ANALYSIS AND INTEGRATION OF BIOLOGICAL DATA: A DATA MINING APPROACH USING NEURAL NETWORKS

Diego Milone, Georgina Stegmayer, Matías Gerard
Reseach Center for Information Tecnologies (CELTIC)
National Scientific and Technical Research Council, Argentina

Laura Kamenetzky, Mariana López, Fernando Carrari
Institute of Biotechnology (IB-INTA)
National Scientific and Technical Research Council, Argentina

Abstract

The volume of information derived from postgenomic technologies is rapidly increasing. Due to the amount of data involved, novel computational methods are needed for the analysis and knowledge discovery into the massive data sets produced by these new technologies. Furthermore, data integration is also gaining attention for merging signals from different sources in order to discover unknown relations. This chapter presents a pipeline for biological data integration and discovery of a-priori unknown relationships between gene expression and metabolite variations. In this pipeline, two standard clustering methods are compared against a novel neural network approach. The neural model provides a simple visualization interface for identification of coordinated patterns variations, independently of the number of produced clusters. Several quality measurements have been defined for the evaluation of the clustering results obtained on a case study involving transcriptomic and metabolomic profiles from tomato fruits. Moreover, a method is proposed for the evaluation of the biological significance of the clusters found. The neural model has shown a high performance in most of the quality measures, with internal coherence in all the identified clusters and better visualization capabilities.

1 Introduction

Nowadays, the biology field is in the middle of a data explosion. A series of technical advances in recent years has increased the amount of data that biologists can record about different aspects of an organism at the genomic, transcriptomic and proteomic levels [Keedwell and Narayanan, 2005].

Nowadays, the discipline of computational biology has allowed biologists to make full use of the advances in computer science and statistics in analysing these information. Due to the amount and nature of the biological data involved (such as noisy and missing data), novel computational methodologies are needed for properly analyzing it. Moreover, as the volume of data continues to grow at a high speed, new challenges appear, such as the need to extract information that was not previously known from these databases to supplement current knowledge. For example, the discovery of hidden patterns of gene expression in microarray 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. In addition, data integration is also gaining attention given the need for merging and extracting knowledge from signals of different sources and nature. Visualization of results is also an important issue for the understanding and interpretation of hidden relationships [Tasoulis et al., 2008].

The discipline 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 on microarrays [Polanski and Kimmel, 2007]. The prediction of classes uses the available information on the expression profiles and the known characteristics of the sets of data 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 formulated [Golub et al., 1999]. Among class discovery techniques, the hierarchical clustering (HC) algorithm is the most commonly used technique in biological data. It 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 constructs 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 [Tasoulis et al., 2008]. 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 (KM)

[Duda and Hart, 2003].

One of the current trends in the field is the integration of two types of biological data: metabolic profiles and transcriptional data from microarrays, with the objective of finding hidden relations among them and to infer new knowledge about the biological processes that involve them [Bino et al., 2004]. 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 data of transcriptome and metabolome in plants, correlating gene transcription profiles with variations profiles of a large number of non-protein molecules, can be used for identifying changes not reflected in the plant morphology [Carrari et al., 2006]. This allows having a snapshot of the metabolic pathways 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 (for example tomato plant) with certain desired characteristics [Lacroix et al., 2008].

The analysis of large biological datasets resulting from the different "omics" fields [1] is usually focussed on three main goals [Lindon et al., 2007]:

1. determine a significant difference between groups related to an effect of interest,
2. visualize differences, trends and relationships between samples and variables, and
3. detect which components (for example, genes) are responsible for the changes.

The new challenges that have arisen in computational biology indicate the need for the development of new data mining techniques to overcome the limitations of existing ones in satisfying these three points [Polanski and Kimmel, 2007]. Among the current proposals in the area, soft computing tools have been mentioned recently [Keedwell and Narayanan, 2005],

[1] Genomics, proteomics, transcriptomics and metabolomics.

in particular artificial neural networks [Kelemen et al., 2008, Tasoulis et al., 2008]. Specifically within artificial neural network models, self-organizing maps (SOM) [Kohonen, 1982, Kohonen et al., 2005] 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 fact, SOMs have been applied to analyse expression profiles in several systems biology studies lately, and it was one of the first machine learning techniques used for these kind of analysis [Quackenbush, 2001].

In [Hirai et al., 2004] a SOM model is proposed for the integrated analysis of *Arabidopsis thaliana* metabolome and transcriptome datasets. A related work [Yano et al., 2006] 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 corroboration rather than knowledge discovery. However, 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 [Saito et al., 2008].

In many cases, 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 lines 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. Or the focus may be the identification of meaningful biological points (markers) that are hidden within large-scale analytical intensity measurements from metabolomic experiments. For these tasks, many software computing tools implementing the use of SOMs have appeared lately, such as [Kaever et al., 2009], which performs data mining on intensity-based profiles using one-dimensional self-organizing maps; or [Tokimatsu et al., 2005] which is a web-based tool for representing quantitative data for individual transcripts and/or metabolites on plant metabolic pathway maps.

Differently from the mentioned approaches, in this chapter we present a methodology for finding relationships among introgression lines compared to a wild type control, instead of data evolving over time. An important contribution of this chapter consists in presenting the pipeline for biological data preprocessing, integration and mining towards the discovery of metabolic pathways. Furthermore, the proposed methodology is oriented towards discovering new and unknown relationships among transcriptional and metabolic data, instead of verifying an a-priori condition or performing a guided analysis. We propose the use of different kind of measurements for evaluating the quality of the clusters found by different clustering techniques. In the case of the neural clustering [Stegmayer et al., 2009], the model 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.

This chapter is organized as follows. First, a brief review of standard clustering algorithms is given. Second, the most relevant quality measures are presented. In Section 4, the pipeline for integration and analysis of introgression lines will be detailed. In this sequence of steps, visualization capabilities, quality measures and biological assesment will be shown for a case of study involving transcriptional and metabolic data of tomato fruits. Finally, the conclussions and future work can be found.

2 A brief review of clustering algorithms

Data mining methods may be categorized as either supervised or unsupervised. In unsupervised methods, no target variable is identified. Instead, the data mining algorithm searches for patterns and structure among all the variables. The most common unsupervised data mining method is clustering [Olson and Delen, 2008]. Clustering refers to the grouping of records, observations or cases into classes of similar objects. A cluster is a collection of records that are similar to one another and dissimilar to records in other clusters. The clustering task does not try to classify, estimate or predict the value of a target variable. Instead, clustering algorithms seek to segment the entire data set into relatively homogeneous subgroups or clusters, where the similarity of the records within the cluster is maximized, and the similarity to records outside this cluster is minimized [Larose, 2005].

The result of clustering can be expressed in different ways. The groups that are identified may be exclusive so that any instance belongs to only one group. Or they may be overlapping so that an instance may fall into several groups. Furthermore, they may be probabilistic, where an instance belongs to each group with a certain probability (or fuzzy membership) [Chakrabarti et al., 2009].

In the following subsections, the fundamentals of the basic clustering data mining algorithms used in this chapter and some basic concepts regarding the validation of the found clusters are presented. We denominate X to the dataset formed by x_i data samples, Ω to the set of samples that have been grouped in a cluster and W to the set of w_i centroids of the clusters in Ω.

2.1 Hierarchical clustering

One of the simplest and most popular unsupervised method in post-genomic data analysis is hierarchical cluster (HC) analysis. This method clusters the data forming a tree diagram, or dendrogram, which shows the relationships between samples according to a proximity matrix. The root node of the dendrogram represents the whole data set, and each leaf node is regarded as a data point. The intermediate nodes describe how the samples are proximal to each other; and the height of the dendrogram usually expresses the distance between each pair of data points or clusters, or a data point and a cluster. The clusters are obtained by cutting the dendrogram at different levels [Larose, 2005].

HC groups data with a sequence of nested partitions, either from singleton clusters to a cluster including all individuals or vice-versa. The former is known as agglomerative HC, and the latter is called divisive HC. Agglomerative HC clustering starts with N clusters, each of which includes exactly one data point. The algorithm then computes the distances between all pairs of data points in the multidimensional parameter space. This is usually computed on a Euclidean basis, though other distance metrics can be used. A series of merge operations is then followed that eventually forces all objects into the same group. In the proximity matrix, the minimal distance between two clusters is searched and those two clusters are

Algorithm 1: Agglomerative HC clustering.

```
Data:
    $X$: dataset
    $k$: number of clusters
Results:
    $\Omega$: clusters
    $W$: centroids
begin
    $N \leftarrow \mathrm{size}(X)$
    Start with $N$ singleton clusters: $\mathbf{w}_i = \mathbf{x}_i, \quad i = 1, \ldots, N$
    Calculate the proximity matrix: $\delta_{ij} = \|\mathbf{w}_i - \mathbf{w}_j\|, \quad 0 < i \neq j \leq N$
    while $N > k$ do
        Search the minimal distance $i^* j^* = \arg \min_{\forall i, \forall j} \{\delta_{ij}\}$
        Combine clusters $\Omega_{i^*}$ and $\Omega_{j^*}$ into a new cluster $\Omega_{ij^*}$
        Remove clusters $\Omega_{i^*}$ and $\Omega_{j^*}$ from $\delta$
        Update $\delta$ by computing the distances to and from $\Omega_{ij^*}$
        Update clusters in the HC tree
        $N \leftarrow N - 1$
    $\Omega \leftarrow$ patterns in the HC tree branches at top level reached
    $W \leftarrow$ centroids of clusters in $\Omega$
end
```

then combined into a new one. After that, the proximity matrix is updated by computing the distances between the new cluster and the other clusters. These steps are repeated until only one cluster remains. The HC method is shown in Algorithm 1.

2.2 k-means

The k-means (KM) algorithm is one of the best-known and most popular clustering algorithms [Forgy, 1965, Duda and Hart, 2003]. The Algorithm 2 shows its functioning. It begins by selecting the desired number of k clusters and assigning their centroids to data points randomly chosen from the training set. At each iteration, data points are classified by assigning them to the cluster whose centroid is closest and then new cluster centroids are computed as the average of all the points belonging to each cluster. This process continues until both the cluster centroids and the class assignments no longer change.

This technique inherently looks for compact, spherical clusters. The KM algorithm has become one of the most widely used clustering approaches finding many applications in post-genomics, especially in the analysis of transcriptomic data [Lindon et al., 2007]. KM assumes that the number of clusters k is already known by the user, which, unfortunately, usually is

Algorithm 2: *k*-means clustering.

```
Data:
    $X$: dataset
    $k$: number of clusters
Results:
    $\Omega$: clusters
    $W$: centroids
begin
    $N \leftarrow \text{size}(X)$
    Randomly initialize centroids: $\mathbf{w}_i = \mathbf{x}_{\text{rnd}(1,\ldots,N)}, \quad i = 1, \ldots, k$
    repeat
        foreach $\mathbf{x}_j \in X$ do
            Search the minimal distance $i^*_j = \arg\min_{\forall i}\{\|\mathbf{w}_i - \mathbf{x}_j\|\}$
            Assign pattern $\mathbf{x}_j$ to cluster $\Omega_{i^*_j}$
        Recalculate centroids $\mathbf{w}_i = \frac{1}{|\Omega_i|} \sum_{\mathbf{x}_j \in \Omega_i} \mathbf{x}_j$
    until $W$ do not change
end
```

not true in practice. Like for cluster initialization, there are also no efficient and universal methods for the selection of k [Xu and Donald C. Wunsch, 2009].

2.3 Neural networks

Neural networks have solved a wide range of problems and have good learning capabilities. Neural network based clustering is closely related to the concept of competitive learning, which is based on the idea of units (neurons) that compete in some way 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 synaptic weight vector [Xu and Donald C. Wunsch, 2009].

Given an input pattern, its similarity (or distance) to the weights vectors is computed. The neurons in the competitive layer then compete with each other, and only the one closest to the input becomes activated or fired. The weight vector of this winning neuron is further moved towards the input pattern. This competitive learning paradigm only allows learning for a particular winning neuron that best matches the given input pattern and it is also known as winner-take-all learning [Haykin, 2007].

Self-organizing maps were introduced in 1982 by Teuvo Kohonen [Kohonen, 1982]. They represent a special class of neural networks that use competitive learning. The goal of self-organizing maps is to 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 [Kohonen et al., 2005]. Thus, SOMs can be appropriate for cluster analysis when looking for underlying hidden patterns in data. A SOM structures the output nodes (neurons) into clusters of nodes, where nodes in closer proximity are more similar to each other than to other nodes that are farther apart. A neighborhood function is defined for each neuron. When competition among the neurons is complete, SOM updates a set of weight vectors within the neighborhood of the winning neuron (see Algorithm 3).

Having finished the training, neighboring input patterns are projected into the lattice, corresponding to adjacent neurons connected to each other through the neighborhood function, giving a clear topology of how the network fits into the input space. Therefore, the regions with a high probability of occurrence of sampled patterns will be represented by larger areas in the feature map [Haykin, 2007]. In this sense, some authors prefer to think of SOM as a method of displaying latent data structures in a visual way rather than through a clustering approach [Xu and Donald C. Wunsch, 2009].

3 Clustering validation for the comparison of algorithms

Given a data set, each clustering algorithm can always produce a partition whether or not there really exists a particular structure in the data. Moreover, different clustering approaches usually lead to different clusters of data, and even for the same algorithm, the selection of a parameter or the presentation order of input patterns may affect the final results. Therefore, effective evaluation standards and criteria are critically important to provide users with a degree of confidence for the clustering results.

The discovery of novel biological knowledge from the analysis of post-genomic data relies upon the use of unsupervised processing methods, in particular clustering techniques. Much recent research in bioinformatics has therefore been focused on the transfer of clustering methods introduced in other scientific fields and on the development of novel algorithms specifically

Algorithm 3: Self-organizing map training.

```
Data:
    $X$: input vector
    $k$: number of neurons for a $k = n \times n$ map
Results:
    $\Omega$: clusters
    $W$: centroids
begin
    $N \leftarrow \text{size}(X)$
    Define neurons neighborhood function $\Lambda(n)$
    Initiliaze the map by choosing random weights values $w_{ij} \in [-0.5, +0.5]$
    repeat
        Select a pattern at random $\mathbf{x}_r = \mathbf{x}_{\text{rnd}(1,\ldots,N)}$
        Search for the winning neuron: $j^* = \arg\min_{\forall j} \{\|\mathbf{w}_j - \mathbf{x}_r\|\}$
        Adapt weights: $\mathbf{w}_j \leftarrow \left\{ \begin{array}{ll} \mathbf{w}_j + \eta\,(\mathbf{x}_r - \mathbf{w}_j) & \text{if } j \in \Lambda_{j^*} \\ \mathbf{w}_j & \text{if } j \notin \Lambda_{j^*} \end{array} \right\}$
    until no significative changes in $\mathbf{w}_j$
    $\Omega_j \leftarrow \{\mathbf{x}_\ell / \|\mathbf{w}_j - \mathbf{x}_\ell\| < \|\mathbf{w}_i - \mathbf{x}_\ell\| \; \forall i \neq j, 0 < i \leq k\}$
end
```

designed to tackle the challenges posed by post-genomic data. To avoid inconsistencies in the results and to assure that the resulting clusters are reflective of the general population, the clustering solution should be validated [Larose, 2005]. The partitions returned by a clustering algorithm are commonly validated using visual inspection and concordance with prior biological knowledge. Suitable computational cluster validation techniques are available in the general data-mining literature, but have been given only a fraction of the same attention in bioinformatics [Handl et al., 2005].

The data-mining literature provides a range of different validation techniques, distinguishing between external and internal validation measures. External validation measures comprise all those methods that evaluate a clustering result based on the knowledge of the correct class labels. Internal measures take a clustering and the underlying dataset as the input, and use information intrinsic to the data to assess the quality of the clustering [Halkidi, 2001].

When clustering a novel biological dataset, cluster validation plays a very different role. A completely objective validation of cluster quality is usually impossible in such a case, but the use of cluster validation during the clustering process can help to improve the quality of results, and increase the confidence in the final result. No reliable method exists to identify the number of clusters in an unknown dataset, and the choice of the best number of clusters may well depend on the clustering method used. A cluster analysis should therefore always be performed for a (sensible) range of different numbers of clusters. Internal validation measures should be used to provide feedback on the quality of the data and to check whether a given partitioning is justified in terms of the underlying data distribution. A good clustering solution will tend to perform reasonably well under multiple measures.

Internal measures can be grouped according to the particular notion of clustering quality that they employ. In [Handl et al., 2005] the following classification of internal measures is proposed:

Type I measures (*Compactness*): it comprises validation measures assessing cluster compactness or homogeneity. Intracluster variance is their most popular representative:

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

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.

Type II measures (*Separation*): this group includes all those measures that quantify the degree of separation between individual clusters. It can be evaluated measuring mean, minimum and maximum Euclidean distance among cluster centroids, according to:

$$\begin{aligned}
\overline{S} &= \frac{2}{k^2 - k} \sum_{i=1}^{k} \sum_{j=i+1}^{k} \|\mathbf{w}_i - \mathbf{w}_j\|_2, && (2)\\
S_m &= \min_{0<i\neq j\leq k} \left\{ \|\mathbf{w}_i - \mathbf{w}_j\|_2 \right\}, && (3)\\
S_M &= \max_{0<i\neq j\leq k} \left\{ \|\mathbf{w}_i - \mathbf{w}_j\|_2 \right\}, && (4)
\end{aligned}$$

where $\overline{S}$ close to zero indicates closer nodes.

Type III measures are *combinations* of the mentioned two types of measures and they are the most popular because they exhibit opposing trends. For example, while compactness improves with an increasing number of clusters, the distance between clusters tends to deteriorate.

The first combined measurement used in this work is the Davies-Bouldin index [Davies and Bouldin, 19

defined as

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\overline{C}_i + \overline{C}_j}{\|\mathbf{w}_i - \mathbf{w}_j\|_2} \right). \qquad (5)$$

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

The other combined measurement used is the internal cluster dispersion rate of the final partition defined as [Mingoti and Lima, 2006]

$$\Upsilon = 1 - \left(\frac{\sum_{j=1}^{k} \left\| \mathbf{w}_j - \left(\frac{1}{N} \sum_{\ell=1}^{N} \mathbf{x}_\ell \right) \right\|_2}{\sum_{i=1}^{N} \left\| \mathbf{x}_i - \left(\frac{1}{N} \sum_{\ell=1}^{N} \mathbf{x}_\ell \right) \right\|_2} \right), \qquad (6)$$

The numerator in (6) corresponds to the sum of the distances among the centroids and the overall sample mean vector; the denominator is the sum of the distances between each pattern and the overall sample mean vector. The smaller the value of Υ, the smaller the intraclass clusters dispersion.

4 Pipeline for integration and analysis of introgression lines

Since the completion of genome sequences, functional identification of unknown genes has become a principal challenge in systems biology. The analysis of biological data, such as, for example, gene expression data and metabolic profiles in plants of agroeconomical interest, is based on the idea that genes (and metabolites) that are involved in a particular metabolic pathway should be co-regulated and therefore should exhibit similar patterns of expression. Thus, a fundamental task for their analysis is to identify groups of data samples showing similar expression patterns. For this task, clustering has become a fundamental approach. It can support the identification of existing underlying relationships among a set of variables such as biological conditions or perturbations. Clustering may, even, allow the discovery of new biological knowledge.

As shown in Section 2, there are several clustering algorithms and most of them, at least indirectly, assume 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 the clusters is spherical [Azuaje and Bolshakova, 2002]. Unfortunately, this type of knowledge may not always be available beforehand in biological data (as in other applications). In general, the application of two or more clustering techniques may provide a basis for the synthesis of accurate and reliable results, especially if similar results are obtained by using different techniques. However, from the application point of view, it is import to be able to quantify the confidence in each method, in particular when new biological inferences could be made from the analyzed data.

In this section of the chapter, we present a pipeline of steps that can be followed to help the discovery of gene-to-gene and metabolite-to-gene networks thanks to the integration of metabolomics with transcriptomics. It is oriented towards the cases where a biological experiment is focused in the study of the differences among several plant genomes, involving an original genome that has been modified by introgression of another species genome.

As part of the process, we propose a specific step that can help in assessing the validity of one clustering algorithm over another one, not only from the point of view of the quality of the clusters found, but also from the biological meaning of the found relations. We introduce several measures that can be applied to the clusters and biological criteria to verify their biological value. This last step presents a strategy that can be of help for the identification of novel metabolic pathways. Figure 1 shows the the proposed process, which consists of the following steps:

1. **IL-Data understanding**: it involves the definition of the biological data, the ILs involved in the study, the kind and number of data types (for example, metabolites and transcripts), the number of experiments and repetitions for each of them, as well as the structure and type of datafiles that contain them (such as raw quantified array data from microarray specific software).

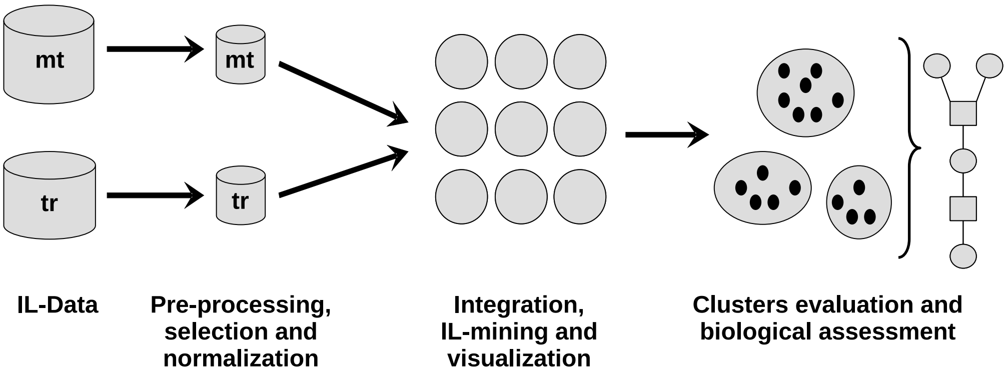

mt
mt
tr
tr
IL-Data
Pre-processing, selection and normalization
Integration, IL-mining and visualization
Clusters evaluation and biological assessment

Figure 1: Pipeline for integration and analysis of biological (metabolic and transcriptomic) data of introgression lines.

2. **Pre-processing, selection and normalization**: it involves the cleaning and error elimination from data, as well as the application of appropiate selection criteria with the objetive of including only sufficiently expressed data in the analysis [Baxter et al., 2005]. This step needs also a normalization of the log ratio of the expression intensity values over the control sample in the case of data coming from microarray experiments [Causton et al., 2003].

3. **Integration, IL-mining and visualization**: with an appropriate normalization, metabolobome and transcriptome data obtained from the same plant material can be integrated into a single multivariate dataset suitable for further analysis with a clustering technique. Several data mining algorithms should be applied, whose relevance is later determined in the next pipeline step. Simple and easy to undertand visualization of results are used for the interpretation of hidden relationships among co-expressed and co-accumulated genes and metabolites, in particular, groups of different kind of patterns. This kind of feature can be useful for a prelimiray evaluation of the clusters.

4. **Clusters evaluation and biological assessment**: For choosing an adequate clustering technique for the data, we propose the comparison of several methods through clustering measurements (that will indicate the quality of the clusters found) and through a biological assesment of the clusters that integrate different kinds of data (that will indicate biological significance of the found aggrupations).

The following subsections describe the proposed steps of the pipeline in detail, through

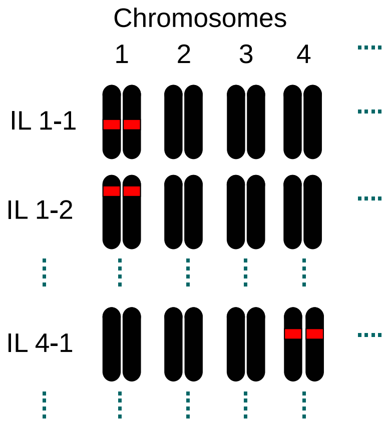


Figure 2: Introgression lines (ILs): the portion of introgressed genome is marked in each IL with a red line.

the application of the proposal to a case of study involving a commercial tomato database[2].

4.1 IL-Data understanding

The metabolome is the final product of a series of gene actions. Hence, metabolomics has a potential to elucidate gene functions and networks, especially when integrated with transcriptomics [Hirai et al., 2005]. Like it was mentioned before, there are many cases where a biological experiment is focused in 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 or transgenic plants overexpressing a gene of interest. 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) [Rieseberg and Wendel, 1993, Lippman and D., 2007].

The case of study presented in this chapter involves the analysis of metabolic and transcriptional profiles data from ILs of tomato (*Solanum lycopersicum*) which posses, at certain chromosomes segments, introgressed portions of a wild tomato species (*Solanum pennelli*). The interest in comparing the cultivated tomato against the different ILs lies on the fact that it has been proven that some wild tomato fruits can be sources of several specific agronomical characters of interest which could be used for the improvement of commercial tomato lines. Figure 2 shows a scheme of the chromosomes of this plant population.

The metabolic data have been obtained analyzing polar extracts of tomato fruits, through

[2] http://ted.bti.cornell.edu/

Gas Chromatography coupled to Mass Spectometry. The peak intensities have been normalized to the quantity of material used. The metabolite profiling technique used allows the identification of approximately 80 primary metabolic compounds [Carrari et al., 2006].

Transcriptional levels have been obtained from hybridization chips having all the genes of the material of interest (all ≈25000 tomato genes) ordered into spots, previously marked with two fluorescence channels. 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 genes. $P = 21$ ILs have been analyzed, with introgressions in chromosomes: 1, 2, 3, 5, 8, 10, 11 and 12 [Baxter et al., 2005].

4.2 Pre-processing, selection and normalization

The following steps aim at filtering error or missing measurements, NaN values, poor quality spots in the microarray images (according to the flags from the image processing program) and spots not expressed in both channels (such as empty spots or spots that do not pass a minimum empirical threshold value). The following lines present a detailed description of the pre-processing, selection and normalization steps performed for each type of data type analyzed.

4.2.1 Metabolite data

Metabolite accumulation measurements are obtained from several replicates of an experiment. Each experiment compares a specific introgression tomato line against a control tomato genome. Metabolites that do not appear in at least two repetitions, are marked as missing data for further analysis. For each metabolite in each IL, the log ratio of the mean of the valid replicates is calculated according to

$$logR_i^m = \log_{10}\left(\frac{\bar{S}_i^m}{\bar{Q}_i^m}\right) \tag{7}$$

where $\bar{S}_i^m$ is the accumulation mean of the valid replicates for metabolite m at the IL i, and $\bar{Q}_i^m$ is the accumulation mean of the valid replicates for the corresponding control measurement. In the selection step, only metabolites with $|logR_i^m| > 0.1$ are kept for data integration and cluster analysis.

4.2.2 Transcriptional Data

Poor quality spots, negative spots, spots not expressed in both channels and empty spots are filtered out. Not expressed spots are detected for IL and control slides according to

$$\bar{F}^t < \bar{B}^t + \alpha\tilde{B}^t \tag{8}$$

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 in the interval $[2, 3]$.

The microarray measurements are normalized using the print-tip Lowess normalization strategy [Causton et al., 2003]. Spots with at least two valid replicated data points are included for analysis using

$$logR^t_{ir} = \log_2\left(\frac{\check{S}^t_{ir}}{\check{Q}^t_{ir}}\right) \tag{9}$$

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

The valid replicates are averaged simply as

$$logR^t_i = \frac{1}{\Gamma^t_i}\sum_{r=1}^{\Gamma^t_i} logR^t_{ir} \tag{10}$$

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

4.2.3 Normalization

After the pre-procesing and selection steps, $M = 71$ metabolites and $T = 1385$ genes have been selected as sufficiently expressed. The resulting log ratios are normalized. For each pattern, the sum of the square of log ratios are set equal to 1 according to

$$x^*_i = \frac{logR^*_i}{\sum_{j=1}^{P}(logR^*_j)^2} \tag{11}$$

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

4.3 Integration, IL-mining and visualization

All normalized data are integrated into a single matrix and arranged in the training set as shown in Table 1. Various ways of integration are possible. For example, before integration of two data sets, the plus/minus sign of one data set can be reversed to obtain negatively correlated items. To find all possible inverted correlations (direct-direct, direct-inverted), the training set includes the original and the inverted versions of all the patterns. Then, each column/dimension-IL is normalized in the range $[0, 1]$ according to a histogram equalization.

For the analysis of these biological data, clustering is implemented, under the assumption that behaviorally similar genes could share common pathways. According to this principle, named "guilt-by-association" [Wolfe et al., 2005], 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.

In this step, several clustering methods are applied to the IL-data. The proposed pipeline 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. Therefore, we have named IL-HC, IL-KM and IL-SOM to the models for clustering in this context.

In the case of the IL-SOM model, each node corresponds to a neuron, for IL-HC each branch is a node and in the case of IL-KM the nodes correspond to the k parts the data are divided by. Several model topologies, map sizes and initialization strategies are possible in IL-SOM. For the map shape we have used a rectangular lattice. The initial vectors are set by principal component analysis, obtaining a learning process independent of the order of input of vectors, and hence reproducible[3].

Let us call *node* to each of the k elements of the grouping method. In the three cases, we identify a node with the index j, the centroid with $\mathbf{w}_j$ and the set of patterns grouped

[3] http://www.cis.hut.fi/projects/somtoolbox/

Table 1: Data integration into a single training set. Original and inverted versions of all the data samples are included.

	IL$_1$	IL$_2$	…	IL$_i$	…	IL$_P$
Transcripts	$x_1^{t_1}$	$x_2^{t_1}$	…	$x_i^{t_1}$	…	$x_P^{t_1}$
	$x_1^{t_2}$	$x_2^{t_2}$	…	$x_i^{t_2}$	…	$x_P^{t_2}$
	⋮	⋮	⋱	⋮	⋮	⋮
	x_1^{t}	x_2^{t}	…	x_i^{t}	…	x_P^{t}
	⋮	⋮	⋱	⋮	⋮	⋮
	x_1^{T}	x_2^{T}	…	x_i^{T}	…	x_P^{T}
Inverted transcripts	$-x_1^{t_1}$	$-x_2^{t_1}$	…	$-x_i^{t_1}$	…	$-x_P^{t_1}$
	$-x_1^{t_2}$	$-x_2^{t_2}$	…	$-x_i^{t_2}$	…	$-x_P^{t_2}$
	⋮	⋮	⋱	⋮	⋮	⋮
	$-x_1^{t}$	$-x_2^{t}$	…	$-x_i^{t}$	…	$-x_P^{t}$
	⋮	⋮	⋱	⋮	⋮	⋮
	$-x_1^{T}$	$-x_2^{T}$	…	$-x_i^{T}$	…	$-x_P^{T}$
Metabolites	$x_1^{m_1}$	$x_2^{m_1}$	…	$x_i^{m_1}$	…	$x_P^{m_1}$
	$x_1^{m_2}$	$x_2^{m_2}$	…	$x_i^{m_2}$	…	$x_P^{m_2}$
	⋮	⋮	⋱	⋮	⋮	⋮
	x_1^{m}	x_2^{m}	…	x_i^{m}	…	x_P^{m}
	⋮	⋮	⋱	⋮	⋮	⋮
	x_1^{M}	x_2^{M}	…	x_i^{M}	…	x_P^{M}
Inverted metabolites	$-x_1^{m_1}$	$-x_2^{m_1}$	…	$-x_i^{m_1}$	…	$-x_P^{m_1}$
	$-x_1^{m_2}$	$-x_2^{m_2}$	…	$-x_i^{m_2}$	…	$-x_P^{m_2}$
	⋮	⋮	⋱	⋮	⋮	⋮
	$-x_1^{t}$	$-x_2^{t}$	…	$-x_i^{t}$	…	$-x_P^{t}$
	⋮	⋮	⋱	⋮	⋮	⋮
	$-x_1^{M}$	$-x_2^{M}$	…	$-x_i^{M}$	…	$-x_P^{M}$

in a node with Ω_j. We use the term "integration node" to reference a node that contains both different kinds of patterns (metabolites and transcripts). For all the methods, we use the Euclidean distance to measure distance between patterns, and clusters are generated for 50 and 200 nodes.

Figure 3 shows the resulting histograms for each method with $k = 50$. As can be seen, IL-HC comprises the vast majority of the patterns in the same branch (Figure 3.a)). This constitutes a major drawback to this technique, both from the perspective of its capabilities as a method of grouping this type of data as from the perspective of the information about the biological processes involved that can be inferred from this grouping. It is important to

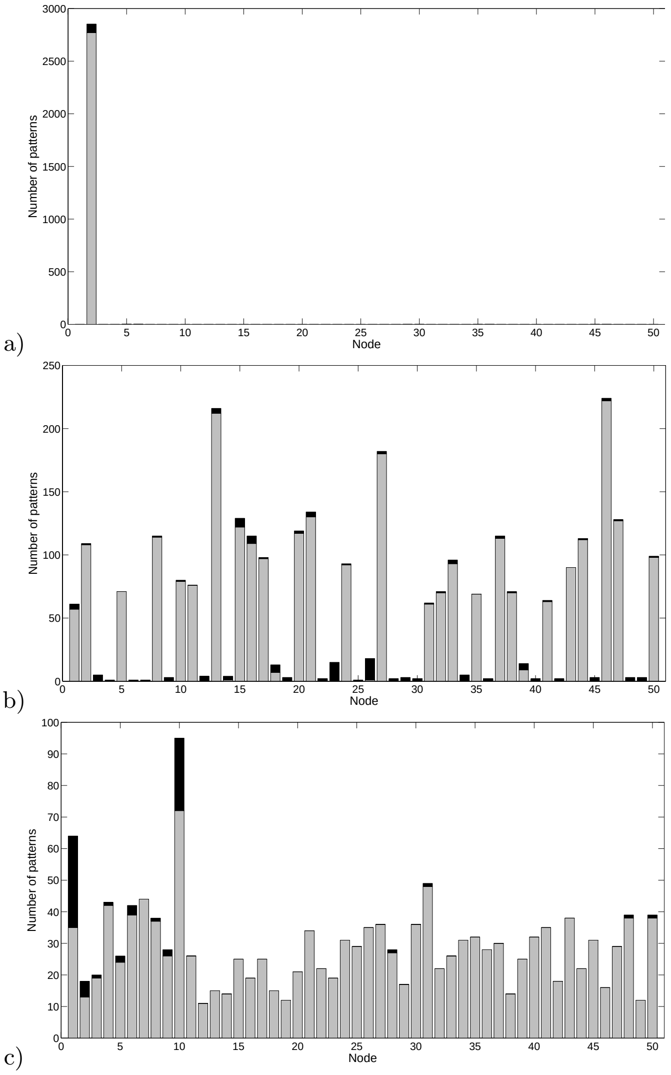


Figure 3: Patterns distribution histograms by clustering method (grey for transcripts, black for metabolites). a) IL-Hierarchical clustering; b) IL-k-means; c) IL-Self-organizing maps.

mention that, regardless of the depth at which the branch lines are cut, the method always tends to group the majority of patterns in a few nodes. Another major inconsistency detected in this case is that the original patterns have been grouped together with their inverted (sign) version, in many cases, in the same node.

When comparing the histograms of IL-KM (Figure 3.b)) and IL-SOM (Figura 3.c)), it can be seen that the distribution in the case of IL-SOM is much more uniform. While IL-KM has several nodes with very few patterns and few nodes with many patterns, the patterns distribution of the IL-SOM is more balanced across all the nodes. This is mainly due to the influence of the neighborhood radius during the self-organizing map training process. In

IL-KM, each node is trained independently from each other; in the case of SOM, instead, an update of the nodes around the winning neuron is performed, which allows the centroids not to spread so much and the patterns can be more homogeneously distributed in regions of neurons with similar centroids. The more balanced distribution of the patterns and the possibility of analyzing individual neurons also extending the analysis to those located nearby, for several radios, is a distinct advantage of the IL-SOM model in comparison with IL-KM.

The notion of a social network and associated methods of analysis and visualizations is influencing an increasing number of application domains, including bioinformatics and systems biology nowadays. A social network consists of a graph $G = (V, E)$ where there are relations among a set of actors or vertices V and edges E representing relations. From the very beginning, visualization played an important role in social network analysis, not only for presentation, but even more so by facilitating data exploration [Brandes, 2008].

We propose here the use of a social graph for visualization and comparison among clustering methods. For example, once the patterns have been assigned to clusters, it is easy to see, by visual inspection, if the direct and corresponding inverted patterns have been grouped consistently, or not.

For the case study example of tomato metabolites and transcriptes, considering only integration nodes for the case $k = 50$, Figure 4 shows, at a glance, that IL-HC has inconsistently grouped direct and inverted sign patterns into one single cluster. IL-KM, instead, has found coherent relationships, but scattered through a different number of integration nodes. Thus, the associations do not always reflected the opposite sign relationships. Differently from the other methods, IL-SOM has always correctly grouped direct and inverted sign data into different (in fact, opposite) clusters in the feature map. For the special case of the IL-SOM, many interesting representations of clusters can be obtained form the projection of the patterns in the lattice of neurons. If the dataset includes the original data plus the original data with inverted sign, the resulting map shows a simmetrical "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 (up-regulated genes and metabolites, down-regulated genes and metabolites) and inverted (down-regulated genes grouped together with up-regulated metabolites) relations among data. There is a specific zone in the map where

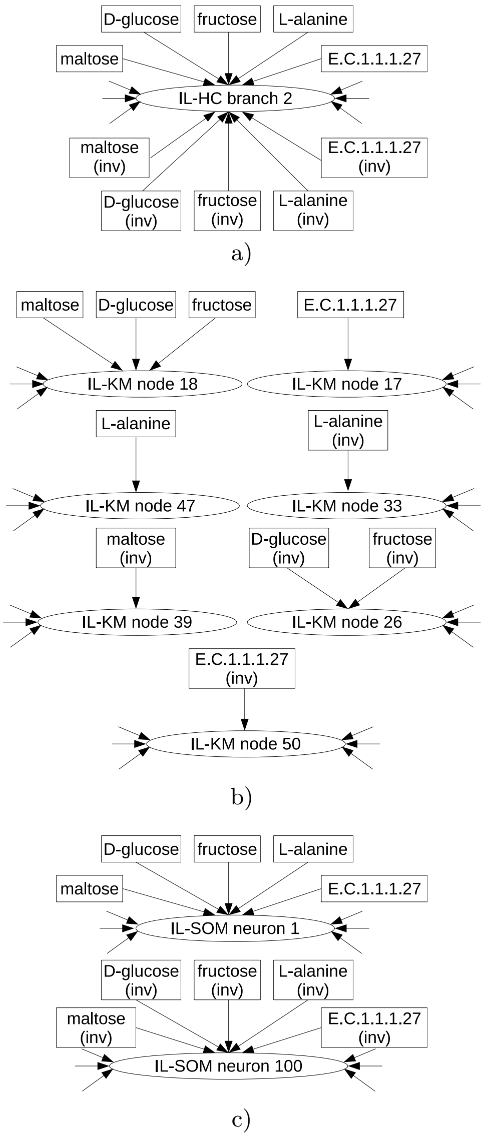
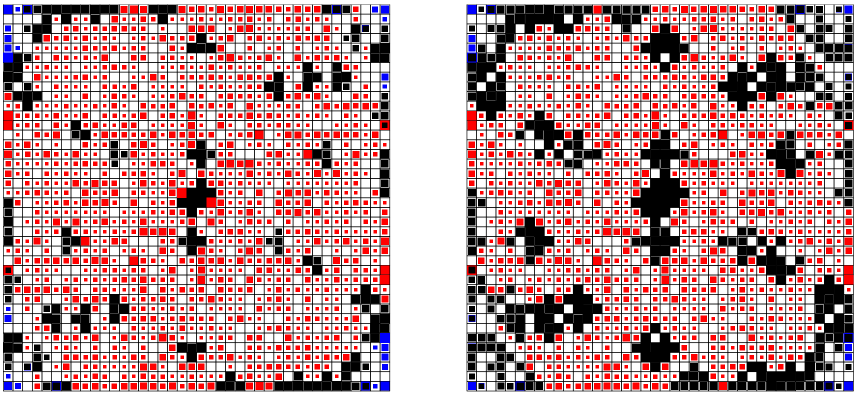


Figure 4: Social net graphs for clusters visualization: a) inconsistent cluster in IL-HC; b) scattered patterns in IL-KM and c) a coherent gruping of patterns in IL-SOM.

the exactly opposite behavior per IL can be found, which is useful for associating specific genes/metabolites to specific IL. For example, one gene similarly expressed in several ILs but up-regulated in a determined IL, which will be down-regulated in it if its sign is inverted. These kind of analysis may be of help for the further inference of a-priori unknown metabolic pathways involving the grouped data.

In a standard 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 [Ultsch, 1999]. It computes the average distance



a) b)

Figure 5: Activation SOM resulting from the integrated analysis of 1385 genes and 71 metabolites from 21 tomato ILs. Map topology of 40x40 neurons with a) $Vn = 1$ and b) $Vn = 2$.

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 the proposed IL-SOM model, instead, provide a simple interface for helping the quickly identification of co-expressed and co-accumulated genes and metabolites through a simple color code for the integration nodes. The focus is on the easily identification of groups of different types of patterns, independently from the number of neurons in a cluster. Furthermore, 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 the identification of neuron clusters. When a Vn is defined, all the neighborhood neurons (according to the neighborhood radius set) are considered as a group and treated altogether accordingly, also for counting whether there are metabolites and transcriptes grouped.

The following visualizations are supported by the IL-SOM model:

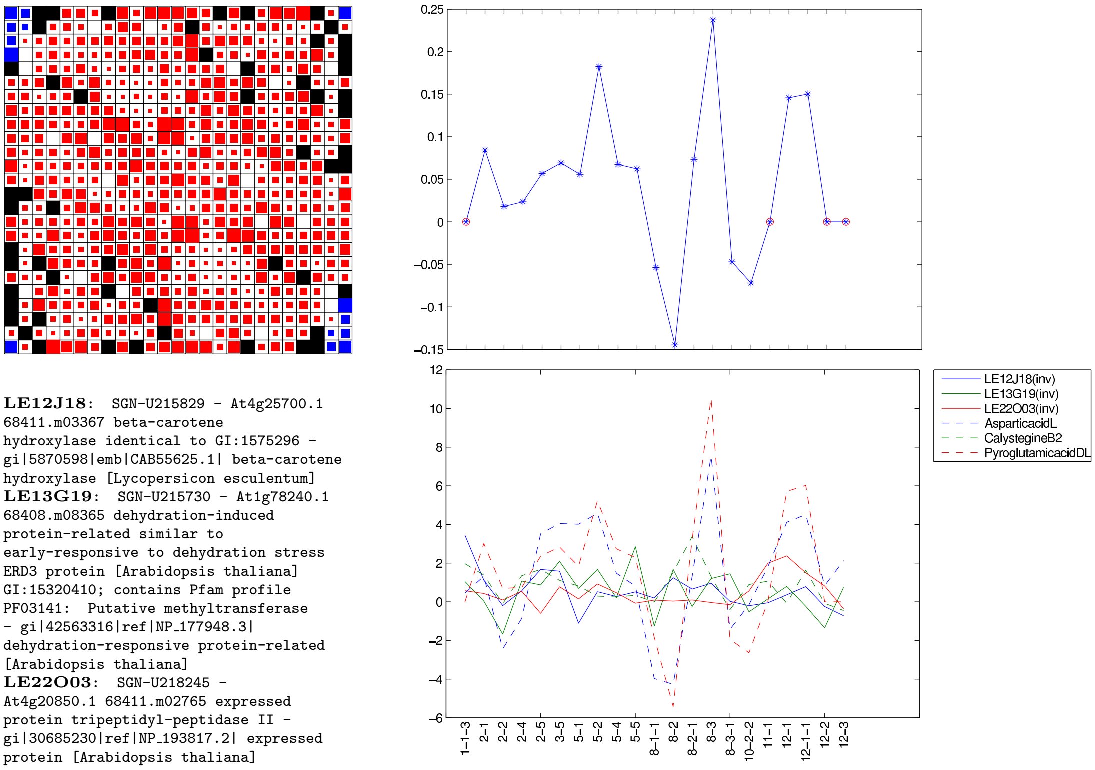
Easy identification of clusters of combined data types: Figure 5 shows different marker colors which indicate the kind of pattern grouped in the neuron: black for combined data types, blue for only metabolites and red for only transcripts. Also the marker size indicates the number of patterns grouped. The figure shows the activation map resulting from the integrated analysis of 21 tomato ILs with a 40x40 neuron topology, with $Vn = 1$ and $Vn = 2$.

Detail view of original data measurements: Figure 6 shows some visualizations provided by the software that has been designed for the IL-SOM. The figure shows, in the upper left, the resulting IL-SOM integrated model for the 21 ILs dataset (map with 20x20 neurons). The curves presented in the right part of the figure show a detail of the normalized patterns which have all been clustered together in the neuron 604: the metabolite *Pyroglutamic Acid.*, and the metabolites *Aspartic acid L* and *Calystegine B2*, together with the inverted sign transcriptes *LE12J18*, *LE13G19* and *LE22O03*. For this metabolite the upper right plot shows its denormalized (original) log ratios. The down left part of the figure shows a decodification of the *LE33K02* transcript according to its probe code, which has been automatically translated into its corresponding Arabidopsis[4] (At.3g16720.1) and Unigene[5](SGN-U217330) annotations[6]. In the denormalized plot, missing data (samples in a IL not having enough valid replicates experiments for the average log ratio calculation) are indicated with a red circle.

Visualization of clusters inside a specific chromosome: 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 an important deviation 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 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 7 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 $Vn = 1$. For example, in the patterns

[4] www.arabidopsis.org
[5] www.sgen.cornell.edu)
[6] http://www.sgn.cornell.edu/



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 6: 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.

inside neuron (1,1), the *Citric acid (Citrate)* 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 capability.

It is quite important to be able to evaluate the quality of several clustering algorithm when applied to biological data, in particular if later biological inferences should be made. This is the objective of the next step of the proposed biological data pipeline.

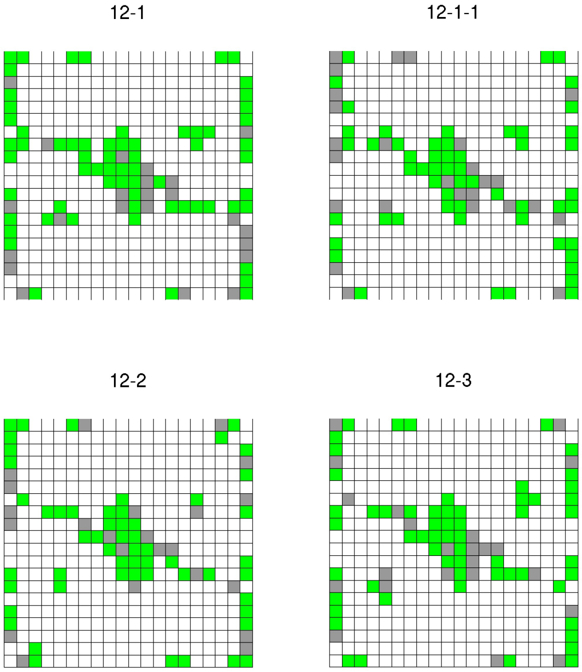


Figure 7: SOM activation map for the tomato chromosome 12, introgression lines 12-1-1, 12-1, 12-2 and 12-3.

4.4 Clusters evaluation and biological assessment

Several clustering algorithms have been employed to analyse the available IL expression data. However, a quality or confidence measure is necessary for each technique to assess the quality of their outcomes. This step provides a cluster validity framework to get insights into the clustering method selection for the biological data under analysis. The results obtained for the case of study indicate that this systematic evaluation approach may support the process for knowledge discovery applications over IL-data.

According to [Guillet and Hamilton, 2007], comparison between clustering methods can be performed using objective and/or subjective measures. An objective measure is based on only the raw data. No additional knowledge about the data is required. A subjective measure is a one that takes into account both the data and the user who uses the data. To define a subjective measure, the user domain or background knowledge about the data is needed. The key issue for mining patterns based on subjective measures is the representation of user knowledge.

For the comparisons among clustering methods in this pipeline step, we propose the use of two categories of measures: i) objective or clustering measurements, that measure the quality of the clusters found with each technique, regardless of their biological meaning; and ii) biological criteria have been defined for the assessment of the clusters that integrate metabolites with transcripts, from the point of view of the metabolic pathways that should involve them.

4.4.1 Comparison based on objective clustering measures

In this step, we compare the applied clustering method according to the clustering measures that have been defined in Section 3. Table 2 shows the results obtained in the comparison of the three methods, for different numbers of nodes. The method of IL-HC has concentrated more than 85% of the patterns in a single node, including direct and inverted data. This is a strong indication that it would not be a valid method for detecting coordinated changes in these kinds of patterns. More compact nodes and having low internode spacing ($\overline{C}$ and $\overline{S}$) have been obtained by IL-SOM. As for separation, it was noted that IL-HC and IL-KM tend to locate a centroid in each of the more distant patterns of the data set analyzed. The self-organizing maps are more robust, keeping the distances between neurons centroids through the neighborhood update during the early stages of the training algorithm. Clearly, by increasing the number of neurons in the map, the degrees of freedom are increased and the outer centroids of the map may get closer to the data patterns that are farther apart from the whole dataset. We can notice that the minimum separation is reduced in the IL-SOM in comparison to the other methods, which allows looking at the map with greater confidence that the changes between nearby nodes are gradual and can form clusters of more biological interest.

With respect to Type III measurements, it has to be considered that intranode dispersion is always greater for IL-SOM, regardless of the number of nodes. This means that the average distance between IL-SOM centroids and the global center of the patterns is lower (in relation to the total patterns dispersion) than the average distance of the other methods centroids. Since IL-HC finds a large node containing most of the patterns and IL-KM forms many spread nodes with a few patterns, all the distances among scattered nodes (and their centroids) are greater. Because of this, the DB index is the lowest for the case of IL-HC. This is not

Table 2: Quality measurements for the clustering methods compared. The two better values for each measure for each k are underlined.

Type	Measure	IL-HC		IL-KM		IL-SOM	
		50	200	50	200	50	200
I	$\overline{C}$	4.728	3.511	4.608	3.167	$\underline{3.433}$	$\underline{3.294}$
II	$\overline{S}$	21.51	13.57	12.29	9.577	$\underline{1.344}$	$\underline{2.118}$
II	S_m	8.635	4.159	1.892	1.369	$\underline{0.239}$	$\underline{0.196}$
II	S_M	45.92	45.92	45.92	45.92	$\underline{3.483}$	$\underline{7.363}$
III	DB	$\underline{2.280}$	$\underline{2.901}$	5.764	3.630	24.42	20.67
III	Υ	0.936	0.842	0.967	0.895	$\underline{0.995}$	$\underline{0.971}$

Table 3: Quality measurements for the clustering methods compared, considering only integration nodes. The two better values for each measure for each k are underlined.

Type	Measure	IL-HC		IL-KM		IL-SOM	
		50	200	50	200	50	200
	N	1	13	26	35	13	21
I	$\overline{C}$	3.787	$\underline{3.356}$	$\underline{3.638}$	3.588	4.104	4.660
II	$\overline{S}$	–	7.493	4.214	6.591	$\underline{1.609}$	$\underline{2.913}$
II	S_m	–	4.372	1.892	2.733	$\underline{0.302}$	$\underline{0.371}$
II	S_M	–	12.51	11.02	15.14	$\underline{3.483}$	$\underline{6.118}$
III	DB	–	$\underline{2.516}$	$\underline{2.472}$	3.274	18.13	11.28
III	Υ	–	0.994	0.993	0.986	$\underline{0.998}$	$\underline{0.995}$

happening in IL-SOM because the distances between centroids are always smaller, since they are better distributed and centroids are not associated to remote and isolated patterns There are more centroids to be distributed for sharing and it is not forced to concentrate many patterns in few centroids. Moreover, since the farther away patterns (probably *outliers*) have to be associates to any centroid, also nodes compactness decreases and hence the DB index is the highest.

Table 3 shows the results obtained in the comparison of the three methods, considering in this case only integration nodes. The interest of this particular analysis lies in the fact that the patterns grouped into these nodes may be parts of the same metabolic pathway. As can be seen, a row has been added to the table with the number of integration nodes found by each technique. As expected, adding more degrees of freedom to the techniques, more nodes of this type are found. IL-KM is the method that founds the higher number of integration nodes having also high cohesion. However, the detail of these clusters indicates that patterns have not been grouped coherently in all cases: there are nodes grouping non-inverted versions of patterns but their inverted versions have been dispersed along several nodes, even mixing non-inverted and inverted versions of data in some cases (which is also inconsistent). On the other hand, IL-HC finds the least amount of nodes, but with the problems highlighted above regarding grouping, in the same node, a direc and an inverted version of the patterns, which is meaningless from the mathematical as well as from the biological viewpoint. IL-SOM, instead, always finds consistently grouped data.

As for Type II measures, IL-SOM has obtained the best rates in terms of minimum, maximum and average integration nodes separation. The IL-HC technique with $k = 50$ has found a single integration node with 98% of the data. Measures of Type II can not be calculated for this case because there is no separation between nodes (there is a single integration node). Similarly, combined measures cannot be calculated for this case because it is meaningless to measure the overlap of a single node. The intercluster dispersion is best for the SOM, although very close to 1.0 in all the other methods. For the same reasons discussed above, the SOM is still having the highest rate of clusters overlap and the lowest spacing between nodes. The DB index favors compact clusters, well separated from each other, which, as already said, is the opposite way in which the self-organizing maps forms the clusters. However, from a biological point of view, it would be useful to have groups 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 can participate in several pathways, simultaneously.

4.4.2 Comparison based on biological assesment criteria.

To assess the significance of the clusters from the viewpoint of their biological meaning, a detailed inspection to verify the membership of the patterns grouped in integration nodes (clusters grouping both metabolites and transcripts data types) to any known metabolic pathway should be performed. In the example given here, only integration nodes found by the three methods have been analyzed.

For this analysis, well known metabolic pathways occurring in tomato fruits have been considered[7], related to energy production (glycolysis and TCA cycle) and few associated reactions, due to their importance in all living organisms and the large amount of available data. Furthermore, except in few cases, the choice of biological processes common to the vast majority of organisms is an important starting point for comparison, because it is assumed that any clustering method used to analyze these data should be able to find such relations.

Figure 8 shows a simplified diagram of the metabolic pathways found for a preliminary analysis with $k = 50$. For the transcripts, the EC codes have been used, corresponding to the standard nomenclature for enzymes. The metabolites and transcripts that are part of the pathway and that are present in the training set have been highlighted with a rectangle and highlighted in bold, respectively. The remaining compounds (in italics) will not be taken into account in this analysis because they have not been measured. In this figure, it is highlighted the number of integrative node in which each compound has been found, distinguishing between the SOM method (node number to the right) and IL-KM method (number to the left). In the case of enzymes that are encoded by more than one gene, all the the nodes in which each gene has been clustered are indicated. To simplify the notation in the following analysis, the compounds that have been grouped together in the same integration node are indicated between brackets $[\cdots]$.

IL-KM has found coherent relationships, but scattered through a different number of integration nodes. For example, this can be observed in the following nodes: [glucose and F6P], [succinate and fumarate], [serine and GABA (4-aminobutyrate)], [EC 4.2.1.2 and 1 gene from EC 4.1.1.31], [malate and 1 gene from EC 1.1.1.1], and [ascorbate and EC 1.1.1.29].

However, the associations do not always reflected the opposite relationships, such as the case of [maltose and glucose], [glutamate and EC 1.1.1.29], [fumarate and EC 4.2.1.2], among others, where for a sign configuration there have been gropued together in the same node, but when the sign is inverted, they have been splitted into different groups. These inconsistencies, if not so significant as in the case of IL-HC, are a limit to the method and throws doubts over the method regarding its applicability to these kind of biological data, specially when looking for unknown relationships among them.

Even when IL-SOM has generated half the number of integrators nodes than IL-KM (as

[7] LycoCyc: http://solcyc.sgn.cornell.edu/LYCO/server.html

can be seen in Table 3), differently from this last one, the patterns associations has been consistent as much for the directed as for the inverted sign cases, and the clusters clearly have associated more compounds in the same pathway in less integration nodes. This has been the case of [maltose, glucose, fructose, F6P, alanine, glycerol 3P, EC 1.1.1.27, EC 4.2.1.2 and 1 gene from EC 4.1.1.31] and of [citrate, glutamate, succinate, malate and sacarose].

The IL-SOM model allows, also, analyzing relationships with different neighborhood radius in the IL-SOM map [Stegmayer et al., 2009], which offers an extra level of analysis in relation to the other methods. If the first neighbors of each neuron are considered (that is to say, $Vn = 1$), another relationships of interest (for the pathway under analysis) can be found, such as [serine, glicerate, GABA (4-aminobutyrate) and EC 4.1.1.31][8], [EC 1.1.1.29 and EC 1.1.3.15][9], and [fumarate and 2 genes from EC 1.1.1.1][10]. In the first group, compounds that have been grouped by IL-KM but not by IL-SOM with neighborhood radius = 0, can be found. Additionally, both genes that codify for the EC 1.1.1.1 enzyme have been now grouped in the same cluster.

A deeper analysis has been performed over the available data, considering now $k = 625$, which correspondas to a IL-SOM of 25 × 25 neurons. Figure 9 shows a schematic diagram of the re-constructed pathways by analyzing, in the new map, neurons 25, 601 and 625 and their neighbors. These neurons were choosen because they grouped most of the glycolytic and TCA cycle intermediates and enzyme enconding genes connecting these pathways with amino acids metabolism. This IL-SOM model, similarly to the presented above, also groups known biochemichal pathways components in few neurons. In many cases, what has been grouped in the same neuron in the IL-SOM map of 20 × 20 neurons, now appears more spread because of the new size of the map and, as previously mentioned, wider visualization radius should be used. As can be seen from the figure, most of the pathway components are grouped in neuron 625 or its neighbors (599[11], 575[12], 621[13]). Conversely, other well characterized intermediates of the TCA cycle were not found in neighboring neurons such as the case of 2-oxoglutare (grouped in neuron 531). This could either be interpreted as a limitation of the

[8] The neighbors of neuron 2 are neurons 1 and 3.

[9] Neurons 27 and 28 are neighbors with $Vn = 1$.

[10] Neurons 31 and 22 are neighbors with $Vn = 1$.

[11] Neurons 625 and 599 are neighbors with $Vn = 2$.

[12] Neurons 625 and 575 are neighbors with $Vn = 2$.

[13] Neurons 625 and 621 are neighbors with $Vn = 4$.

"guilt-by-association" assumption or as a restriction in the SOM model and surely further investigations are needed to elucidate this point.

Furthermore, this model allows the discovery of a new relationship such as the case of the *glutaredoxin* (*EC 1.20.4.1*) encoding gene found in neuron 602[14]. The function of this enzyme is widely described in the literature in many organisms [Holmgren, 1989], but its metabolic linkage with primary carbon metabolism is not known in tomato fruits. This example highlights the potential of the method used here to propose experiments in order to test new emerging hypotheses.

In comparison, IL-KM (as IL-SOM) has grouped these TCA fundamental compounds (*malate* and *2-oxoglutarate*) in different clusters, and similarly to the previous analysis, the patterns are scattered all over the cluster space. In the figure, only IL-KM coherent groupings are indicated with its corresponding cluster number on the left for the following metabolites [fructose and glucose]. On all other cases, the metabolites and enzimes are spread over several different clusters.

5 Conclusions and perspectives

This chapter has presented a pipeline for data integration and discovery of a-priori unknown relationships among introgression lines transcriptional and metabolic data. The pipeline includes four steps: 1) IL-Data understanding, 2) Pre-processing, selection and normalization, 3) Integration, IL-mining and visualization; and 4) Clusters evaluation and biological assessment. Each step of the proposed methodology has been explained in detail, through a case study, which involved genes microarray measurements and metabolite profiles from tomato fruits.

In this pipeline, two standard clustering methods (hierarchical clustering and k-means) are compared against a novel neural network approach named IL-SOM, oriented towards IL-data mining, also providing simple visualizations for identification of co-expressed genes and co-accumulated metabolites. The methods have been compared through objective measures, to analyze the quality of the found clusters. Moreover, a way of measuring their biological significance has been proposed as well, which was addressed from the perspective of the use-

[14] Neurons 601 and 602 are neighbors with $Vn = 1$.

fulness of the groupings to identify those patterns that change in coordination and therefore belong to common pathways of metabolic regulation. The IL-SOM model has shown high performance in most objective quality measures, plus the maximum coherence from the viewpoint of the biological significance of the relationship between metabolites and transcripts obtained.

As future work, the proposed pipeline will be extended to allow finding new relationships within known metabolic pathways. This will be done by checking integration nodes against online available metabolic pathways (for example, the Kyoto Encyclopedia of Genes and Genomes[15]) for finding candidates genes and metabolites belonging to new metabolic pathways.

[15] www.genome.jp/kegg/

References

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

[Baxter et al., 2005] Baxter, C. J., Sabar, M., Quick, W. P., and Sweetlove, L. J. (2005). 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.*, 56:1591–1604.

[Bino et al., 2004] 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., and Sumner, L. (2004). Potential of metabolomics as a functional genomics tool. *Trends Plant Sci*, 9(9):418–425.

[Brandes, 2008] Brandes, U. (2008). Social network analysis and visualization. *IEEE Signal Processing Magazine*, 25(6):147–151.

[Carrari et al., 2006] Carrari, F., Baxter, C., Usadel, B., Urbanczyk-Wochniak, E., Zanor, M.-I., Nunes-Nesi, A., Nikiforova, V., Centero, D., Ratzka, A., Pauly, M., Sweetlove, L. J., and Fernie, A. R. (2006). 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.*, 142:1380–1396.

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

[Chakrabarti et al., 2009] Chakrabarti, S., Cox, E., Frank, E., Gting, R., Han, J., Jiang, X., Kamber, M., Lightstone, S., Nadeau, T., Neapolitan, R., Pyle, D., Refaat, M., Schneider, M., Teorey, T., and Witten, I. (2009). *Data Mining. Know it All.* Elsevier.

[Davies and Bouldin, 1979] Davies, D. and Bouldin, D. (1979). A cluster separation measure. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1(4):224–227.

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

[Forgy, 1965] Forgy, E. (1965). Cluster analysis of multivariate data: efficiency vs. interpretability of classifications. *Biometrics*, 21(1):768–780.

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

[Guillet and Hamilton, 2007] Guillet, F. and Hamilton, H. J., editors (2007). *Quality Measures in Data Mining*, volume 43 of *Studies in Computational Intelligence*. Springer.

[Halkidi, 2001] Halkidi, M. e. a. (2001). On clustering validation techniques. *Journal of Intelligent Informations Systems*, 17(1):107–145.

[Handl et al., 2005] Handl, J., Knowles, J., and Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212.

[Haykin, 2007] Haykin, S. (2007). *Neural Networks: A Comprehensive Foundation (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[Hirai et al., 2005] Hirai, M., Klein, M., Fujikawa, Y., Yano, M., Goodenowe, D., Yamazaki, Y., Kanaya, S., Nakamura, Y., Kitayama, M., Suzuki, H., Sakurai, N., Shibata, D., Tokuhisa, J., Reichelt, M., Gershenzon, J., and Saito, K. (2005). Elucidation of gene-to-gene and metabolite-to-gene networks in arabidopsis by integration of metabolomics and transcriptomics. *J Biological Chemistry*, 280(27):25590–25595.

[Hirai et al., 2004] Hirai, M. Y., Yano, M., Goodenowe, D. B., Kanaya, S., Kimura, T., Awazuhara, M., Arita, M., Fujiwara, T., and Saito, K. (2004). 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*, 101:10205–10.

[Holmgren, 1989] Holmgren, A. (1989). Thioredoxin and glutaredoxin systems. *J. Biol. Chem.*, 264:13963–13966.

[Kaever et al., 2009] Kaever, A., Lingner, T., Feussner, K., Gbel, C., Feussner, I., and Meinicke, P. (2009). Marvis: a tool for clustering and visualization of metabolic biomarkers. *BMC Bioinformatics*, 10:92+.

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

[Kelemen et al., 2008] Kelemen, A., Abraham, A., and Chen, Y. (2008). *Computational Intelligence in Bioinformatics*. Springer.

[Kohonen, 1982] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

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

[Lacroix et al., 2008] Lacroix, V., Cottret, L., Thebault, P., and Sagot, M.-F. (2008). An introduction to metabolic networks and their structural analysis. *IEEE Transactions on computational biology and bioinformatics*, 5(4):594–617.

[Larose, 2005] Larose, D. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley-Interscience.

[Lindon et al., 2007] Lindon, J. C., Nicholson, J. K., and Holmes, E., editors (2007). *The Handbook of Metabonomics and Metabolomics*. Elsevier.

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

[Mingoti and Lima, 2006] Mingoti, S. A. and Lima, J. O. (2006). Comparing som neural network with fuzzy c-means, k-means and traditional hierarchical clustering algorithms. *European Journal of Operational Research*, 174(3):1742–1759.

[Olson and Delen, 2008] Olson, D. and Delen, D. (2008). *Advanced Data Mining*. Springer.

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

[Quackenbush, 2001] Quackenbush, J. (2001). Computational analysis of microarray data. *Nat Rev Genet*, 2(6):418–427.

[Rieseberg and Wendel, 1993] Rieseberg, L. and Wendel, J. (1993). *Introgression and its consequences in plants*, volume 1. Oxford University Press.

[Saito et al., 2008] Saito, K., Hirai, M. Y., and Yonekura-Sakakibara, K. (2008). Decoding genes with coexpression networks and metabolomics - majority report by precogs. *Trends in Plant Science*, 13:36–43.

[Stegmayer et al., 2009] Stegmayer, G., Milone, D., Kamenetzky, L., Lopez, M., and Carrari, F. (2009). Neural network model for integration and visualization of introgressed genome and metabolite data. In *International Joint Conference on Neural Networks*.

[Tasoulis et al., 2008] Tasoulis, D., Plagianakos, V., and Vrahatis, M. (2008). *Computational Intelligence in Bioinformatics*, volume 94 of *Studies in Computational Intelligence*. Springer.

[Tokimatsu et al., 2005] Tokimatsu, T., Sakurai, N., Suzuki, H., Ohta, H., Nishitani, K., Koyama, T., Umezawa, T., Misawa, N., Saito, K., and Shibata, D. (2005). Kappa-view: a web-based analysis tool for integration of transcript and metabolite data on plant metabolic pathway maps. *Plant Physiology*, 138(3):1289–1300.

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

[Wolfe et al., 2005] Wolfe, C. J., Kohane, I. S., and Butte, A. J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, 6.

[Xu and Donald C. Wunsch, 2009] Xu, R. and Donald C. Wunsch, I. (2009). *Clustering*. Wiley and IEEE Press.

[Yano et al., 2006] Yano, M., Kanaya, S., Altaf-Ul-Amin, M., Kurokawa, K., Hirai, M. Y., and Saito, K. (2006). Integrated data mining of transcriptome and metabolome based on bl-som. *Journal of Computer Aided Chemistry*, 7:125–136.

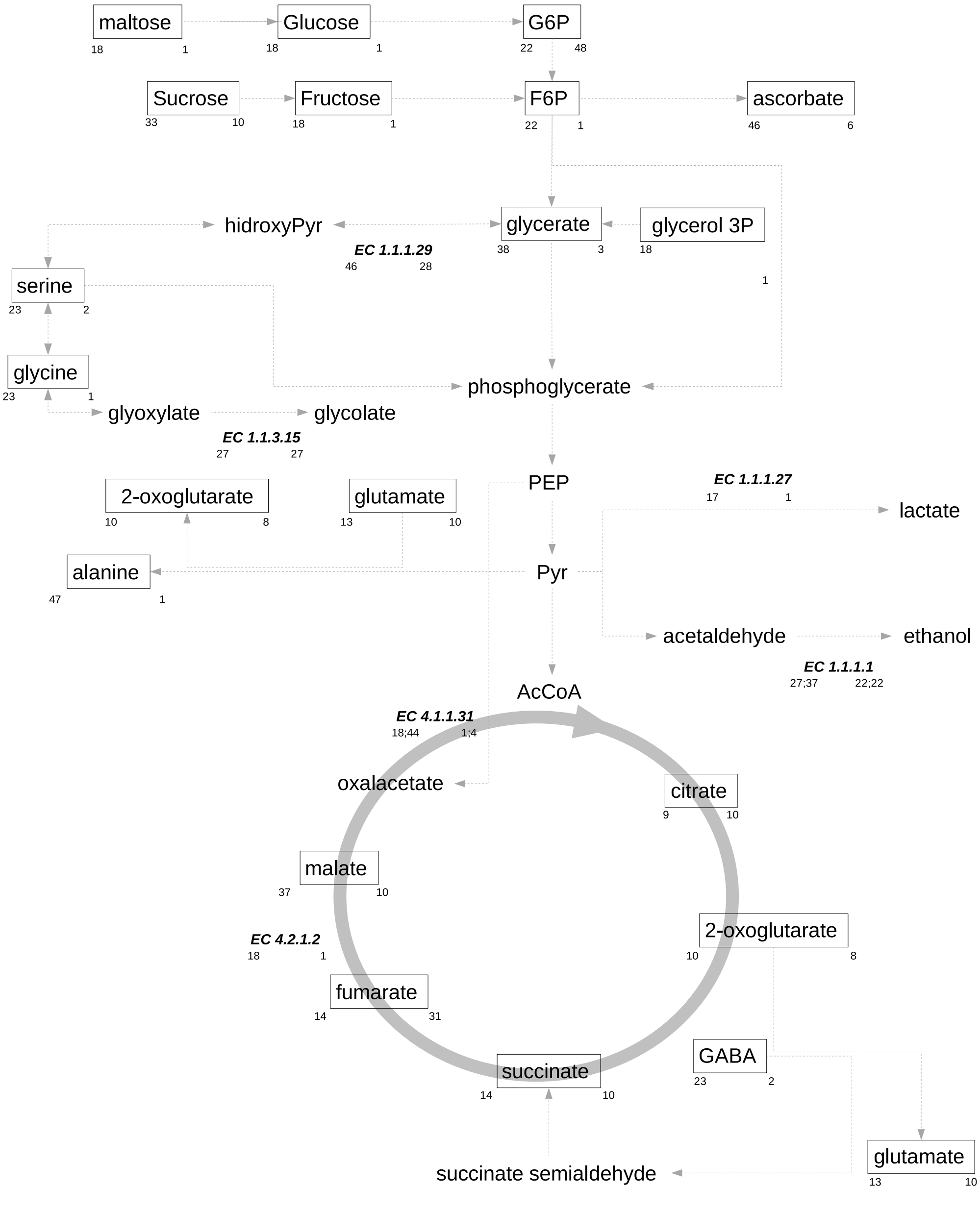


Figure 8: Simplified schematics of glycolisis, TCA cycle and associated reactions. The "EC" codes correspond to standard enzyme coding.

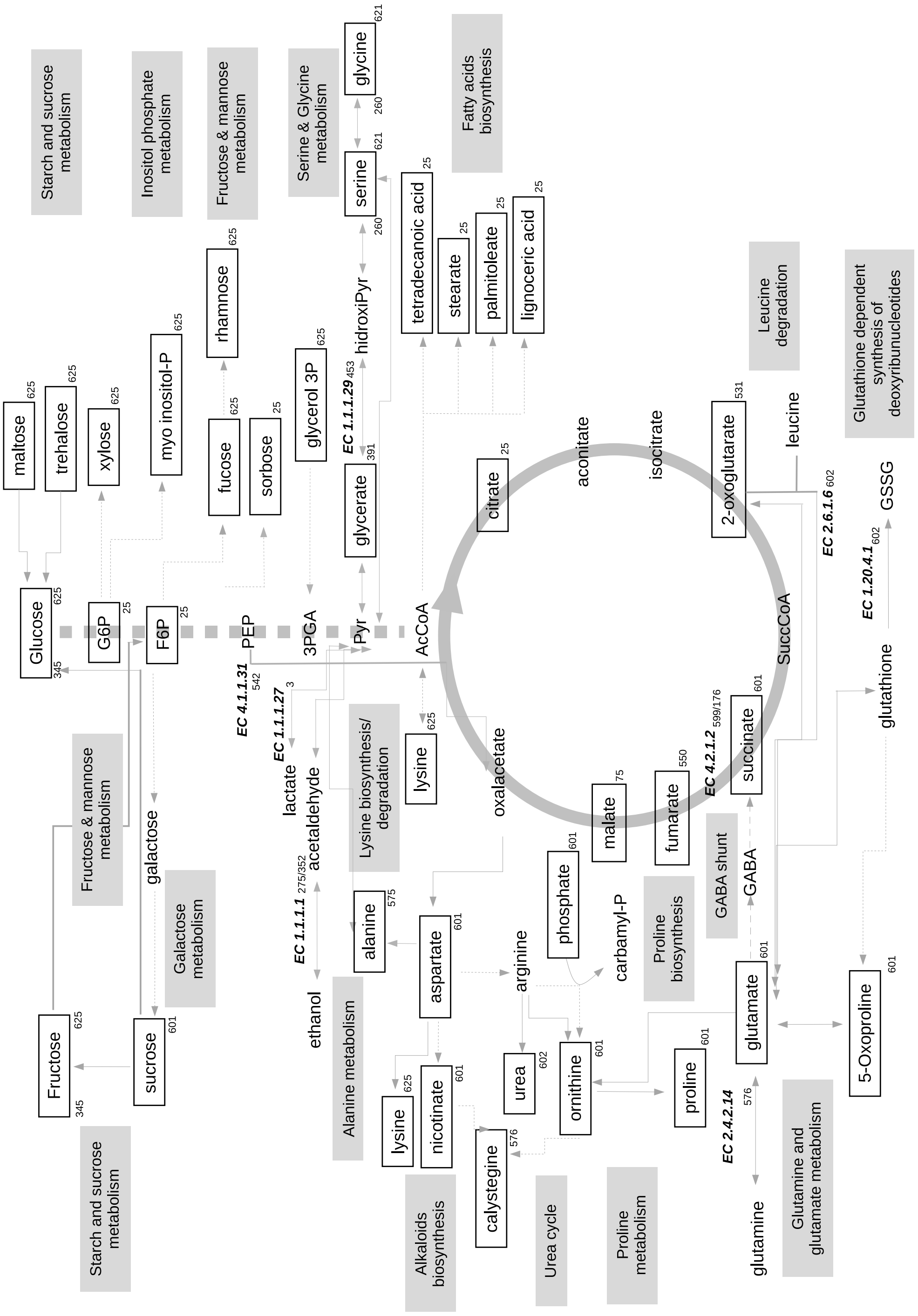


Figure 9: Primary carbon metabolism in tomato fruits. The glycolytic and tricarboxilic acid (TCA) pathways are represented. Full arrows indicate single reactions and dotted arrows represent multiple chemical reactions. Associated biological pathways are highlighted in gray boxes.