STRATEGY FOR THE DETECTION
OF FUNCTIONAL CLUSTERS OF GENES
USING DATA MINING TECHNIQUES

I. INTRODUCTION

Large amounts of gene expression data obtained with diverse techniques are available nowadays. Thus, it is necessary to find and implement methods to analyze these data and understand the biological processes involved [4]. As a first step, data mining techniques such as clustering algorithms have been used for grouping genes with similar expression patterns as well as experiments [5][6][7][8].

The clustering of genes could provide insights about unknown genes grouped with known genes [4][10]. In addition, the expression data may be related with biological knowledge to see how gene clusters are related to the biological knowledge. This relation is the first step in validation of the obtained clusters. In this work, we show a useful, general and flexible methodology to find gene clusters based on microarray [1] and RNA_Seq [2] expression data, highlighting relationships between genes and regulatory activities for expression experiments conducted in particular conditions of interest. In order to detect relationships between genes, we used data mining algorithms associated to biological knowledge on gene promoters. Once groups were constructed we searched for representative promoters in each group to categorize and validate the obtained results. The relationships between genes in each group can be further developed by constructing a gene-network.

II. STRATEGY

2.1 Data

The procedure presented here can be used for gene expression data, such as Microarray or RNA-Seq data.

2.2 Methodology

1. Gene Selection: A fraction of all genes for which data is available could be filtered. This fraction can be differentially expressed genes: highly expressed genes, low expressed genes, etc. In our application example, we selected genes with expression levels above the 3rd quartile.

2. Selection of similarity measure and clustering algorithm: The selection of the similarity measure depends on the clustering algorithm chosen because not all the algorithms need a measure. We propose the use of various clustering algorithms (linear and non-linear approaches).

3. Clustering of genes: To obtain the gene clusters we run the chosen algorithms several times varying the parameters in order to gain greater diversity in the results and further analysis.

4. Association and categorization of the constructed clusters with biological knowledge: In order to find patterns in the constructed clusters from the previous step, we propose to use knowledge on gene promoters of the organism under study to search for regulatory relationships and to identify those genes with similar behavior and therefore being controlled by the same promoter. Then we categorize each cluster, identifying which is the most representative promoter in each one.

5. Selection and validation of clusters: We select the most interesting clusters according to the representative promoter importance in the specific research. Then, in order to validate them, we plot them to check the relationship between the gene expression profiles in each cluster. Also we search the inter-genic regions of genes in those clusters and we identify promoters through the use of various algorithms that are designed for this purpose and verify that the genes within the same cluster are related to the same promoter, which in its turn needs to be the representative promoter in that cluster. At the end of this step, new promoters could be identified.
III. EXPERIMENTATION

In this work, we applied the methodology described on microarray and RNA_SEQ expression data obtained with pathogenicity experiments on the plant *Arabidopsis thaliana*. We selected highly expressed genes by filtering those that were above third quartile of the gene expression average. Then, we selected two linear clustering algorithms: Expectation Maximization and K-means and one non-linear algorithm: KK-means [3] and we applied them several times on the data varying the parameters. Then, we categorized the clusters associating them with gene promoters and we found the representative promoters in each cluster.

IV. RESULTS

After applying the methodology on microarray and RNA_SEQ expression data, we were able to identify interesting promoters in the constructed clusters like DPBF1.2.binding.site.motif and MYB4 binding site due its relation with pathogen resistance in plants, which was the condition of interest in this work. These representative promoters were present in the clusters of each of the clustering algorithms applied and they were interesting for their biological characteristics, i.e. response to acid abscisic and stress respectively. On the other hand, the gene expression profiles within a selected cluster showed a similar behavior and were coexpressed along the experiments in both microarray and RNA_SEQ data.

Figure 2 shows an example of the obtained results applying the k-means algorithm and varying the number of clusters on microarray data. Figure 2 (a) presents the table with three values for k and four clusters. It is possible to see the representative promoters in each cluster and its occurrence value in the cluster. Figure 2 (b) shows the gene expression profiles whose representative promoter is DC3 in one of the clusters obtained by the k-means algorithm with k=77.

![Figure 2](image-url)

**Figure 2.** Results of the algorithm k-means on microarray expression data. (a). Table with the representative promoters in four clusters, between brackets it is present its occurrence value on the gene total number in the cluster. (b) Example of gene expression profiles in one cluster with DC3, the interesting representative promoter.

V. CONCLUSIONS

We were able to observe how the methodology presented was useful to analyze expression data obtained with molecular biology techniques using data mining algorithms for the construction of gene clusters and the identification of representative promoters that were highly present in the formed clusters and were related to pathogen resistance in plants, which was directly involved in the research experiments. Moreover, the methodology is general and flexible to be applied to other organisms for which gene expression data and additional biological knowledge is available.

REFERENCES