effects on gene expression and metabolite accumulation in each line may provide important clues regarding the metabolic pathways that involve the introgressed segments [14]. A recent advance in this field has reported probabilistic associations and visualizations of genes, metabolites and phenotypes for this kind of datasets [15].

For the analysis of these biological data, clustering is implemented under the assumption that behaviorally similar samples could share common pathways. According to this principle, named ”guilt-by-association”, a set of genes involved in a biological process is co-expressed under the control of the same regulatory network [16]. One of the assumptions is that if an unknown gene is co-expressed with a known one from a homogeneous biological source, the first one is probably involved in the same metabolic pathway (for a review see [17]). A similar reasoning can be applied to metabolites [18] and the integration of both types of data [19].

In this paper, we present the *omeSOM tool, which trains a two-dimensional self-organizing map (SOM), as an alternative for the analysis and interpretation of large amounts of data from different nature such as gene expression and metabolite profiling [20]. The raw dataset used in this work consisted of profiles of ripen tomato fruits harvested from a population of introgression lines derived from a cross between the tomato domesticated species *Solanum lycopersicum* and the wild species *Solanum pennellii* [21]. This work adds a new analytical dimension providing a specialized tool for grouping and searching new relationships between metabolites and transcripts.

The paper is structured as follows: first, implementation and software features are described and then, a discussion of the *omeSOM clustering is presented. After that, the visualization tools available, together with a final discussion in a biological context, are presented.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone, G. Stegmayer, L. Kamenetzky, M. López, J. Giovannoni, Je Min Lee & 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. 4, pp. 438-448, 2010.

Implementation and software features

The *omeSOM tool has been implemented in the Matlab®/Octave [22] programming language. We used a standard toolbox for SOM training, provided by the original developers of this neural network model [23]. The software packages and documentation can be downloaded from the project home page http://sourceforge.net/projects/sourcesinc/files/omesom/.

The *omeSOM software provides the following main options and visualizations:

- **Create *omeSOM model**: creating an *omeSOM model requires an input file with the *.data* extension, for example *datasetname.data* (a detailed explanation of the required format file is given below). The map size should be typed by the user in the command line.
- **Search**: any input data point can be located on the *omeSOM. This function returns the neuron number where a given metabolite name/transcript code has been grouped.
- **Neurons map**: several views of a trained map are possible, showing transcript (red), metabolite (blue) and both molecular entities (black) grouped into neurons. Detailed plots of normalized and un-normalized data are shown. Additionally, in the case of transcripts, their corresponding Arabidopsis [24] and Solanaceae Unigene [25] annotations can be retrieved. Also, a list of metabolic pathways [26] associated with each metabolite is shown.
- **3-colors map**: a specific view of the map is shown, painting the neurons according to a color scale that easily indicates those grouping transcripts and metabolites which are 1 standard deviation out of the neuron mean.
- **Neurons error measure**: a typical measure of clustering quality (cohesion) is calculated for each neuron and shown graphically over the feature map with different marker sizes.
- **Neurons having pseudo-zeros**: there are special situations where some metabolite may show undetectable levels in a specific genotype, having however valid measurements for many others.

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

Results and Discussion

The case study used to test the *omeSOM software applicability involves the analysis of fruit transcriptional and metabolite profiles data from a set of tomato ILs derived from a cross between *Solanum*

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone, G. Stegmayer, L. Kamenetzky, M. López, J. Giovannoni, Je Min Lee & 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. 4, pp. 438-448, 2010.

lycopersicum and its wild relative *Solanum pennelli*. An exemplary dataset can be downloaded from http://sourceforge.net/projects/sourcesinc/files/omesom/data.

IL-dataset input file

Table 1 shows an example of an input dataset appropriate for *omeSOM. The input matrix must have the following format: a first row with the number of genotypes studied; a second one may have a comment line enclosing the name of each genotype. From the third row on, each line must have the measurements (x) for each IL of a single molecule (m for metabolite, t for transcript).

Each measurement is an average log value ($logR_i^*$), where * stands for the metabolite or transcript at the genotype i, calculated from the relative measurements of the compounds studied for valid experiments, where there are measurements for at least two technical replicates. The resulting log ratios are normalized. For each pattern, the sum of the square of log ratios is set equal to 1 according to

$$x_i^* = \frac{logR_i^*}{\sum_{j=1}^{P}(logR_j^*)^2} \qquad (1)$$

where P is the total number of genotypes studied.

Several data integrations are possible. For example, before integration of two datasets, the plus/minus sign of one dataset can be reversed to obtain negatively correlated items. To find all possible relations[1, 2], the training set should include the original and the inverted sign versions of all the data samples.

For the case study, the metabolic data were obtained analyzing polar extracts of tomato fruits, through Gas Chromatography coupled to Mass Spectrometry (GC-MS). The peak intensities were normalized to the internal standards added and to the mass of the tissue sample processed [27]. The metabolite profiling technique used allows the identification of approximately 80 primary metabolic compounds [10]. Transcriptional levels were obtained from TOM2 chips (long oligo arrays representing approximately 12,000 tomato unigenes) ordered into spots, previously marked by hybridization with two fluorescence probes. The tomato gene expression database used contains annotation and sequence information of oligo array. A total of $P = 21$ ILs were analyzed, with introgressions in chromosomes: 1, 2, 3, 5, 8, 10, 11 and 12. After pre-processing and selection steps, $M = 71$ metabolites and $T = 1385$ transcripts reached the threshold value to be considered valid data [28].

[1] 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 $\downarrow m \leftrightarrow \downarrow m$).

[2] 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 $\downarrow m \leftrightarrow \uparrow m$).

Table 1: Input training set containing measurements for T transcripts and M metabolites from P ILs. Original and inverted versions of all the data samples are included in the example.

P						
IL$_1$	IL$_2$	$\dots$	IL$_i$	$\dots$	IL$_P$	
$x_1^{t_1}$	$x_2^{t_1}$	$\dots$	$x_i^{t_1}$	$\dots$	$x_P^{t_1}$	Transcript1
$x_1^{t_2}$	$x_2^{t_2}$	$\dots$	$x_i^{t_2}$	$\dots$	$x_P^{t_2}$	Transcript2
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\dots$
x_1^T	x_2^T	$\dots$	x_i^T	$\dots$	x_P^T	TranscriptT
$-x_1^{t_1}$	$-x_2^{t_1}$	$\dots$	$-x_i^{t_1}$	$\dots$	$-x_P^{t_1}$	Transcript1(inv)
$-x_1^{t_2}$	$-x_2^{t_2}$	$\dots$	$-x_i^{t_2}$	$\dots$	$-x_P^{t_2}$	Transcript2(inv)
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\dots$
$-x_1^T$	$-x_2^T$	$\dots$	$-x_i^T$	$\dots$	$-x_P^T$	TranscriptT(inv)
$x_1^{m_1}$	$x_2^{m_1}$	$\dots$	$x_i^{m_1}$	$\dots$	$x_P^{m_1}$	Metabolite1
$x_1^{m_2}$	$x_2^{m_2}$	$\dots$	$x_i^{m_2}$	$\dots$	$x_P^{m_2}$	Metabolite2
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\dots$
x_1^M	x_2^M	$\dots$	x_i^M	$\dots$	x_P^M	MetaboliteM
$-x_1^{m_1}$	$-x_2^{m_1}$	$\dots$	$-x_i^{m_1}$	$\dots$	$-x_P^{m_1}$	Metabolite1(inv)
$-x_1^{m_2}$	$-x_2^{m_2}$	$\dots$	$-x_i^{m_2}$	$\dots$	$-x_P^{m_2}$	Metabolite2(inv)
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\dots$
$-x_1^M$	$-x_2^M$	$\dots$	$-x_i^M$	$\dots$	$-x_P^M$	MetaboliteM(inv)

*omeSOM Clustering

Neural network-based clustering is closely related to the concept of competitive learning, which is based on the idea of units (neurons) that compete to respond to a given subset of inputs. The nodes in the input layer admit input patterns and are fully connected to the output nodes in the competitive layer. Each output node corresponds to a cluster and is associated with a prototype or weight vector. Given an input pattern, its distance to the weight vectors is computed and only the neuron closest to the input becomes activated. The weight vector of this winning neuron is further moved towards the input pattern. This competitive learning paradigm is also known as winner-takes-all learning [29].

Self-organizing maps (SOMs) represent a special class of neural networks that use competitive learning. Their aim is to represent complex high-dimensional input patterns into a simple low-dimensional discrete map, with neurons that can be visualized in a two-dimensional lattice structure, while preserving the proximity relationships of the original data as much as possible [30]. Therefore, SOMs can be appropriate for cluster analysis when looking for underlying hidden patterns in data. A neighborhood function is defined for each neuron and when competition among the neurons is complete, SOMs update a set of

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone, G. Stegmayer, L. Kamenetzky, M. López, J. Giovannoni, Je Min Lee & 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. 4, pp. 438-448, 2010.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone, G. Stegmayer, L. Kamenetzky, M. López, J. Giovannoni, Je Min Lee & 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. 4, pp. 438-448, 2010.

weight vectors within the neighborhood of the winning neuron.

The *omeSOM tool builds a SOM model oriented towards discovering unknown relationships among transcriptional and metabolite data, showing previously unknown clusters of coordinated up-regulated and down-regulated patterns in each tomato genotype. Several model topologies, map sizes and initialization strategies are possible. The initial vectors are set by principal component analysis, obtaining a learning process independent of the order of input of vectors, and hence reproducible. The model learning method is the batch training algorithm [30], where the whole training set is gone through at once and only after this the map is updated with the net effect of all the samples. Comparison between each pattern $\mathbf{x}^*$ and each neuron weight vector $\mathbf{w}_j$ is measured through the standard Euclidean distance $d(\mathbf{x}^*, \mathbf{w}_j) = \|\mathbf{x}^* - \mathbf{w}_j\|_2$. We use a gaussian neighborhood function of the form $g_{ij} = e^{-\frac{\delta_{ij}^2}{2r^2}}$, where δ_{ij} is the distance between neuron i and neuron j on the map grid and r is the neighborhood radius.

*omeSOM Visualizations

An appropriate visualization of the resulting characteristics map, painting the neurons according to the type of data grouped, is proposed for helping in the rapid identification of combined data types. The setting of several possible visualization neighborhoods (Vn) of a neuron is also helpful for the easy detection of groups of combined data types, avoiding the need for an identification procedure.

For the special case of the *omeSOM, many interesting representations of clusters can be obtained from the projection of the 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 down-left zones of the map group exactly the same data but having opposite sign. This allows seeing, at once, direct (both genes and metabolites are up-regulated or down-regulated) and inverted (down-regulated genes grouped together with up-regulated metabolites) relations among data. There is a specific zone in the map where the exactly opposite behavior for each IL can be found, which is useful for associating specific genes/metabolites to specific genotype.

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

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone, G. Stegmayer, L. Kamenetzky, M. López, J. Giovannoni, Je Min Lee & 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. 4, pp. 438-448, 2010.

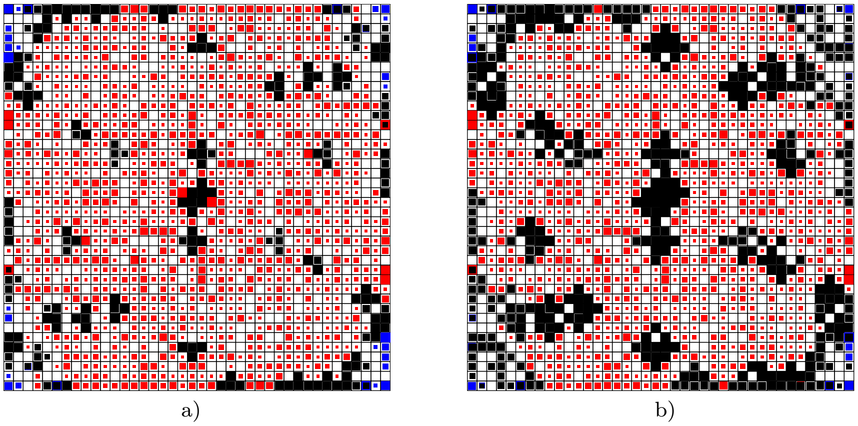


Figure 1: Activation SOM resulting from the integrated analysis of 1385 genes and 71 metabolites from 21 tomato ILs. Map topology of 40x40 neurons with a) $Vn = 0$ and b) $Vn = 1$.

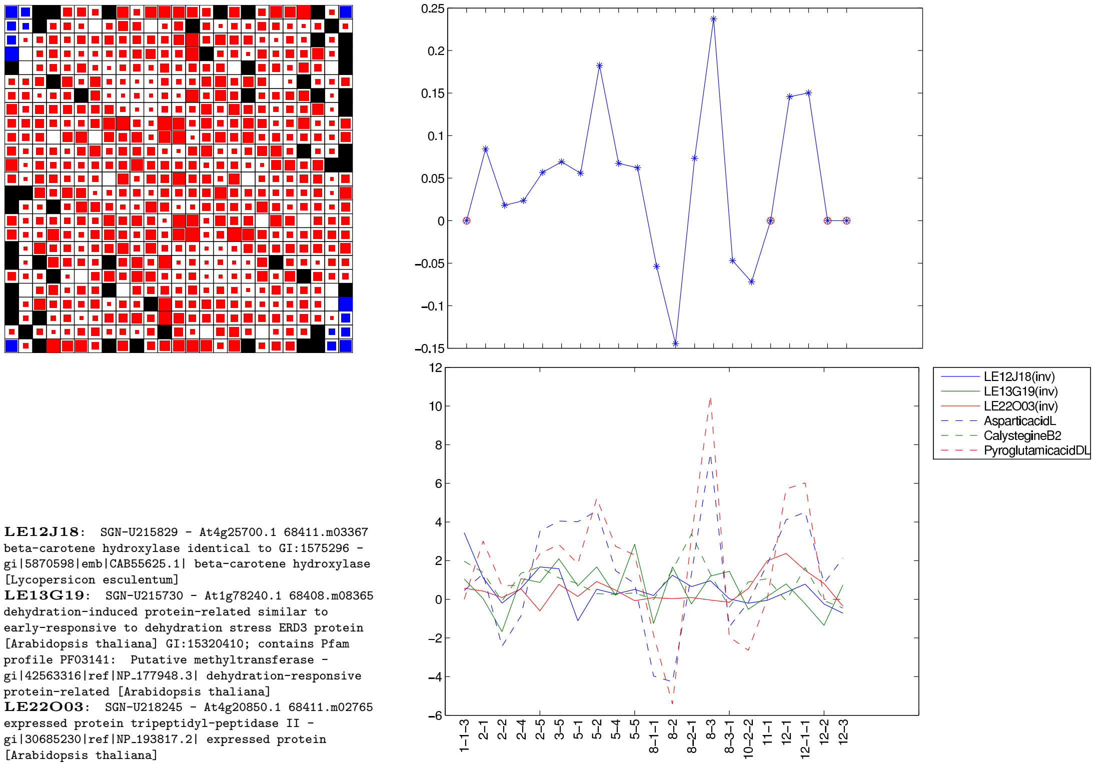
The visualizations provided by the proposed *omeSOM model, instead, provide a simple interface for helping in the rapid identification of co-expressed genes and co-accumulated metabolites through a simple color code. The focus is on the easy identification of groups of different patterns, independently of the number of neurons in a cluster. Furthermore, setting of several possible visualization neighborhoods (Vn) of a given neuron is also helpful for the easy detection of groups of combined data types, avoiding the need for the identification of neuron clusters. When a Vn is defined, all the neighboring neurons (according to the neighborhood radius set) are considered as a group and treated altogether accordingly, also for counting whether there are metabolites and transcripts grouped.

The following visualizations are supported by *omeSOM:

Easy identification of clusters of combined data types: Figure 1 shows different marker colors which indicate the kind of pattern grouped in the neuron: black for combined data types, blue for metabolites and red for transcripts grouped alone. Also the marker size indicates the relative number of patterns grouped. The figure shows the activation map resulting from the integrated analysis of 21 tomato ILs with a 40x40 neuron topology, with with $Vn = 0$ and $Vn = 1$.

Detail view of original data measurements: Figure 2 shows the resulting *omeSOM integrated model for the 21 IL dataset (a), with the same color code as that shown in Figure 1. Curves presented in b show a detail of the normalized patterns which have all been clustered together in neuron 604: three metabolites: *L-aspartic acid*, *calystegineB2* and *D/L-pyroglutamic acid*, together with three inverted sign transcripts: *LE12J18*, *LE13G19* and *LE22O03*. Figure 2 c shows the non-normalized (original)

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone, G. Stegmayer, L. Kamenetzky, M. López, J. Giovannoni, Je Min Lee & 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. 4, pp. 438-448, 2010.



LE12J18: SGN-U215829 - At4g25700.1 68411.m03367 beta-carotene hydroxylase identical to GI:1575296 - gi|5870598|emb|CAB55625.1| beta-carotene hydroxylase [Lycopersicon esculentum]
LE13G19: SGN-U215730 - At1g78240.1 68408.m08365 dehydration-induced protein-related similar to early-responsive to dehydration stress ERD3 protein [Arabidopsis thaliana] GI:15320410; contains Pfam profile PF03141: Putative methyltransferase - gi|42563316|ref|NP_177948.3| dehydration-responsive protein-related [Arabidopsis thaliana]
LE22O03: SGN-U218245 - At4g20850.1 68411.m02765 expressed protein tripeptidyl-peptidase II - gi|30685230|ref|NP_193817.2| expressed protein [Arabidopsis thaliana]

Figure 2: Integration model visualizations. Top left figure: the resulting SOM integrated model for the 21 ILs dataset, having 25x25 neurons with $Vn = 0$. Bottom right: detail of the normalized cluster patterns values clustered together in neuron 604. Top right: detail of the de-normalized (original) values for the metabolite Pyroglutamic Acid (5-oxoproline)). Bottom left: probes codes decodification.

log ratios of *pyroglutamic acid*. Red circles indicate missing values for metabolites (samples in an IL not having enough significantly detected replicate experiments for the average log ratio calculation). In the case of a transcript, red circles indicate non-significant expression levels with respect to the control genotype. The down-left panel shows the annotation of one transcript according to its probe code which is automatically linked into the Arabidopsis (At) and Unigene (SGN-U) annotations and metabolite pathways.

KEGG pathways associated with grouped compounds: Data grouped by neurons are checked against metabolic pathways available online, such as the Kyoto Encyclopedia of Genes and Genomes [32], for finding candidate genes belonging to metabolic pathways. For each metabolite, a list of KEGG pathways where it participates can be easily visualized in the same interface.

Visualization of clusters inside a specific chromosome segment: Another possibility is the visualization of clusters from all ILs belonging to the same chromosome. This allows the comparison

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone, G. Stegmayer, L. Kamenetzky, M. López, J. Giovannoni, Je Min Lee & 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. 4, pp. 438-448, 2010.

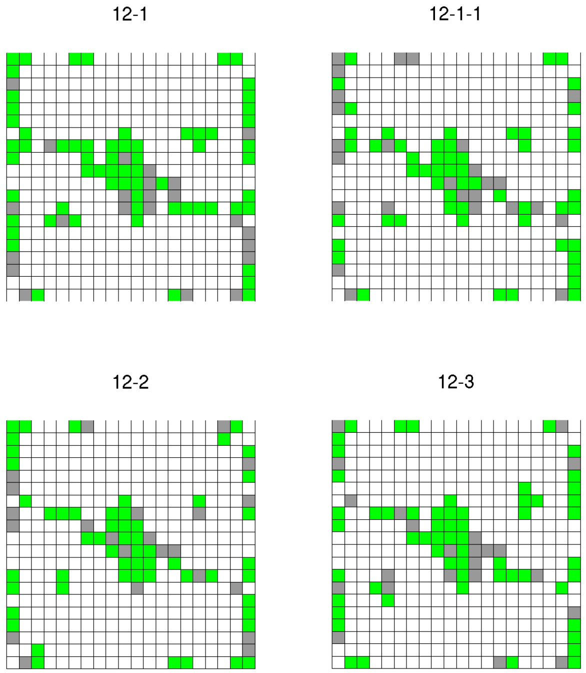


Figure 3: SOM activation map for the tomato chromosome 12, introgression lines 12-1-1, 12-1, 12-2 and 12-3.

of pattern expressions according to a color scale that paints only neurons having patterns with an important deviation from the neuron mean, for each dimension/IL. That is, neurons where at least one pattern has a value greater than the mean plus one standard deviation in the corresponding IL is depicted in green. If in this IL there is at least one pattern in the neuron with a value lower than the mean plus one standard deviation, the neuron is painted in gray. The variations in the expression levels of the grouped patterns may provide useful information regarding genes/metabolites specifically associated with a certain introgressed chromosomal segment. This tool could help in the association of metabolite and transcript networks with genetic maps. Figure 3 presents the output of the 3-colors map function which shows the activation of a 20x20 map for all ILs comprising chromosome 12. The comparison of these maps allows identifying those ILs showing distinctive neurons. This feature might facilitate mapping those genetic factors involved in the clusters.

Quality evaluation of clusters of combined data types: It is quite important to be able to evaluate the quality of a clustering algorithm when applied to biological data, in particular if later biological inferences should be made. Inside *omeSOM, a typical clustering measure is calculated for each neuron and shown graphically over the feature map with different marker sizes, when the feature Neurons error measure is selected. This measure comprises validation measures assessing cluster compactness or homogeneity [33]. Intracluster variance is their most popular representative:

$$\overline{C}_j = \frac{1}{|\Omega_j|} \sum_{\forall \mathbf{x}_i \in \Omega_j} \|\mathbf{x}_i - \mathbf{w}_j\|_2, \qquad (2)$$

where $|\Omega_j|$ is the number of patterns in node j. As a global measure of compactness, the average over all nodes is calculated $\overline{C} = \frac{1}{k}\sum_j \overline{C}_j$. Values of $\overline{C}$ close to 0 indicate more compact nodes.

Conclusions

The *omeSOM model is oriented towards discovering unknown relationships between data, as well as providing simple visualizations for the identification of co-expressed genes and co-accumulated metabolites. It has a user-friendly interface and provides several visualization tools easy to understand by non-expert users. A case study which involved gene expression measurements and metabolite profiles from tomato fruits was here presented to show the application of the proposed tool. The interest in comparing the cultivated tomato against the different ILs lies on the fact that, as it has been proven, some wild tomato relatives can be sources of several agronomical characters which could be used for the improvement of commercial tomato lines. Therefore, *omeSOM can be proposed as a software tool designed to give support to the data mining task applied to both basic research and breeding programs.

Availability and requirements

- Project name: *omeSOM.
- Project home page: **http://sourceforge.net/projects/sourcesinc/files/omesom/**.
- Operating system(s): Microsoft Windows and Linux.
- Programming language: Matlab/Octave.
- Other requirements: SOM toolbox.
- License: opensource, free for academic use.

Authors contributions

DM and GS implemented the *omeSOM graphical user interface, the clustering algorithm and the clustering measurement, and wrote the manuscript. JG provided the transcript case study dataset. LK,

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone, G. Stegmayer, L. Kamenetzky, M. López, J. Giovannoni, Je Min Lee & 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. 4, pp. 438-448, 2010.

ML and FC provided the metabolite case study dataset, tested the software and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by CONICET (PIP 1122008), ANPCyT (PICT 2008 00100, PICT 2009 00140, PICT 2009 00152) and INTA (PE 243542)

References

1. Polanski A, Kimmel M: *Bioinformatics*. Springer-Verlag, NY 2007.
2. Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, Lander E: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science* 1999, **286**:531–537.
3. Olson D, Delen D: *Advanced Data Mining*. Springer 2008.
4. Xu R, II DW: *Clustering*. Wiley and IEEE Press 2009.
5. Larose D: *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley-Interscience 2005.
6. Keedwell E, Narayanan A: *Intelligent Bioinformatics: The Application of Artificial Intelligence Techniques to Bioinformatics Problems*. Wiley 2005.
7. Kelemen A, Abraham A, Chen Y: *Computational Intelligence in Bioinformatics*. Springer 2008.
8. Tasoulis D, Plagianakos V, Vrahatis M: *Computational Intelligence in Bioinformatics*, *Volume 94 of* Studies in Computational Intelligence. Springer 2008.
9. Bino R, Hall R, Fiehn O, Kopka J, Saito K, Draper J, Nikolau B, Mendes P, Roessner-Tunali U, Beale M, Trethewey R, Lange B, Wurtele E, Sumner L: **Potential of metabolomics as a functional genomics tool**. *Trends in Plant Science* 2004, **9**:418–425.
10. Carrari F, Baxter C, Usadel B, Urbanczyk-Wochniak E, Zanor M, Nunes-Nesi A, Nikiforova V, Centero D, Ratzka A, Pauly M, Sweetlove L, Fernie A: **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 Physiology* 2006, **142**:1380–1396.
11. Li Z, Fu B, Gao Y, Xu J, Ali J, Lafitte H, Jiang Y, Rey JD, Vijayakumar C, Maghirang R, Zheng T, Zhu L: **Genome-wide Introgression Lines and their Use in Genetic and Molecular Dissection of Complex Phenotypes in Rice (Oryza sativa L.)**. *Plant Molecular Biology* 2005, **59**:33–52.
12. Rieseberg L, Wendel J: *Introgression and its consequences in plants*, *Volume 1*. Oxford University Press 1993.
13. Lippman Z, Semel Y, Zamir D: **An integrated view of quantitative trait variation using tomato interspecific introgression lines**. *Current Opinion in Genetics and Development* 2007, **17**:1–8.
14. Jingyuan JF, Joost JK, Bouwmeester H, America T, Francel WV, Jane LW, Michael HB, de Vos Ric C, Dijkstra M, Richard AS, Johannes F, Koornneef M, Vreugdenhil D, Breitling R, Ritsert CJ: **System-wide molecular evidence for phenotypic buffering in Arabidopsis**. *Nature genetics* 2009, **41**(2):166–167.
15. Joung J, Corbett A, Fellman S, Tieman D, Klee H, Giovannoni J, Fei Z: **Plant MetGenMAP: an integrative analysis system for plant systems biology**. *Plant Physiology* 2009, **151**:1758–1768.
16. Wolfe C, Kohane I, Butte A: **Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks**. *BMC Bioinformatics* 2005, **6**:227–237.
17. Usadel B, Obayashi T, Mutwil M, Giorgi F, Bassel G, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart N: **Co-expression tools for plant biology: opportunities for hypothesis generation and caveats**. *Plant, Cell & Environment* 2009, **32**(12):1633–1651.
18. Kaever A, Lingner T, Feussner K, Gobel C, Feussner I, Meinicke P: **MarVis: a tool for clustering and visualization of metabolic biomarkers**. *BMC Bioinformatics* 2009, **10**:92–100.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone, G. Stegmayer, L. Kamenetzky, M. López, J. Giovannoni, Je Min Lee & 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. 4, pp. 438-448, 2010.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone, G. Stegmayer, L. Kamenetzky, M. López, J. Giovannoni, Je Min Lee & 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. 4, pp. 438-448, 2010.

19. Junker B, Klukas C, Schreiber F: **VANTED: A system for advanced data analysis and visualization in the context of biological networks**. *BMC Bioinformatics* 2006, **7**:109–121.

20. Stegmayer G, Milone D, Kamenetzky L, Lopez M, Carrari F: **Neural Network Model for Integration and Visualization of Introgressed Genome and Metabolite Data**. In *IEEE International Joint Conference on Neural Networks*, *Volume 1* 2009:2983–2989.

21. Eshed Y, Zamir D: **An introgression line population of Lycopersicon pennellii in the cultivated tomato enables the identification and fine mapping of yield associated QTL**. *Genetics* 1995, **141**:1147–1162.

22. **GNU Octave** [http://www.gnu.org/software/octave/].

23. **SOM Toolbox** [http://www.cis.hut.fi/projects/somtoolbox/].

24. **Arabidopsis annotations** [http://www.arabidopsis.org].

25. **Solanaceae Unigene annotations** [http://www.sgn.cornell.edu].

26. **KEGG: Kyoto Encyclopedia of Genes and Genomes** [http://www.genome.jp/kegg/].

27. Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie A: **Gas chromatography mass spectrometry-based metabolite profiling in plants**. *Nature Protocols* 2006, **1**:387–396.

28. Milone D, Stegmayer G, Gerard M, Kamenetzky L, Lopez M, Carrari F: *Analysis and integration of biological data: a data mining approach using neural networks*, IGI Global 2010 . in press.

29. Haykin S: *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, Inc. 2007.

30. Kohonen T, Schroeder M, Huang T: *Self-Organizing Maps*. Springer-Verlag New York, Inc. 2005.

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

32. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Research* 2000, **28**:27–30.

33. Handl J, Knowles J, Kell D: **Computational cluster validation in post-genomic data analysis**. *Bioinformatics* 2005, **21**(15):3201–3212.