Neural Network Model for Integration and Visualization of Introgressed Genome and Metabolite Data

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

Abstract— **The volume of information derived from post-genomic technologies is rapidly increasing. Due to the amount of data involved, novel computational models are needed for introducing order into the massive data sets produced by these new technologies. Data integration is also gaining increasing attention for merging signals in order to discover unknown pathways. These topics require the development of adequate soft computing tools. This work proposes a neural network model for discovering relationships between gene expression and metabolite profiles of introgressed lines. It also provides a simple visualization interface for identification of coordinated variations in mRNA and metabolites. This may be useful when the focus is on the easily identification of groups of different patterns, independently of the number of formed clusters. This kind of analysis may help for the inference of a-priori unknown metabolic pathways involving the grouped data. The model has been used on a case study involving data from tomato fruits.**

I. Introduction

PROCESSING and discovery of relationships in the vast amount of data to be analyzed in certain bioinformatics areas represent major challenges today. The discovery of hidden patterns of gene expression in microarray data and metabolite profiles from plants of economic importance to agro-biotechnology, is a current challenge because the use of any algorithm for pattern recognition suffers from the so-called curse of dimensionality. Moreover, the volume of information from genomic experiments is increasing at a high speed and due to the amount and nature of the biological data involved (such as noisy and missing data) novel computational models are needed for properly analyzing the data sets produced. Data integration is also gaining attention for merging signals from different sources and of different nature. Moreover, visualization of results is an important issue for understanding and interpreting the hidden relationships in data [1].

In response to the need of analyzing large amounts of biological data, mainly statistical methods have been early adapted to bioinformatics for feature reduction such as t-test statistical analysis and linear discriminant analysis (LDA). Principal components analysis (PCA) has also been applied but its main drawback is that it creates a new feature space which can be difficult to interpret biologically [2]. An alternative to feature reduction is feature selection in which, given an original dataset, a subset of features is chosen. This selection is oriented to keep the more representative features, those that allow building a classifier in a dimension tractable by more conventional models. Many methods have been proposed in the literature for this task, including the construction of a criteria ranking for ordering and a search algorithm for selecting the best features [3].

Bioinformatics has evolved over time, mainly from the development of data mining techniques and their application to automatic prediction and discovery of classes, two key tasks for the analysis and interpretation of gene expression data from microarray experiments [2]. The prediction of classes uses the available information on the expression profiles and the known characteristics of the datasets or experiments to build classifiers for future data. On the contrary, in the case of classes discovery, data are explored from the viewpoint of the existence or not of unknown relations and a hypothesis to explain them is proposed [4]. Among class discovery techniques, the hierarchical clustering algorithm is a deterministic method based on a pairwise distance matrix. This algorithm establishes small groups of genes/conditions that have a common expression pattern and then construct a dendrogram, sequentially, on the basis of the distances between feature vectors. Clusters are obtained by pruning the tree at some level, and the number of clusters is controlled by deciding at which level of the hierarchy of the tree the splitting is performed [1]. Regarding non-hierarchical algorithms, the distances are calculated from a predetermined number of clusters and the genes are iteratively placed in different groups until minimizing each cluster internal spread. The more representative algorithm of this type is k-means [5].

Regardless the wide availability of information provided by metabolic profiles and microarrays studies, the knowledge extraction required to study them is not a trivial task. One of the current trends is the integration of two types of biological data: metabolic profiles and transcriptional data from microarrays, with the purpose of finding hidden correlations among them and to infer new knowledge about the biological processes under study [6]. For example, a problem of interest is how to be able to evaluate the presence of genes associated with regulatory mechanisms in metabolic pathways. This is especially important in plants due to the disponibility of primary and secondary metabolites and the wide variety of genes associated with these pathways. In particular the integration of transcriptome and metabolome data from plants, correlating gene transcription profiles with variations profiles of a large number of non-protein molecules, can be used for identifying silent phenotypes changes [7].

This allows having a snapshot of the metabolic pathways

G. Stegmayer is with CIDISI, CONICET (email: gstegmayer@santafe-conicet.gov.ar). D. Milone is with sinc(i), FICH-UNL, CONICET (email: d.milone@ieee.org). L. Kameneztky, M. López and F. Carrari are with GEFEIA, INTA-IB, CONICET (email: {lkamenetzky,mlopez,fcarrari}@cnia.inta.gov.ar).

from the changes in transcription profiles and the simultaneous analysis of metabolites and their variation in response to a given condition. A metabolic network can be formally defined as a collection of objects and the relationships between them. The objects can be chemical compounds (metabolites), biochemical reactions, enzymes (proteins) and genes. The identification of links between genes, proteins and reactions is not a trivial task, and is of particular interest for the reconstruction of a metabolic network, which could be involved in obtaining a final product with certain desired characteristics [8].

The organization of the paper is the following. Section II presents related work and outlines the main features of the proposed model. Section III explains the data preprocessing task. In Section IV, the proposed neural network model for data integration and visualization is described. Section V shows the results and their discussion. Finally, in Section VI, conclussions and future work are drawn.

II. Related work and Proposed Model Outline

These new challenges that have arisen in bioinformatics highlight the need for the development of new techniques to overcome the limitations of existing ones [2]. Among the current proposals, soft computing tools have been mentioned recently [9], in particular artificial neural networks [10], [1]. Specifically within these kind of models, self-organizing maps (SOM) [11], [12] have proven to be adequate for handling large data volume and projecting them in low dimensional maps while showing, at the same time, hidden relationships.

In [13] a SOM model is proposed for the integrated analysis of Arabidopsis thaliana metabolome and transcriptome datasets. A related work [14] shows that the clustering performance of SOM helped in the elucidation of a metabolic mechanism responding to sulfur deficiency. The results showed that functionally related genes were clustered in the same or neighbor neurons. The examination of each cluster "by hand" helped in the deduction of putative functions of genes involved in glucosinolate biosynthesis. However, the experiments and the model were specifically set for following the evolution of a previously-established condition (sulfur and nitrogen defficiency) over time, and therefore it was used for hypotheses validation rather than knowledge discovery.

In many cases, however, the biological experiment does not involve time evolution of a particular condition, but the interest focuses on the study of the differences among several plant genomes. It may involve an original genome that has been modified by introgression of wild species alleles (cisgenic plants) or transgenic plants overexpressing a gene of interest. An introgression line (IL) is defined as a genotype that carries genetic material derived from a similar species, for example a "wild" relative. The use of introgression lines allows the study and creation of new varieties by introducing exotic traits and constitute a useful tool in crop domestication (and breeding) [15], [16]. The experiments presented in this work involve the analysis of metabolite and transcriptional profiles data from tomato IL carrying allelic substitutions from a wild species. Figure 1 shows a scheme of the chromosomes of this plant population.

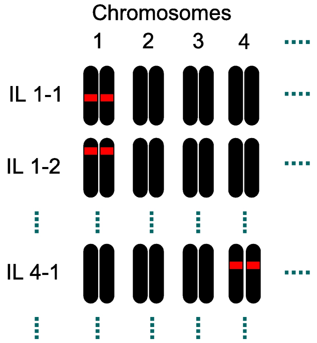


Fig. 1. Introgression lines (ILs): the portion of introgressed genome are marked in each IL with a red line.

Differently from the mentioned approaches, in our work we propose a SOM model for finding relationships among ILs compared to a wild type control (IL-SOM), instead of data evolving over time. Furthermore, the proposed model is oriented towards discovering new and unknown relationships among transcriptional and metabolic data, instead of verifying an a-priori condition. It also provides a simple visualization interface for the identification of co-expressed and co-accumulated genes and metabolites. The focus is on the easily identification of groups of different kind of patterns, independently from the number of formed clusters. This kind of analysis may be useful for later inference of unknown metabolic pathways involving the grouped data.

III. Data Preprocessing

The following subsections describe the preprocessing, filtering and selection steps [17] that have been applied to each type of data patterns before integration, with the objective of including only sufficiently expressed data in the analysis [18].

A. Metabolite Data

Metabolite accumulation measurements are obtained from fruits harvested at mature stages in field experiments and subjected to gas chromatography-mass spectrometry analyses. Metabolites that do not appear in at least two repetitions are not considered for further analysis. For each metabolite in each IL, the log ratio of the mean of the valid replicates is calculated according to

$$logR_i^m = \log_{10}\left(\frac{\dfrac{1}{N_i^m}\displaystyle\sum_{r=1}^{N_i^m} S_{ir}^m}{\dfrac{1}{\mathit{\Gamma}_i^m}\displaystyle\sum_{r=1}^{\mathit{\Gamma}_i^m} C_{ir}^m}\right) \qquad (1)$$

where S_{ir}^m is the accumulation for the metabolite m, in the replicate r at IL i, C_{ir}^m is the accumulation for the corresponding control measurement, N_i^m is number of valid IL measurement and $\mathit{\Gamma}_i^m$ is number of valid control replicates.

In the selection step only metabolites with $|logR_i^m| > \rho$ are kept for data integration and cluster analysis.

B. Transcriptional Data

Transcriptional levels are obtained from hybridization chips having spotted arrayed probes representing different genes. Poor quality spots, negative spots, spots not expressed in both channels and empty spots were filtered out. Not expressed spots are detected for IL and control slides according to

$$\bar{F}^t < \bar{B}^t + \alpha \tilde{B}^t \qquad (2)$$

where $\bar{F}^t$ is the foreground signal mean for the transcript t, $\bar{B}^t$ is the spot mean background, $\tilde{B}^t$ is the spot background standard deviation and α is a quality parameter to be empirically set.

Spots with at least two replicated data points are included for analysis using

$$logR_{ir}^t = \log_2\left(\frac{\check{S}_{ir}^t}{\check{C}_{ir}^t}\right) \qquad (3)$$

where $\check{S}_{ir}^t$ is the foreground signal median for the transcript t, in the replicate r at IL i, and $\check{C}_{ir}^t$ is the foreground signal median for the transcript t, in the corresponding control replicate r for the same IL i.

These measures are normalized using the print-tip Lowess normalization strategy [17] and the valid replicates were averaged simply as

$$logR_i^t = \frac{1}{N_i^t}\sum_{r=1}^{N_i^t} logR_{ir}^t \qquad (4)$$

where N_i^t is number of valid replicates for the transcript t at the IL i.

C. Combined Data Normalization

The resulting log ratios from the preprocessing and selection steps are normalized and combined before feeding the SOM model. For each pattern, the sum of the square of log ratios are set equal to 1 according to

$$R_i^* = \frac{logR_i^*}{\sum_{j=1}^{P}(logR_j^*)^2} \qquad (5)$$

where $*$ stands for m or t and P is the total number of available ILs.

All normalized data are arranged in the training set as shown in Table I. To find all possible inverted correlations, the training set includes the original and the inverted versions of the patterns. Then, each column/dimension-IL is normalized in the range $[0, 1]$ acording to a histogram equalization.

TABLE I
ARRANGEMENT OF THE PATTERNS IN THE TRAINING SET.

	IL$_1$	IL$_2$	...	IL$_i$	...	IL$_P$
Transcripts	$R_1^{t_1}$	$R_2^{t_1}$	...	$R_i^{t_1}$	...	$R_P^{t_1}$
	$R_1^{t_2}$	$R_2^{t_2}$	...	$R_i^{t_2}$	...	$R_P^{t_2}$
	⋮	⋮	⋱	⋮	⋮	⋮
	R_1^t	R_2^t	...	R_i^t	...	R_P^t
	⋮	⋮	⋱	⋮	⋮	⋮
	R_1^T	R_2^T	...	R_i^T	...	R_P^T
Inverted	$-R_1^{t_1}$	$-R_2^{t_1}$	...	$-R_i^{t_1}$	...	$-R_P^{t_1}$
transcripts	$-R_1^{t_2}$	$-R_2^{t_2}$	...	$-R_i^{t_2}$	...	$-R_P^{t_2}$
	⋮	⋮	⋱	⋮	⋮	⋮
	$-R_1^t$	$-R_2^t$	...	$-R_i^t$	...	$-R_P^t$
	⋮	⋮	⋱	⋮	⋮	⋮
	$-R_1^T$	$-R_2^T$	...	$-R_i^T$	...	$-R_P^T$
Metabolites	$R_1^{m_1}$	$R_2^{m_1}$	...	$R_i^{m_1}$	...	$R_P^{m_1}$
	$R_1^{m_2}$	$R_2^{m_2}$	...	$R_i^{m_2}$	...	$R_P^{m_2}$
	⋮	⋮	⋱	⋮	⋮	⋮
	R_1^m	R_2^m	...	R_i^m	...	R_P^m
	⋮	⋮	⋱	⋮	⋮	⋮
	R_1^M	R_2^M	...	R_i^M	...	R_P^M
Inverted	$-R_1^{m_1}$	$-R_2^{m_1}$	...	$-R_i^{m_1}$	...	$-R_P^{m_1}$
metabolites	$-R_1^{m_2}$	$-R_2^{m_2}$	...	$-R_i^{m_2}$	...	$-R_P^{m_2}$
	⋮	⋮	⋱	⋮	⋮	⋮
	$-R_1^t$	$-R_2^t$	...	$-R_i^t$	...	$-R_P^t$
	⋮	⋮	⋱	⋮	⋮	⋮
	$-R_1^M$	$-R_2^M$	...	$-R_i^M$	...	$-R_P^M$

IV. MODEL FOR DATA INTEGRATION AND VISUALIZATION

For the analysis of biological data, clustering is implemented under the assumption that behaviorally similar genes could share common pathways. According to this principle, named "guilt-by-association" [19], a set of genes involved in a biological process are co-expressed under the control of the same regulation network. This way, if an unknown gene is co-expressed with known genes in a biological process, this unknown gene is probably involved in the same metabolic pathway. Similar reasoning can be applied to metabolites.

The proposed SOM in this work is based on the idea that such a model can make tractable the problem of computational analysis and interpretation of large amounts of data from different nature, such as gene expression and metabolic profiles, for finding relationships among introgressed lines instead of data evolving over time. This allows discovery of previously unknown relationships among those transcripts and metabolites, which could lead to the inference of metabolic networks involving them.

Several model topologies and initialization strategies are possible. We have used different map sizes. For the map shape we have used a rectangular sheet with rectangular lattice. The initial vectors are set by principal component analysis. As a consequence, the result of the learning process becomes independent of the input vectors order, and hence it is reproducible. That is to say, given a dataset, the same map representation is always obtained for it.

The model learning method is the batch training algorithm [12], 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[1]. Comparison between each pattern $\mathbf{R}^*$ and each neuron weight vector $\mathbf{w}_j$ is measured through the standard euclidean distance

$$d(\mathbf{R}^*, \mathbf{w}_j) = \|\mathbf{R}^* - \mathbf{w}_j\|_2 \tag{6}$$

The updating is done by simply replacing the prototype vector (or winning neuron) with a weighted average over the samples, where the weighting factors are the neighborhood function values. We have used a gaussian neighborhood function of the form

$$g_{ij} = e^{-\frac{\delta_{ij}^2}{2r^2}} \tag{7}$$

where δ_{ij} is the distance between neuron i and neuron j on the map grid and r is the neighborhood radius.

The proposed model is oriented towards discovering unknown relationships among transcriptional and metabolite data, showing previously unknown clusters of coordinated up-regulated and down-regulated patterns in each IL. Furthermore, the dataset may include the original data plus the original data with inverted sign. For example, one gene similarly expressed in several ILs but up-regulated in a specific line, will be down-regulated in it if its sign is inverted. Thus, the resulting map shows a simmetrical "triangular" configuration, that is, the top-right and down-left zones of the map group exactly the same data but having opposite sign. The inclusion of original and inverted patterns allows seeing, at once, direct (up-regulated genes and metabolites, down-regulated genes and metabolites) and inverted (down-regulated genes grouped together with up-regulated metabolites) relations among data. These kind of analysis may be of help for the further inference of a-priori unknown metabolic pathways involving the grouped data.

V. Results and Discussion

A. Case Study

The case of study for the proposed model involves a dataset containing metabolites and transcripts from tomato (*Solanum lycopersicum*) which posses, at certain chromosomes segments, introgressions lines of a wild tomato species (*Solanum pennelli*). The interest in comparing the domesticated tomato variety against the different ILs lies on the fact that it has been proven that wild tomato germoplasm are valuable sources of several agronomical characters of interest which could be used for the improvement of commercial tomato lines.

The metabolic data has been obtained from the analysis of extracts of the plant tissue of interest, through Gas Chromatography coupled with Mass Spectometry (GC-Tof-MS). The peak intensities are normalized to the quantity of material used. Metabolite identification and quantification were performed as previously described [7]. For this data there were 4 replicates and metabolites having less than 2 valid replicates were removed ($2 \leq N_i^m, \Gamma_i^m \leq 4, \forall i, m$). We used a $\rho = 0.1$ for filtering according to the accumulation levels.

Transcriptional levels were obtained from hybridization chips having all the genes of the material of interest (ca. 8,000 genes) ordered into spots. The tomato gene expression database used contains annotation and sequence information of all probes on the tomato oligo array by microarray hybridization mRNA expression techniques for ≈ 13000 spots comprising the above mentioned 8,000 genes. For this data there were 8 replicates and genes not having at least one IL with at least 2 valid replicates have been removed ($2 \leq N_i^t \leq 8$). We used the quality parameter $\alpha = 2.0$.

After data preprocesing, $M = 37$ metabolites and $T = 668$ genes were selected as sufficiently expressed. The following $P = 21$ ILs have been analyzed: 1-1-3, 2-1, 2-2, 2-4, 2-5, 3-5, 5-1, 5-2, 5-4, 5-5, 8-1-1, 8-2-1, 8-2, 8-3-1, 8-3, 10-2-2, 11-1, 12-1-1, 12-1, 12-2 and 12-3 [18].

B. IL-SOM Integrated Visualization and Analysis

An appropriate visualization of the resulting characteristics map, painting the neurons according to the type of data grouped in them, is proposed for helping the quickly identification of combined data types. Thus, we define a visualization neighborhood for the evaluation of combined pattern types (metabolite/transcript) associated to a neuron. Differently from the radius r in the standard SOM neighborhood, the visualization neighborhood has a radius Λ. The setting of several possible radius in the visualization neighborhoods of a neuron is helpful for cluster identification. Instead of defining small maps (5x5 or 10x10 neurons), bigger maps (20x20 or 40x40) with $\Lambda > 0$ allow the dynamic analysis of clusters formation without re-training the SOM.

Figure 2 shows different marker colors which indicate the kind of pattern grouped in the neuron: black for combined data types, blue for only metabolites and red for only transcripts. Also the marker size indicates the number of patterns grouped. The figure shows the activation map resulting from the integrated analysis of 21 tomato ILs with a 40x40 neuron topology and $\Lambda = 1$. A detailed examination of the neurons marked in black in the top-left corner of the resulting map, indicating mixed data types grouping, shows that the transcript LE13L21 groups together with the metabolites Citrate, Fructose, Glucose-16-anhydro-beta, Glucose and Xylose, which are all involved in primary metabolic pathways. Therefore, this analysis detected a coordinated pattern of changes of these metabolites and highlighted a gene as a candidate to participate in this pathway.

In fact, in most cluster analyses, groups are not known a priori, and the interest is focused on finding them without the help of a response variable, like in [20]. Also, visualization of results may be difficult when the number of objects to

[1] www.cis.hut.fi/projects/somtoolbox/

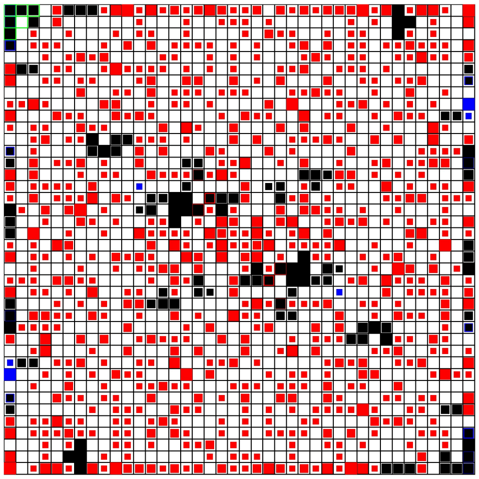
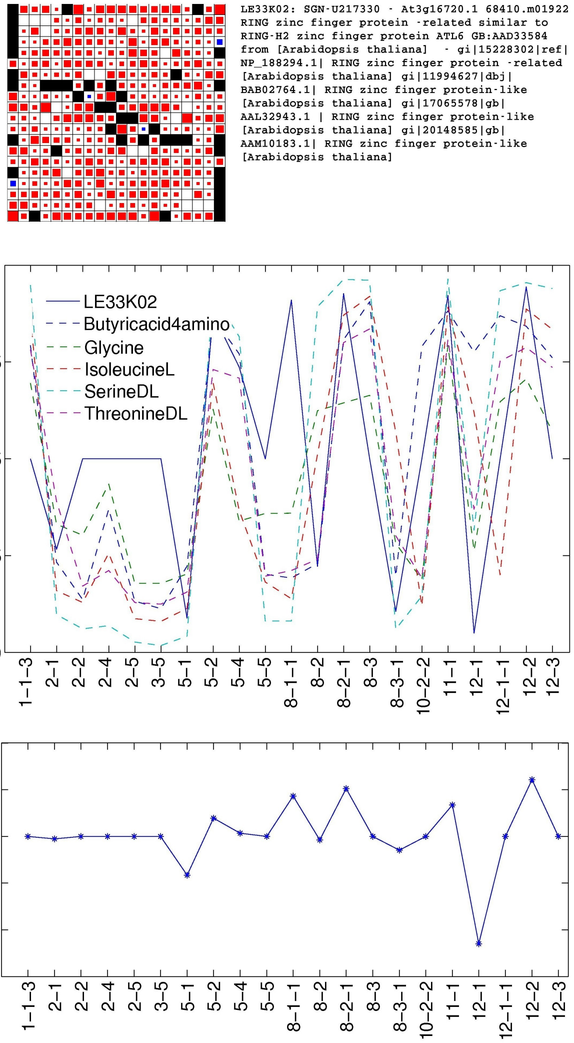


Fig. 2. Activation IL-SOM map resulting from the integrated analysis of 668 genes and 37 metabolites from 21 tomato ILs. Map topology of 40x40 neurons with $\Lambda = 1$.

group is large. In a SOM map, 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 [21]. It computes the average distance between the codebook vectors of adjacent nodes on the map and displays this value at the top of each node position, 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 visualization here proposed, instead, provides a simple visualization interface for the identification of co-expressed and co-accumulated genes and metabolites through a simple color code and the use of visualization neighborhoods. The focus is on the easily identification of groups of different types of patterns, independently from the number of neurons in a cluster. The setting of several possible Λ radius avoids the need for a cluster identification procedure because with the proposed color code, groups of combined data types are easily detected and marked.

Figure 3 shows some visualizations provided by the soft computing tool that has been designed in this work. The figure shows, in the upper left, the resulting SOM integrated model for the 21 ILs dataset (map with 20x20 neurons). The curves presented in the middle part of the figure shows a detail of the normalized patterns which have all been clustered together in the neuron in row 20 and column 5: the transcript LE33K02, and the metabolites Butyric acid 4 amino, Glycine, IsoleucineL, Serine DL and Threonine DL. For this transcript the last down plot shows its denormalized (original) log ratios. The upper right part shows a decodification of the LE33K02 transcript according to its probe code, which has been automatically translated into its corresponding Arabidopsis (At.3g16720.1) and Unigene



```
LE33K02: SGN-U217330 - At3g16720.1 68410.m01922
RING zinc finger protein -related similar to
RING-H2 zinc finger protein ATL6 GB:AAD33584
from [Arabidopsis thaliana]  - gi|15228302|ref|
NP_188294.1| RING zinc finger protein -related
[Arabidopsis thaliana] gi|11994627|dbj|
BAB02764.1| RING zinc finger protein-like
[Arabidopsis thaliana] gi|17065578|gb|
AAL32943.1 | RING zinc finger protein-like
[Arabidopsis thaliana] gi|20148585|gb|
AAM10183.1| RING zinc finger protein-like
[Arabidopsis thaliana]
```

Fig. 3. Integration model visualizations. Top left figure: the resulting IL-SOM integrated model for the 21 ILs dataset, having 20x20 neurons with $\Lambda = 0$. Middle plot: detail of the normalized cluster patterns values clustered together. Down plot: detail of the de-normalized (original) values for the transcript LE33K02 clustered in the neuron. Top right: probes codes decodification.

(SGN-U217330) annotations[2].

Another possibility is the visualization of clusters inside a specific chromosome, for all the included ILs in it. This allows the comparison of patterns expressions according to a color scale that paints only neurons having patterns with a value that deviate from the neuron mean, for each dimension/IL. That is, the 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 grey. The variations in the expression levels of the grouped patterns may provide useful information

[2] http://www.sgn.cornell.edu/

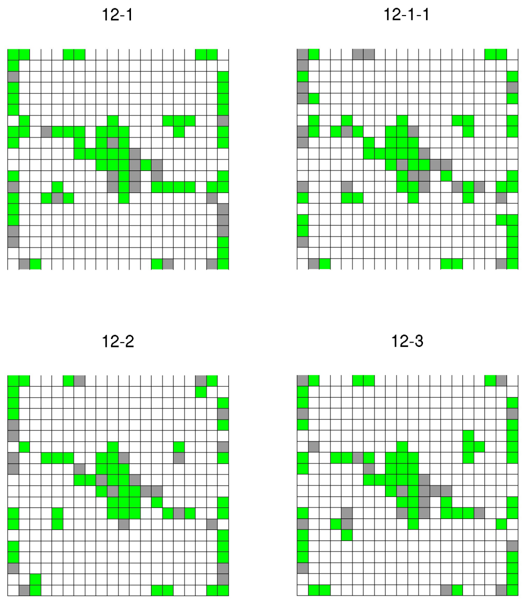


Fig. 4. IL-SOM activation map for the tomato chromosome 12: 12-1-1, 12-1, 12-2 and 12-3 ILs.

regarding genes/metabolites specifically associated to certain mechanisms in each particular IL. The visualization of these patterns outliers (or special patterns), IL against IL, may highlight interest characteristics possed by a specific IL that may differentiate it from the other ones.

Figure 4 shows the activations per IL of a 20x20 map for the tomato chromosome 12: 12-1, 12-1-1, 12-2 and 12-3 ILs, with $\Lambda = 1$.For example, in the patterns inside neuron (1,1), the Citric acid metabolite can be found. A detailed analysis of the grouped pattern shows that the Citric acid is the only responsible for the neuron being painted green in the IL 12-1 (while its values on the other ILs remain inside the average). This is an important clue regarding the metabolite location being highly tighted to this specific IL. Other similar relations can be drawn with this type of visualization tool.

VI. Conclusions and Future Work

This work has proposed a neural network model for finding relationships among the circuit of metabolic regulation in tomato fruits by using genetic materials with specific introgressed alleles from exotic germoplasm. The model is oriented towards discovering unknown relationships between transcriptional and metabolic data, also providing simple visualizations for identification of co-expressed genes and co-accumulated metabolites. The model and visualizations presented in this work can be used for inspecting the neurons activated only with both types of data, which can be easily identified through the simple color code. A case study with the application of the proposed model has been presented as well, which involved genes expression measurements and metabolite profiles from tomato fruits. Several examples were shown, including the detection of a group of metabolites involved in a well known metabolic pathway.

As future work, it would be interesting to compare the proposed model with the neural gas algorithm, as an alternative to obtain the maps. Also, the proposed model will be used for finding relationships within the network that regulate tomato fruit metabolism. This way, the grouped patterns could be checked against the available metabolic pathways databases available online (e.g. the Kyoto Encyclopedia of Genes and Genomes) for finding possible metabolic pathways involving them.

References

[1] D. Tasoulis, V. Plagianakos, and M. Vrahatis, *Computational Intelligence in Bioinformatics*, ser. Studies in Computational Intelligence, A. Kelemen, A. Abraham, and Y. Chen, Eds. Springer, 2008, vol. 94.

[2] A. Polanski and M. Kimmel, *Bioinformatics*. Springer-Verlag, NY, 2007.

[3] L. Wang and X. Fu, Eds., *Data Mining with Computational Intelligence*, ser. Advanced Information and Knowledge Processing. Springer, 2005.

[4] 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*, vol. 286, pp. 531–7, 1999.

[5] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. Wiley, 2003.

[6] 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.

[7] 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.

[8] V. Lacroix, L. Cottret, P. Thebault, and M.-F. Sagot, "An introduction to metabolic networks and their structural analysis," *IEEE Transactions on computational biology and bioinformatics*, vol. 5, no. 4, pp. 594–617, 2008.

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

[10] A. Kelemen, A. Abraham, and Y. Chen, *Computational Intelligence in Bioinformatics*. Springer, 2008.

[11] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, pp. 59–69, 1982.

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

[13] 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.

[14] 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.

[15] L. Rieseberg and J. Wendel, *Introgression and its consequences in plants*, R. G. Harrison, Ed. Oxford University Press, 1993, vol. 1.

[16] S. Y. Lippman, Z.B. and Z. D., "An integrated view of quantitative trait variation using tomato interspecific introgression lines," *Current Opinion in Genetics and Development*, vol. 17, pp. 1–8, 2007.

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

[18] 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, Jun. 2005.

[19] 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, 2005.

[20] 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.

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