An Experimental Study on Microarray Expression Data from Plants under Salt Stress by using Clustering Methods

Houda Fyad*, Fatiha Barigou, Karim Bouamrane

Laboratoire d'informatique d'Oran (LIO), Département d'informatique, Université Oran1 Ahmed Benbella (Algeria)

Received 9 June 2019 | Accepted 13 February 2020 | Published 27 May 2020

Abstract

Current Genome-wide advancements in Gene chips technology provide in the "Omics (genomics, proteomics and transcriptomics) research", an opportunity to analyze the expression levels of thousand of genes across multiple experiments. In this regard, many machine learning approaches were proposed to deal with this deluge of information. Clustering methods are one of these approaches. Their process consists of grouping data (gene profiles) into homogeneous clusters using distance measurements. Various clustering techniques are applied, but there is no consensus for the best one. In this context, a comparison of seven clustering algorithms was performed and tested against the gene expression datasets of three model plants under salt stress. These techniques are evaluated by internal and relative validity measures. It appears that the AGNES algorithm is the best one for internal validity measures for the three plant datasets. Also, K-Means profiles a trend for relative validity measures for these datasets.

Keywords

Clustering Methods, Clustering Validity Indices, Gene Chips Analysis, Gene Expression, Plant Datasets.

DOI: 10.9781/ijimai.2020.05.004

I. Introduction

Abiotic stresses significantly reduce agricultural productivity worldwide. Plant growth and crop productivity are affected by environmental factors, especially saline stress [1]. Therefore, it is important to know the genes implicated in tolerance to salinity [1]. Furthermore, technological advances in the field of genomics, such as DNA sequencing, have generated a wealth of genetic information [2]. Such information includes expression profile levels of thousands of genes under various experimental conditions [3]. Hence, a better biological view of the presumed gene functions can be obtained.

Therefore, a wide range of machine learning methods such as clustering have been developed [4], [5]. They are being used in a variety of applications, such as cancer diagnosis [6], pharmacovigilance [7] and plant breeding [8]. "Omics research" has thereafter relied on clustering techniques to group genes. The main objective of clustering techniques is exploring the results of DNA chips to classify and group identical expression profiles [4], identify co-expressed genes [4], find their biological functions [8], [9], and explain their regulatory mechanisms [9], [10].

In Gene chips data analysis, some classical clustering techniques were implemented. One of these is the Hierarchical algorithm commonly called UPGMA [11], [12]. It generates dendrograms and heat maps that display and intuitive visualization of genes and their relationships [12]-[15]. Other clustering methods, called Partitioning methods like K-Means, PAM and CLARA were developed. Their purpose is to partition the gene expression dataset into (k) coherent clusters with same biological characteristics [16]-[18]. Model based clustering methods, such as Self Organization Map (SOM) is also another clustering technique. Their aim is similar to K-Means and Hierarchical algorithms. The advantage of the SOM algorithm is its ability to visualize and optimize the high-dimensional data on an output map of neurons with similar gene functions [19].

Some of these methods have been combined. For instance, Hierarchical algorithm with SOM algorithm designated Self-Organizing Tree Algorithm (SOTA) [20] and Self-Dynamically Growing Self-Organizing Tree (DGSOT) [20] algorithms were developed for improving clustering performance when there is noisy data and determining a good quality of partition of the gene expression data [20], [21].

Another combination of the Hierarchical algorithm that was associated with the K-Means algorithm is called Hierarchical K-Means [22]. This combination took advantage of both algorithms. The hierarchical algorithm provided a tree structure of groups that was used by the K-Means algorithm to determine relevant and compact gene expression groups [22].

Many more algorithms were implemented in order to enhance the convergence and efficiency of the clustering result. There are for example, Fuzzy clustering [17, 23], Fuzzy clustering based on Local Approximation of MEmbership (FLAME) [23], Graph-based clustering method like MST [24], Grid-based clustering method (STING, CLIQUE) [24], Density-based clustering method (OPTCS, DBSCAN) [24], Gaussians and Spectral Clustering methods [24].

* Corresponding author.
E-mail address: houdafyad82@gmail.com

While all these algorithms have been compared in different studies, there was no clear agreement on the most appropriate clustering algorithm to be used for clustering genes with their associated expression profiles [25]-[27].

Mostly, each clustering method has its own parameters for calculating clusters. The decision to use a particular method for clustering will depend on the nature of the datasets being studied and what the researcher expects to achieve using that method [26]-[29].

Based on these considerations, we decided to conduct a comparative study of seven most commonly used clustering algorithms on gene expression datasets from three model plants under saline stress. These methods are evaluated based on both internal and relative validity measures. The main objective of this study is to address biologists' concerns about the most appropriate algorithm to be used for achieving the desired gene clustering.

The remaining of this paper is organized as follows: Section II presents an overview of clustering techniques used in gene expression. Section III, is dedicated to gene expression experiments, the choice of clustering techniques and the clustering Validity concepts. Section IV, provides the results and discussion of the performance of the respective algorithms. Finally, Section V concludes by summarizing findings and identifying possible future work.

II. Related Work

In the previous section, we have mentioned the importance of analyzing and studying gene expression data with clustering techniques. These techniques have helped to answer several biological questions.

Hierarchical algorithms (HC) are the earliest ones used in gene expression data. Eisen et al. [12] used HC for an empirical analysis to classifying and visualizing gene expression on yeast Saccharomyces cerevisiae datasets. Dendrogram tree and Heat maps are well-known HC graphic tools that illustrate the correlation of these genes.

Alizadeh et al. [13] applied the same method on Diffuse large B-cell Lymphoma (DLBCL), HC has permitted the discovery of new molecular subtypes with three different genetic signatures.

Bajsa et al. [14] focused on the determination of the transcription level of the cellular pathways on the model plant Arabidopsis. The result of the heat map with dendrogram revealed the up-regulated gene and down-regulated ones in different time courses under salt stress.

Hossen et al. [15] analyzed the clustering proximity effect on two types of gene expression datasets (Affimetrix and cDNA). The authors implemented seven Hierarchical algorithms (Single Linkage, Complete Linkage, Average Linkage, Ward, Centroid, Median, Mcquitty) according to five proximity measures (Euclidean, Manhattan, Pearson, Spearman and, Cosine). The Ward method with Cosine distance was outperforming on both types of datasets.

Takahashi et al. [16] studied gene expression of 4 varieties of Wheat, the analysis concerned different levels of salinity tolerance. K-Means Clustering algorithm optimized to 3 the number of clusters: Cluster I included genes expressed as an early response that occurred within 24 hours under control conditions. Cluster II assembled genes expressed during the second day under control conditions and Cluster III included the genes expressed in the late response that occurred on the third day. The Hierarchical clustering (with Pearson correlation and average linkage) method and Principal component analysis were used for visualization of results [16].

Gasch et al. [17], worked on the Yeast gene expression profile. K-Means and Fuzzy C-means (FCM) have established the expression profile during seven periods of the cell cycle in Yeast. They validate their results with the Davis Bouldin Index (DBI). FCM has achieved the DBI of 0.31452 for K=3 and 0.37822 for K=4 which is better than K-Means clustering.

The study conducted by Ge et al. [18], with Hierarchical clustering and K-Means methods allowed identifying eight (08) distinctive gene groups regulated by abiotic stress in Glycine soja. The authors successfully discovered the corresponding co-regulated genes and their functions.

FLAME is an extension of Fuzzy clustering based on the Local Approximation of Membership that was implemented for microarray data [23]. The advantage of the FLAME algorithm compared to the FCM algorithm is its ability to define various and homogeneous groups of genes, and to give a relevant subdivision of biological functions patterns [23].

SOM algorithm was applied to Yeast Sporulation, Human Fibroblasts Serum and Rat CNS datasets [19]. This method provides a better result for the recognition and classification of the features in complex and multidimensional datasets. Luo et al. [30] have used the SOTA algorithm to discover Transcription Factor (TF) gene families in *Medicago sativa* during ABA treatment. In that case, 82 TF genes families were distributed into four clusters with the number of genes equal respectively 15, 34, 18 and, 14.

The comparison study between the following Clustering algorithms (HC, K-Means, and SOM) was performed on Solanum tuberosum genes showing differential expression in abiotic stress [25]. The author in this study, obtains almost the same number of the clusters for these different algorithms.

López-Kleine et al. [26] applied AGNES, DIANA, K-Means (with Euclidean and Manhattan distances) and SOM for clustering the genes involved in pathogen resistance on Tomato. The results showed that AGNES, K-Means, and SOM grouped these genes into two clusters: genes implicated or no in plant resistance. DIANA was abandoned because almost all genes were assigned to one cluster.

In other comparison work [27], Hierarchical algorithms with single, complete and average linkage, K-Means, Gaussians Clustering methods (FMC), Spectral Clustering (SP) methods and a Nearest Neighbour-based methods were evaluated on 35 gene expression datasets of various cancerous tissue types. FMC and K-Means were the most appropriate methods to recover the true structure of this kind of datasets.

Singh et al. [28] assessed the efficiency of K-Means (KM), Density-based clustering (DBC) and expectation maximization (EM) methods by using the sum of squared error, log-likelihood measures. These methods were tested on SRBCT, Lymphoma and three different Leukemia datasets. The results showed that EM algorithm gives the best result with log-likelihood measurement. KM and DBC algorithms produced similar results with regards to the sum of squared error measurement.

Bihari et al. [29] compared the performance of KM, HC clustering, SOM and DBSCAN on Iris flower gene expression data. The comparison results of these methods were validated by using internal and external indices. According to the experimental analysis KM is more appropriate for gene clustering.

In Table I, we summarize some algorithms that we have mentioned in the state of the art. We will give their main characteristics with respect to the following parameters: (i) influence of noisy data, (ii) ability to work properly with large dataset and (iii) algorithm computation time.

III. Methods

This section describes the different experiences conduced to analyze and compare the clustering methods for plant genes expression

TABLE I. Characteristics of Various Clustering Algorithms

"n" is the number of points in the dataset, "k" is the number of clusters, "l" is the number of iterations, "m" is the number of initial sub-clusters produced by the graph partitioning algorithm.

Clustering method	Algorithm	Type of data	Sensitive to noisy data	Dealing with high dimensional data	Scale	Computational time
Hierarchical	**AGNES**: a bottom-up approach [30], [44].	Numerical	Not very sensitive	No	NA	$O(n^2)$
	DIANA: a top-down approach [31], [44].	Numerical	Not very Sensitive	Yes	NA	$O(2^n)$
	BIRCH: agglomerative hierarchical based clustering algorithm [32], [44].	Numerical	Very Insensitive	No	Yes	$O(n)$
	CURE: has been developed to handle a huge volume of data, insensitive to outliers and capable of working with clusters of different shapes and sizes [33], [44].	Numerical	Insensitive	Yes	Yes	$O\ (n^2+nm_m m_a+\ n^2 log\ n)$
	ROCK: uses the concept of the number of links between two records to assess the similarity of the categorical attributes of the dataset [34], [44].	Categorical	Not very Sensitive	No	Yes	$O\ \ (n^2 log\ n)$
	CAMELEON: based on a dynamic model for merging clusters. It calculates the interconnectivity and the proximity of two clusters in order to discover the similarity between them [35], [44].	Numerical⁄ Categorical	NA	No	Yes	NA
Partitioning	**K-MEANS**: is a method which aims to divide the dataset elements into groups that are well separated from each other [36], [44].	Numerical	Sensitive	No	Yes	$O(Ikmn)$
	PAM: algorithm aims to find a sequence of objects called medoids that are located in the center of clusters. It is a more robust partitioning algorithm against outliers than the k-means partitioning algorithm [37], [44].	Numerical	Not very Sensitive	No	Yes	$O(k(n-k)^2)$
	CLARA: was developed in order to deal with large datasets. It does not work with the whole set of data, but with a small portion of the data which is chosen randomly [38], [44]	Numerical	Not very Sensitive	No	Yes	$O\ (k(40+k)^2+\ k(n-k))$
	CLARANS: is an extension of the CLARA algorithm. It is a combination of sampling techniques with the PAM algorithm [39], [44].	Numerical	Sensitive	No	Yes	$O(kn^2)$
Model based	**SOM**: consists in projecting the large data space observed on a 2 or 3 dimensional space called a map. This map is composed of groups of neurons connected together according to the concept of neighborhood [40], [44].	Numerical	Not very sensitive	Yes	Yes	$O(I)$
Fuzzy based	**FCM**: allows assigning an element to one or more clusters [41], [44].	Numerical	Sensitive	No	Yes	$O(n)$
Grid based	**CLIQUE**: finds clusters in subspaces of high density data [42], [44].	Numerical	Not very Sensitive	No	Yes	$O(n+k^2)$
Density based	**DBSCAN**: groups in the neighborhood of a point having a given radius (ε) a minimum number of points (MinPts) [43], [44].	Numerical	Very insensitive	No	No	$O(m\ log\ m)$

(NA) information is not mentioned by authors.

data. The proposed workflow is described in Fig. 1

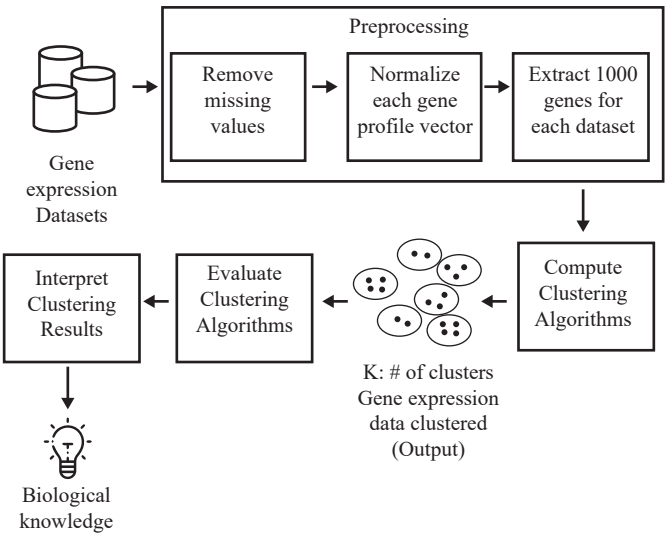

Fig. 1. Flowchart of the Experimental Process.

A. Datasets Selection

Three datasets of expression data relating to plants *Arabidopsis thaliana*, *Solanum lycopersicom* (Tomato) and *Medicago truncatula* under salt stress were considered. In this experimental study, we choose to work with these datasets because they are based on model plants. *Arabidopsis thaliana* is regarded to be the first most studied and investigated model plant. *Solanum lycopersicom* and *Medicago truncatula* are also model plants, each representing a family of plant species. Datasets for these model plants cover a broad spectrum of gene expressions.

1. Dataset 1: Arabidopsis Thaliana (A. Thaliana) Salt Stress

This dataset describes the salt stress experiment of model *Arabidopsis thaliana* leaves using Affymetrix Array, 2 samples of leaves from 3 genotypes of *A. thaliana* with and without 100 mM NaCl. This dataset shows the salt-stress influence on leaves from these 3 genotypes. The experiment results explain a global change on related genes and provide an insight into the molecular mechanisms underlying variation in salt stress responses [45]. The *Arabidopsis thaliana* dataset was downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16765.

2. Dataset 2: Solanum Lycopersicom (Tomato) Salt Stress

This dataset describes the salt stress experiment of an old Tomato leaves using Affymetrix Array, 6 samples of leaves-old with 200 mM NaCl for 5 h, 6 samples of leaves-old without 200 mM NaCl for 5 h. This dataset compares the salt-stress influence analysis on leaves-old from 2 genotypes of Tomato. The experiment results that the Wild tomato genotype is significantly more salt-tolerant than a Cultivar, *Solanum lycopersicom* [46]. The *Solanum lycopersicom* dataset was downloaded from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16401.

3. Dataset 3: Medicago Truncatula (M. Truncatula) Salt Stress

This dataset describes the time-course salt stress experiment of model legume *Medicago truncatula* roots using Affymetrix Array, 6 samples of Medicago truncatula seedlings grew in two weeks in hydroponics media with 200mM NaCl salt stress at 0, 6, 24, 48 hours, 12 samples other of Medicago truncatula seedlings 3 days Petri dishes with 180mM NaCl salt stress at 0, 1, 2, 5, 10, 24 hours. This dataset reveals the salt stress effect on Medicago truncatula seedlings [47]. The *Medicago truncatula* dataset was downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14029.

These datasets are retrieved from the Gene Expression Omnibus database [48]. Table II gives all the information concerning these three datasets.

B. Preprocessing

As shown in Fig. 1, before gene clustering, it is necessary to pre-process datasets (removing missing values). And then, every gene vector is normalized according to whether its mean is equal to 0 and its standard deviation has a value of 1 [49]. For more homogeneity with the number of genes present in the Tomato dataset, only 1000 genes of the *Arabidopsis thaliana* and the *Medicago truncatula* datasets were randomly extracted for analysis. They were randomly selected to eliminate selection bias. This selection is due to the fact that the number of genes annotated on tomato is less important than for *Arabidopsis thaliana* and the *Medicago truncatula* in this type of dataset.

C. Clustering

For analyzing and evaluating the three datasets cited before, we used the open-source R environment which contains a variety of functions for data clustering. Among the clustering algorithms, we have chosen seven one: Hierarchical algorithms (AGNES, DIANA), Partitioning algorithms (K-MEANS, PAM and CLARA), Fuzzy Clustering (FANNY). These categories of methods are functions defined in R package named "cluster". Model-based Clustering (SOM) in this category of methods depends on R packages "kohonen" and "mclust". All these seven methods are contained in the R package named "clValid" that includes some validity measures that we used for testing our three datasets. These algorithms are the most commonly used, as the time complexity is low and they offer an easy interpretation of results by biologists. The code source link of each clustering method used and their validation is: http://github.com/Projet-82/New-Project/blob/master/Clustering_eva-lunation_codesource.R.

TABLE II. Dataset Description

Data set	#Genes	#Samples	Genotypes	Salt-Stress concentration	Time points
1	**A. thaliana_salt stress**				
	15 288	18	6 Ws	0 mM NaCl 100 mM NaCl	NA
			6 Col	0 mM NaCl 100 mM NaCl	NA
			6 Col(gl)	NaCl 100 mM NaCl	NA
2	**Tomato_salt stress**				
	1 000	12	6 Money maker	0 mM NaCl 200 mM NaCl	5 hours
			6 PI365967	0 mM NaCl 200 mM NaCl	5 hours
3	**M. truncatula_salt stress**				
	2 394	18	6 NA	180 mM NaCl	0, 1, 2, 5, 10, 24 hours
			12 NA	200 mM NaCl	0, 6, 24, 48 hours

D. Evaluation

To evaluate and compare the clustering algorithms, we consider two important concepts: Cohesion and separation. The Cluster cohesion measures how closely related are objects in a cluster. [50]. Cluster separation measures how distinct a cluster is from other clusters [50].

For this study, the following measures are used to assess the quality and consistency of the clusters on terms of the cohesion and separation of clusters resulting from different clustering algorithms: Connectivity index [50], [51], Dunn index [50], [51], and Silhouette coefficient [50], [51]. These 3 measurements are called internal measures.

In the other hand, the stability measures compare the results from clustering based on the full data to clustering based on removing each column, one at a time. These 4 measures work especially well if the data are highly correlated, which is often the case in high-throughput genomic data. They included Average proportion of non-overlap (APN) [50], [51], Average distance (AD) [50], [51], Average distance between means (ADM) [50], [51], and the figure of merit (FOM) [50], [51]. These lasts are called relative measures.

1. Connectivity Index

It measures how much neighbouring data points have been ranked in the same cluster [50], [51]. It is calculated by the following formula:

$$Conn(c)=\sum_{i=1}^{N}\sum_{j=1}^{L}x_{i,nn(j)} \tag{1}$$

With

$$x_{i,nn_{i(l)}}=\begin{cases}\frac{1}{j}, if \nexists c_k : i\in c_k \wedge nn_{i(j)}\in c_k\\ 0, otherwise\end{cases} \tag{2}$$

Where:

"K" is the total number of clusters.

"N" is the total number of rows (observations).

$nn(j)$ is the j^{th} nearest neighbour of the data point.

"L" is the parameter determining the number of neighbours that contribute to connectivity measure. Connectivity should be minimal.

2. Dunn Index

Dunn's goal is to identify dense and well-isolated clusters. It describes the proportion between the minimum and the maximum distances separating the clusters [50], [51]. It is computed by the following formula:

$$D=\frac{\min_{1\le i\le j\le n} d(i,j)}{\max_{1\le i\le j\le n} d'(k)} \tag{3}$$

Where:

$d(i\,j)$ describes the two cluster's distance i and j.

$d'(k)$ measures the intra-group distance of cluster k.

$d(i, j)$ is the inter-group distance. In this case the distance corresponds to the centroids distance.

3. Silhouette Coefficient

The silhouette width coefficient defines the compactness based on the paired distance between all items in the cluster, and the separation based on paired distance between all items on the cluster and all items in the nearest cluster [50], [51]. The Silhouette score is defined as:

$$S(i)=\frac{b(i)-a(i)}{\max\{a(i)-b(i)\}} \tag{4}$$

Where:

$a(i)$ is the average distance of gene i to other genes in the same cluster. $b(i)$ is the average distance of gene i to genes in its nearest neighbour cluster. The average of $S(i)$ across all genes reflects the overall quality of the clustering result.

4. Average Proportion of Non-overlap (APN)

The APN measure calculates the average proportion between observations that are not affected in their similar cluster by grouping together the complete data and grouping together the data with one column removed [50], [51]. The APN measure is denoted as follows:

$$APN(K)=\frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}\left(1-\frac{n(C^{i,l}\cap C^{i,0})}{n(C^{i,0})}\right) \tag{5}$$

Where:

"K" is the total number of clusters.

"M" is the total number of columns (attributes)

"N" is the total number of rows (observations).

"$n(C^{i,0})$" represents the cluster that contains observation i using the original clustering (based on all available data).

"$C^{i,l}$" represents the cluster that contains observation i where the clustering is based on the dataset with column removed.

5. Average Distance (AD)

The mean distance between observations that are not assigned in a similar cluster by grouping based on complete data and grouping based on data with one column deleted is estimated by the AD measure [50], [51] which is denoted as follows:

$$AD(K)=\frac{1}{NM}\sum_{i=1}^{N}\sum_{l=1}^{M}\frac{1}{n(C^{i,l}\cap C^{i,0})}\left[\sum_{i\in C^{i,0},j\in C^{i,l}}\left(dist(g_i,g_j)\right)\right] \tag{6}$$

Where:

$dist(g_i, g_j)$ is a distance (e.g. Euclidean, Manhattan, etc.) between two expression genes profiles i and j.

"K" is the total number of clusters.

"M" is the total number of columns (attributes).

"N" is the total number of rows (observations).

"$n(C^{i,0})$" represents the cluster that contains observation i using the original clustering (based on all available data).

"$C^{i,l}$" represents the cluster that contains observation i where the clustering is based on the dataset with column removed.

6. Average Distance between Means (ADM)

The ADM measure calculates the mean distance between cluster centers that are not assigned in a similar cluster by grouping based on complete data and grouping based on data with one column [50], [51]. The ADM measure is denoted as follows:

$$ADM(K)=\frac{1}{NM}\sum_{i=1}^{N}\sum_{l=1}^{M}dist(\bar{x}_{c^{i,l}},\bar{x}_{c^{i,0}}) \tag{7}$$

Where:

"M" is the total number of columns (a collection of samples, time points…).

"N" is the total number of rows (observations).

$\overline{x}_{C^{i,0}}$ is the mean of the observations in the cluster which contains observation i, when clustering is based on the full data.

$\overline{x}_{C^{i,l}}$ is the mean of the observations in the cluster which contains observation i, when clustering is based on the dataset with column removed. Currently, ADM only uses the Euclidean distance.

7. Figure of Merit (FOM)

The intra-cluster mean variance of the suppressed column observations is computed by the FOM measurement, the resulting classification is being based on the remaining samples (not cleared). This estimates the average error using predictions based on cluster averages [50], [51]. For a particular left-out column *l*, the FOM is:

$$FOM(l,K) = \sqrt{\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in C_k(l)} dist(x_{i,l},\overline{x}_{C_k(l)})} \quad (8)$$

Where:

"K" is the total number of clusters.

"N" is the total number of rows (observations).

$x_{i,l}$ is the value of the i^{th} observation in the l^{th} column in the cluster.

$\overline{x}_{C_k(l)}$ is the average of the cluster $C_k(l)$. Currently, the only distance available for FOM is Euclidean.

IV. Results & Discussion

In this section, we comment on and discuss the results obtained. Tables III, IV and V describe the various performance measures and validity indices corresponding to the three best clustering algorithms applied on our three datasets. The rest of the clustering algorithms are not shown.

A. Dataset 1: A. Thaliana Salt Stress

It can be seen from the results of Table III that in the case of Hierarchical algorithms, AGNES gives, for an optimal number of clusters **K=4**, high performance with a lower Connectivity Index equal to **25.049**, best Silhouette index score equal to **0.488** and with best Dunn index value score equal to **0.228**. This algorithm performance is followed by K-Means and DIANA algorithms for the Dunn index, with a value of **0.0576** (resp. **0.0579**) for a cluster number equal to 10. SOM is less rated than DIANA by the index of Silhouette which is worth **0.2818** (resp. **0.5068**) for **K = 2**. Also, we found that FANNY algorithm does not provide any results because of its inability to generate measurable clusters.

On the other side, the cluster stability measures describes that the Model based Clustering algorithm, SOM presents high performance too with AD measure value equal to **3.6581** and FOM measure value equal to **0.6988** for **K = 10**.

AGNES, followed by DIANA gives good performance with an APN value equal to **0.0047** (respectively **0.0056**). And on the contrary, DIANA is followed by AGNES with an ADM value equal to **0.0518** (resp. **0.0677**) for **K = 2**. PAM is better than K-Means in AD measure with a value equal to **3.7822** (or **3.8037**). K-Means is better than PAM in FOM measurement with a value equal to **0.7083** (respectively **0.7196**) for **K=10**.

B. Dataset 2: Tomato Salt Stress

From the results presented in Table IV, we can observe that concerning Hierarchical Clustering algorithms, for the optimal numbers of clusters **K=4**, AGNES gives high performance with lower Connectivity index value equal to **25.128**, best Silhouette index score equal to **0.7229** and with best Dunn index value score equal to **0.124**. This performance is followed by the DIANA algorithm for the Dunn and Silhouette indices, whose value is **0.0648** (resp. **0.7161**) for a cluster number equal to 4 (resp.2). SOM is lower than DIANA, for the Silhouette index is worth **0.7161** (resp. **0.7122**) for **K = 2**.

TABLE III. Evaluation of the 3 Best Clustering Techniques on A. Thaliana Salt Stress Dataset

Algorithm rank		1	2	3
		Algorithm [parameter K] score		
Internal validation	Conn. index	AGNES[K=4] 25.048	DIANA[K=2] 37.637	K-Means[K=2] 147.272
	Dunn index	AGNES[K=4] 0.2281	K-Means[K=10] 0.0579	DIANA[K=10] 0.0576
	Silhouette index	AGNES[K=4] 0.4880	DIANA [K=2] 0.5068	SOM[K=2] 0.2818
Relative validation	APN measure	AGNES[K=2] 0.0047	DIANA[K=2] 0.0056	K-Means[K=2] 0.0236
	AD measure	SOM[K=10] 3.6581	PAM [K=10] 3.7822	K-Means[K=10] 3.8037
	ADM Measure	DIANA[K=2] 0.0518	AGNES[K=2] 0.0677	K-Means[K=2] 0.1045
	FOM measure	SOM[K=10] 0.6988	K-Means[K=10] 0.7083	K-Means[K=10] 0.7196

On another side, the relative measures show that the partitioning Clustering algorithm, PAM produces high performance too with AD measure equal to **0.992** and FOM measure equal to **0.324** for number of clusters **K=10**, DIANA as well performed with APN measure equal to **0.0076** followed by FANNY and K-Means values equal to **0.0111** (resp. **0.0116**) for an optimal number of **K = 2**. PAM done a good result with an AD measure score equal to 0.9924 and FOM with a score equal to **0.3241** for **K=10**.

CLARA obtains a good result too with AD and ADM measures value equal to **1.0751** (resp. **0.0817**) for **K=10**. FANNY presents a good performance with an APN and AD measures with values equal to 0.0111 (resp. **1.0690**) for **K=2** (resp. **K=10**) and K-Means presents the same behavior for ADM and FOM values equal to **0.0804** (resp. **0.3280**) for the same number of clusters.

TABLE IV. Evaluation of the 3 Best Clustering Techniques on Tomato Salt Stress Dataset

Algorithm rank		1	2	3
		Algorithm [parameter K] Score		
Internal validation	Conn. index	AGNES[K=4] 25.128	CLARA[K=2] 30.249	K-Means[K=2] 35.0825
	Dunn index	AGNES[K=4] 0.1239	DIANA[K=4] 0.0648	CLARA[K=10] 0.0353
	Silhouette index	AGNES[K=4] 0.7229	DIANA[K=2] 0.7161	SOM[K=10] 0.7122
Relative validation	APN measure	DIANA[K=2] 0.0076	FANNY [K=2] 0.0111	K-Means[K=2] 0.0116
	AD measure	PAM[K=2] 0.9924	FANNY[K=10] 1.0690	CLARA[K=10] 1.0751
	ADM measure	FANNY[K=2] 0.0538	K-Means[K=2] 0.0804	CLARA[K=10] 0.0817
	FOM measure	PAM[K=10] 0.3241	K-Means[K=10] 0.3280	SOM[K=10] 0.3290

C. Dataset 3: M. Truncatula Salt Stress

From the results reported in Table V, the Hierarchical Clustering algorithms, for the optimal numbers of clusters **K = 4**, AGNES presents high performance with lower Connectivity index value equal to **2.9290**, a best Dunn and with best Silhouette index of **0.8296** (resp. **0.9587**). This performance is followed by the DIANA method with Connectivity and Silhouette indices equal to **5.2869** and. **0.9406**, respectively, for a cluster number equal to **2**.

On the other side, the relative stability describes that DIANA as well performed with a value of APN measure equal to **0.0001**. This performance is followed by AGNES and CLARA with values equal to **0.0007** (resp. **0.0053**) and the same behavior with the inverse ordered algorithms is shown with ADM measure equal respectively for CLARA **0.0288** and AGNES **0.0307** with a cluster number of **K = 2**. K-Means performs in FOM measure with a value equal to **0.3487** followed by PAM and SOM with a value equal to **0.3497** (resp. **0.4059**) with a cluster number of **K = 10**.

TABLE V. Evaluation of the 3 Best Clustering Techniques on M. Truncatula Salt Stress Dataset

Algorithm rank		1	2	3
		Algorithm [parameter K] score		
Internal validation	Conn. index	AGNES[K=4] 2.9290	DIANA[K=2] 5.2869	K-Means[K=2] 24.5222
	Dunn index	AGNES[K=4] 0.8296	DIANA[K=2] 0.3359	SOM[K=6] 0.0044
	Silhouette index	AGNES[K=4] 0.9587	DIANA[K=2] 0.9406	K-Means[K=10] 0.8822
Relative validation	APN measure	DIANA[K=2] 0.0001	DIANA[K=2] 0.0007	K-Means[K=2] 0.0053
	AD measure	PAM[K=10] 0.9523	PAM[K=10] 0.9945	K-Means[K=10] 1.0745
	ADM measure	DIANA[K=2] 0.0001	AGNES[K=2] 0.0288	K-Means[K=2] 0.0307
	FOM measure	K-Means[K=10] 0.3487	K-Means[K=10] 0.3497	PAM[K=10] 0.4059

According to the three internal validity measures Connectivity, Silhouette and Dunn index value, Hierarchical Clustering (AGNES) appears to be the most efficient with **K = 4** clusters for the three datasets examined (Fig. 2, 3 and 4). However, the number of plant genes categories as found by authors who have submitted these different datasets is higher than 4 categories.

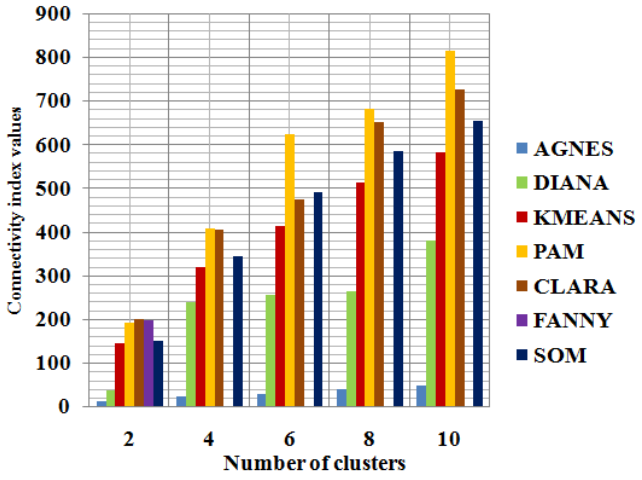

Fig. 2. Performance of Connectivity index using A. thaliana salt stress dataset.

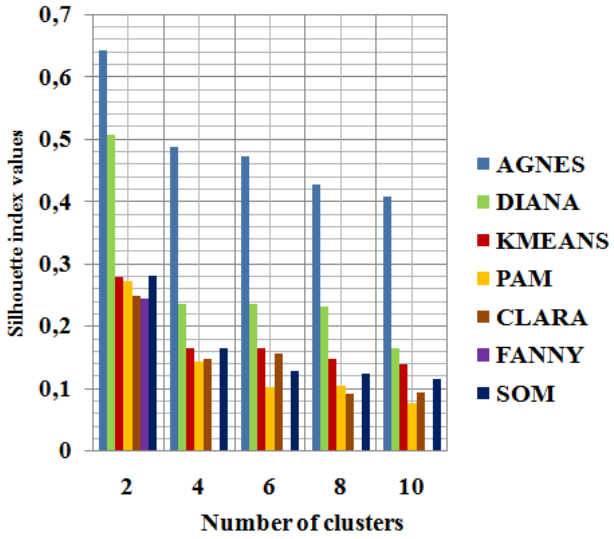

Fig. 3. Performance of Silhouette index using A. thaliana salt stress dataset.

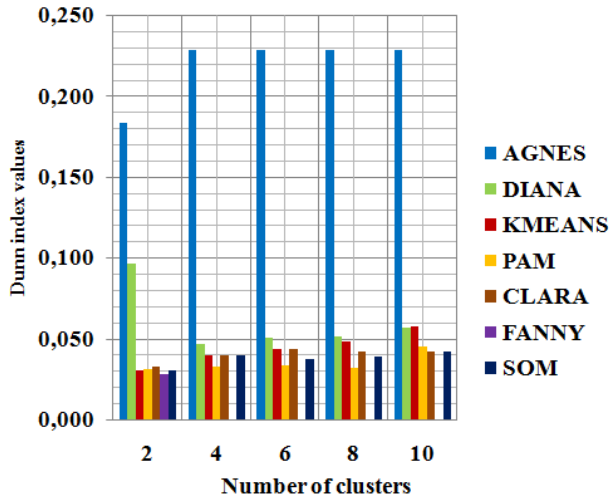

Fig. 4. Performance of Dunn index using A. thaliana salt stress dataset.

On the other side, the relative validity measures report that SOM algorithm performs well for dataset 1, with AD measure value equal to **3.6581** and FOM measure value equal to **0.6988** (Fig. 5 and 6). PAM and K-Mean provide good results for dataset 2 and 3 with the same number of clusters equal to **10** (Fig. 5 and 6). Also, this number of clusters would correspond biologically to the number of gene families found.

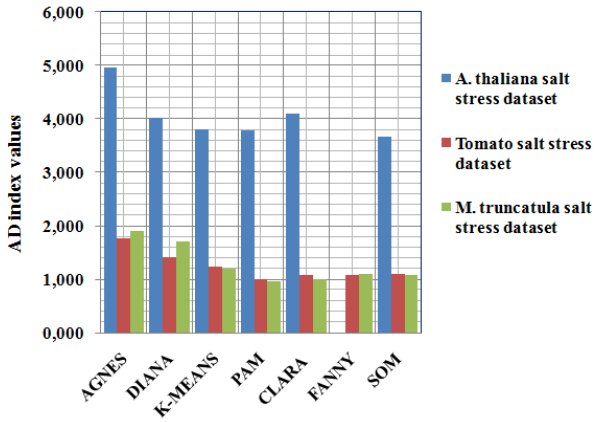

Fig 5. Performance of AD index with K=10.

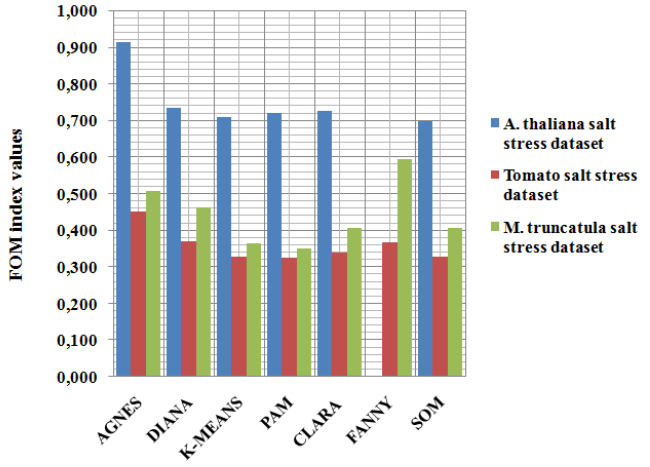

Fig 6. Performance of FOM index with K=10.

Finally, this study demonstrates that, according to the values of the validity indices (internal and relative) and the number of optimal clusters:

- For dataset 1, the SOM algorithm is the most efficient with the relative validation indices (ADM and FOM) for an optimal cluster number of 10. This cluster number is compatible with the biological reality of different gene families obtained by the submitters of this dataset. This algorithm has confirmed its performance in other datasets of complex organisms such as: Human Fibroblasts Serum and Rat CNS datasets [19].
- For datasets 2 and 3, PAM and K-Means algorithms are also distinguished by their performance for the same relative validation indices (ADM and FOM) and for the same cluster number equal to 10 compatible with the biological reality of different gene families obtained by the experimenters. These algorithms revealed interesting results in different kind of datasets: cancerous tissue types [52] and on the plant functions [53].

When we consider only the values of validity indices (internal and relative) without taking into count the cluster number expected by biologist:

- For all datasets, the AGNES algorithm presents the best internal indices values (Connectivity, Dunn and Silhouette) with an optimal number of clusters K=4. We also note that according to the relative validity indices (APN, ADM and FOM), the K-Means algorithm seems to be suitable for the three datasets with an APN index value of **0.0236** for dataset 1, which decreases to **0.0116** for dataset 2 and 0.0053 for dataset 3 when the number of clusters is set to 2. The ADM index also shows a decreasing trend with values of **0.1045** for dataset 1, and **0.0804** (respectively **0.0307**) for datasets 2 and 3 respectively with the same number of clusters K=2. For the last index, which is FOM, it is equal to **0.7083** for dataset 1 and decreases to **0.3280** for dataset 2 and **0.3497** for dataset 3 for K=10. And here the number of clusters is in adequacy with the biological reality of the families of genes.

However, firstly it should be kept in mind that these datasets are not reference datasets and therefore they are not necessarily "potentially groupable" which may explain the mismatch between the number of optimum clusters obtained and the number of expected clusters by biologists. Secondly, we have had to retain only a number of 1000 genes for each datasets. This reduced size of the gene sample is due to the number of genes annotated on tomato which is less than that of the other two plants. This fact may have contributed to this mismatch or competed to make the datasets less groupable.

V. Conclusion

In this paper, seven clustering algorithms were compared and evaluated on three sets of gene expression data from plants subjected to salt stress. The purpose was to determine the best performing algorithm that produces the optimal number of clusters reflecting the biological reality.

The results showed that the SOM algorithm allows a good distribution of genes for dataset 1. The partitioning algorithms PAM and K-Means for datasets 2 and 3 lead to the same results but with slightly lower validity index values. When we take into account only the internal validity indices, we see that the AGNES algorithm presents for the three data sets, the best values (Connectivity, Dunn and Silhouette) with a number of clusters equal to 4. In this case, we also note that the values of the relative validity indices allow the emergence of a trend indicating an acceptable performance of K-Means for the three sets of data.

This work has certain limitations: (i) The number of genes studied: Only 1000 genes are selected. (ii) Noise and outliers are inherent in the expression data. Clustering methods can be affected by this phenomenon. But, although K-Means is generally deemed as a sensitive method to outliers, it appears in this study that it is not the case. Because we obtained for this later a result with acceptable indices values and with an optimal cluster number identical to the one expected by biologists.

These results provide guidance for future work. The use of AGNES and K-Means clustering methods may be recommended for the analysis of this type of datasets. The additional orientation would be to associate the expression profiles (numerical aspect) with the corresponding annotations described by the ontologies (semantic aspect) in order to provide enrichment in the gene clustering.

Acknowledgment

The authors would like to thank the Directorate General of Science Research and Technological Development (DGRSDT), Ministry of Higher Education and Scientific Research of Algeria for their support in this work.

References

[1] Sharma. (2016). "Computational gene expression profiling under salt stress reveals patterns of co-expression", Genomics data, Vol. 7, pp. 214-221. DOI: https: //doi.org/10.1016/j.gdata.2016.01.009.

[2] F. M. Afendi, N. Ono, Y. Nakamura, K. Nakamura, L. K. Darusman, N. Kibinge, A. H. Morita, K. Tanaka, H. Horai, and M. Altaf-Ul-Amin. (2013), "Data mining methods for omics and knowledge of crude medicinal plants toward big data biology", Computational and Structural Biotechnology Journal, Vol. 4, No. 5, pp. e201301010. DOI: https: //dx.doi.org/10.5936/csbj.201301010.

[3] L. M. O. Mesa, L. F. N. Vasquez, and L. Lopez-Kleine. (2012). "Identification and analysis of gene clusters in biological data". In 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops, pp. 551-557.

[4] N. Pasquier, C. Pasquier, L. Brisson, and M. Collard. (2008). "Mining gene expression data using domain knowledge", International Journal of Software and Informatics (IJSI), Vol. 2, No. 2, pp. 215-231.

[5] K. Raza. (2012), "Application of data mining in bioinformatics", arXiv preprint arXiv: 1205.1125.

[6] R. Govindarajan, J. Duraiyan, K. Kaliyappan, and M. Palanisamy. (2012). "Microarray and its applications", Journal of pharmacy & bioallied sciences, Vol. 4, No. (Supp2), pp.S310. DOI: https: //dx.doi.org /10.4103/0975-7406.100283.

[7] W. Shannon, R. Culverhouse, and J. Duncan. (2003). "Analyzing microarray data using cluster analysis". Pharmacogenomics, Vol. 4, No. 1, pp. 41-52. DOI: https: //doi.org/10.1517/phgs.4.1.41.22581.

[8] W. A. Rensink and C. R. Buell. (2005). "Microarray expression profiling resources for plant genomics", Trends in plant science, Vol. 10, No. 12, pp. 603-609. DOI: https: //dx.doi.org/10.1016/j.tplants.2005.10.003.

[9] S. Y. Rhee and M. Mutwil. (2014). "Towards revealing the functions of all genes in plants". Trends in plant science, Vol. 19, No. 4, pp. 212-221. DOI: https: //dx.doi.org/10.1016/j.tplants.2013.10.006.

[10] K. Byron and J. T. Wang. (2018). "A comparative review of recent bioinformatics tools for inferring gene regulatory networks using time-series expression data". International journal of data mining and bioinformatics, Vol. 20, No. 4, pp. 320-340. DOI: https: //doi.org/10.1504/IJDMB.2018.094889.

[11] Y. Loewenstein, E. Portugaly, M. Fromer, and M. Linial. (2008), "Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space", Bioinformatics, Vol. 24, No. 13, pp. i41-i49. DOI: https: //doi.org/10.1093/bioinformatics/btn174.

[12] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. (1998). "Cluster analysis and display of genome-wide expression patterns" Proceedings of the National Academy of Sciences, Vol. 95, No. 25, pp. 14863-14868.

[13] A.Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, and X. Yu. (2000). "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling", Nature, No. 6769, Vol. 403, pp. 503. DOI: https: //doi.org/10.1038/35000501.

[14] J. Bajsa, Z. Pan, and S. O. Duke. (2011). "Transcriptional responses to cantharidin, a protein phosphatase inhibitor, in Arabidopsis thaliana reveal the involvement of multiple signal transduction pathways" Physiologia plantarum, Vol. 143, No. 2, pp. 188-205. DOI: https: //doi.org/10.1111/j.1399-3054.2011.01494.x.

[15] A. Hossen, H. A. Siraj-Ud-Doulah, and A. Hoque. (2015). "Methods for evaluating agglomerative hierarchical clustering for gene expression data: a comparative study", Computational Biology and Bioinformatics, Vol. 3, No. 6, pp. 88-94. DOI: https: //doi.org/10.11648/j.cbb.20150306.12.

[16] F. Takahashi, J. Tilbrook, C. Trittermann, B. Berger, S. J. Roy, M. Seki, K. Shinozaki, and M. Tester. (2015). "Comparison of leaf sheath transcriptome profiles with physiological traits of bread wheat cultivars under salinity stress", PLoS One, Vol. 10, No. 8, pp. e0133322. DOI: https: //doi.org/10.1371/journal.pone.0133322.

[17] P. Gasch and M. B. Eisen. (2002). "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering", Genome biology, Vol. 3, No. 11, pp. research0059. 1, 2002. DOI: https: //doi.org/10.1186/gb-2002-3-11-research0059.

[18] Y. Ge, Y. Li, Y.-M. Zhu, X. Bai, D.-K. Lv, D. Guo, W. Ji, and H. Cai. (2010). "Global transcriptome profiling of wild soybean (Glycine soja) roots under NaHCO 3 treatment", BMC plant biology, Vol. 10, No. 1, pp. 153. DOI: https: //doi.org/10.1186/1471-2229-10-153.

[19] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. (1999), "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation", Proceedings of the National Academy of Sciences, Vol. 96, No. 6, pp. 2907-2912, 1999. DOI: https: //doi.org/10.1073/pnas.96.6.2907.

[20] S. Babichev, V. Lytvynenko, M. A. Taif, and A. Sharko. (2016). "Hybrid model of inductive clustering system of high-dimensional data based on the sota algorithm". No. 843, pp. 173-179.

[21] T. Deepika, and R. Porkodi. (2015). "A survey on microarray gene expression data sets in clustering and visualization plots". Int J Emerg Res Manag Technol, Vol. 4, No. 3, pp. 56-66.

[22] M. S. Hasan, and Z. H. Duan. "Hierarchical k-Means: A Hybrid Clustering Algorithm and Its Application to Study Gene Expression in Lung Adenocarcinoma". In Emerging Trends in Computer Science and Applied Computing, chap 4. Quoc Nam Tran and H. Arabnia, Eds. Boston: Morgan Kaufmann, 2015, pp. 51-67. DOI: https: //doi.org/10.1016/B978-0-12-802508-6.00004-1.

[23] C. Muruganantha, and D. Ramyachitra. (December 2014). "An Empirical Analysis of Flame and Fuzzy C-Means Clustering for Protein Sequences". International Journal of Computational Intelligence and Informatics Vol. 4, No. 3, pp. 214-220.

[24] J. Oyelade, I. Isewon, F. Oladipupo, O. Aromolaran, E. Uwoghiren, F. Ameh, M. Achas, and E. Adebiyi. (2016). "Clustering algorithms: Their application to gene expression data". Bioinformatics and Biology insights, Vol. 10, pp. BBI. S38316. DOI: https: //doi.org/10.4137/BBI.S38316.

[25] A. Sharma. (2016). "Computational gene expression profiling under salt stress reveals patterns of co-expression", Genomics data, Vol. 7, pp. 214-221. DOI: https: //doi.org/10.1016/j.gdata.2016.01.009.

[26] L. López-Kleine, J. Romeo, and F. Torres-Avilés. (2013). "Gene functional prediction using clustering methods for the analysis of tomato microarray data". In 7th International Conference on Practical Applications of Computational Biology & Bioinformatics. pp. 1-6. Springer, Heidelberg. DOI: https: //doi.org/10.1007/978-3-319-00578-2_1.

[27] N. Belacel, Q. Wang, and M. Cuperlovic-Culf. (2006), "Clustering methods for microarray gene expression data", Omics: a journal of integrative biology, Vol. 10, No. 4, pp. 507-531. DOI: https: //doi.org/10.1089/omi.2006.10.507.

[28] A. Bihari, S. Tripathi, and A. Deepak. (2019). "Gene Expression Analysis Using Clustering Techniques and Evaluation Indices". Available at SSRN 3350332.

[29] A. A. Singh, A. E. Fernando, and E. J. Leavline. (2016). "Performance Analysis on Clustering Approaches for Gene Expression Data". International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, No. 2, pp. 196-200. DOI: https: //doi.org/10.17148/IJARCCE.2016.5242.

[30] D. Luo, Y. Wu, J. Liu, Q. Zhou, W. Liu, Y. Wang ... & Z. Liu. (2019). "Comparative transcriptomic and physiological analyses of Medicago sativa l. indicates that multiple regulatory networks are activated during continuous ABA treatment". International journal of molecular sciences, Vol. 20, No. 1, pp.47. DOI: https: //doi.org/10.3390/ijms20010047.

[31] Rousseeuw, P. J., & Kaufman, L. (1990). Finding groups in Hoboken: Wiley Online Library.

[32] Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. Paper presented at the ACM Sigmod Record.

[33] Guha, S., Rastogi, R., & Shim, K. (1998). CURE: an efficient clustering algorithm for large databases. Paper presented at the ACM Sigmod Record.

[34] Guha, S., Rastogi, R., & Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. Information systems, Vol. 25, No. 5, pp. 345-366. DOI: https://doi.org/10.1016/S0306-4379(00)00022-3.

[35] Karypis, G., Han, E.-H. S., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. Computer, Vol. 8, pp. 68-75. DOI: https: //doi.ieeecomputersociety.org/10.1109/MC.2005.258

[36] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Paper presented at the Proceedings of the fifth Berkeley symposium on mathematical statistics and probability.

[37] Kaufman, L., Rousseeuw, P., & Dodge, Y. (1987). Clustering by Means of Medoids in Statistical Data Analysis Based on the: L1 Norm,~ orth-Holland, Amsterdam.

[38] Deepa, M. S., & Sujatha, N. (2014). Comparative Studies of Various Clustering Techniques and Its Characteristics. International Journal of Advanced Networking and Applications, Vol. 5, No.6, pp. 2104.

[39] Ng, R. T., & Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. IEEE Transactions on Knowledge & Data Engineering, Vol. 5, pp. 1003-1016. DOI:https://doi.ieeecomputersociety.org/10.1109/TKDE.2002.1033770.

[40] Kohonen, T. (2013). Essentials of the self-organizing map. Neural networks, Vol. 37, pp. 52-65. DOI: https://doi.org/10.1016/j.neunet.2012.09.018

[41] Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. DOI: https:/doi: 10.1080/01969727308546046

[42] Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (2005). Automatic subspace clustering of high dimensional data. Data Mining and Knowledge Discovery, Vol. 11, No. 1, pp. 5-33. DOI: https://10.1007/s10618-005-1396-1

[43] Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Paper presented at the Kdd.

[44] Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Lin, C.-T. (2017). A review of clustering techniques and developments. Neurocomputing, Vol. 267, pp. 664-681. DOI: https://doi.org/10.1016/j.neucom.2017.06.053.

[45] Z. Chan, R. Grumet, and W. Loescher. (2011). "Global gene expression analysis of transgenic, mannitol-producing, and salt-tolerant Arabidopsis thaliana indicates widespread changes in abiotic and biotic stress-related

genes", Journal of Experimental Botany, Vol. 62, No. 14, pp. 4787-4803. DOI: https://doi.org//10.1093/jxb/err130.

[46] W. Sun, X. Xu, H. Zhu, A. Liu, L. Liu, J. Li, and X. Hua. (2010). "Comparative transcriptomic profiling of a salt-tolerant wild tomato species and a salt-sensitive tomato cultivar", Plant and Cell Physiology, Vol. 51, No. 6, pp. 997-1006, 2010. DOI: https://doi.org/10.1093/pcp/pcq056.

[47] D. Li, Y. Zhang, X. Hu, X. Shen, L. Ma, Z. Su, T. Wang, and J. Dong. (2011). "Transcriptional profiling of Medicago truncatula under salt stress identified a novel CBF transcription factor MtCBF4 that plays an important role in abiotic stress responses". BMC plant biology, Vol. 11, No. 1, pp. 109. DOI: https://doi.org/10.1186/1471-2229-11-109.

[48] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, and M. Holko. (2012). "NCBI GEO: archive for functional genomics data sets—update". Nucleic acids research, Vol. 41, No. D1, pp. D991-D995. DOI: https://doi.org/10.1093/nar/gks1193.

[49] A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G. C. Tseng. (2006). "Evaluation and comparison of gene clustering methods in microarray analysis", Bioinformatics, Vol. 22, No. 19, pp. 2405-2412. DOI: https://doi.org/10.1093/bioinformatics/btl406

[50] G. Brock, V. Pihur, S. Datta, and S. Datta. (2011). "clValid, an R package for cluster validation", Journal of Statistical Software (Brock et al., March 2008).

[51] Punitha, K. (2019). Extraction of Co-Expressed Degs From Parkinson Disease Microarray Dataset Using Partition Based Clustering Techniques. Paper presented at the 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT). DOI: https://ieeexplore.ieee.org/document/8869140

[52] M. C. de Souto, I. G. Costa, D. S. de Araujo, T. B. Ludermir, and A. Schliep. (2008). "Clustering cancer gene expression data: a comparative study", BMC bioinformatics, Vol. 9, No. 1, pp. 497. DOI: https://doi.org/10.1186/1471-2105-9-497.

[53] X. Yu, G. Yu, and J. Wang. (2017). "Clustering cancer gene expression data by projective clustering ensemble", PLoS One, Vol. 12, No. 2, pp. e0171429. DOI: https://doi.org/10.1371/journal.pone.0171429.

Houda Fyad

Houda Fyad is an Assistant Professor in computer science at University of Oran 2, Algeria. She received her engineering degree in computer science department (2006) from University of Oran1. She also received her master degree (2011) with specialization Informatique & Automatique from the same university. Currently she is preparing her PhD thesis within the Computer Science Department in the University of Oran1 with specialization Diagnostic and Decision-Making Assistance and Human Interaction Machine. In research field, she works on Machine Learning, Data mining, Bioinformatics and Ontologies.

Fatiha Barigou

Fatiha Barigou is a university lecturer at Computer Science Department at Université Oran 1. She is a research member of the AIR team in the LIO laboratory. She does research in Text Data Mining, Big data and Artificial Intelligence. Her current projects are Sentiment Analysis, AI, Fog and Cloud Computing in healthcare.

Karim Bouamrane

Karim Bouamrane received the PhD Degree in computer science from the Oran University in 2006. He is Professor of computer Science at the same university. He is the head of computer science laboratory (LIO) and Decision and piloting system team. His current research interests deal with decision support system in maritime transportation, urban transportation system, production system, and application of bio-inspired based optimization metaheuristic. He participates in several scientific committees' international/national conferences in Algeria and others countries in the same domain and collaborates in Algerian-French scientific projects. He is co-author of more than 40 scientific publications.