Analysis of low-correlated spatial gene expression patterns: a clustering approach in the mouse brain data hosted in the Allen Brain Atlas

P. Rosati[1], *C.A. Lupaşcu*[2] *and D. Tegolo*[3]*

[1] *Innogea srl, Via Principe di Belmonte 102, Palermo, Italy*
[2] *Institute of Biophysics, National Research Council, Via Ugo La Malfa 153, Palermo, Italy*
[3] *Department of Mathematics and Computer Science, University of Palermo, Via Archirafi 34, Palermo, Italy*
* *E-mail: domenico.tegolo@unipa.it*

Abstract: The Allen Brain Atlas (ABA) provides a similar gene expression dataset by genome-scale mapping of the C57BL/6J mouse brain. In this paper, we describe a method to extract the spatial information of gene expression patterns across a set of 1047 genes. The genes were chosen from among the 4104 genes having the lowest Pearson correlation coefficient used to compare the expression patterns across voxels in a single hemisphere for available coronal and sagittal volumes. The set of genes analysed in this paper is the one discarded in the article by Bohland et al., which was considered to be of a lower consistency, not a reliable dataset. Following a normalisation task with a global and local approach, voxels were clustered using hierarchical and partitioning clustering techniques. Cluster analysis and a validation method based on entropy and purity were performed. We analyse the resulting clusters of the mouse brain for different numbers of groups and compare them with a classically-defined anatomical reference atlas. The high degree of correspondence between clusters and anatomical regions highlight how gene expression patterns with a low Pearson correlation coefficient between sagittal and coronal sections can accurately identify different neuroanatomical regions.

1 Introduction

In the last two decades, the evolution of technology has allowed answers to be given to some queries arising from biological studies. Some of them have been restricted to the own field of science involving biological researchers, some others have been supported by a multidisciplinary approach in which data analysis plays a formal and substantial role [1–3]. Without prejudice to the generality, only a few principle technologies need to be cited so as to understand the quality of the evolution in gene expression. Microarrays with its two-colour fluorescence hybridisation allow the monitoring of the expression of many genes [4]. Starting from the methodologies proposed by F. Sanger et all in [5] a new approach named *Next Generation Sequencing* was developed during the last decade. Shendure and Ji in [6] dissert on this methodology, highlighting the positive performances concerning the Sanger model, and during the year 2010, McKenna et al. in [7] discuss their GATK framework in which it is easy to develop efficient and robust tools to analyse next-generation DNA sequencers.

New research topics in science and technology aimed at improving the ability to store large amounts of data. Their analysis, interpretation and their integration have characterised the recent challenges of to researchers.

Over the last decades, methods and techniques, acquired from scientific disciplines related to biology, reformulated adequately to extracting correlations between heterogeneous data, have contributed to the interpretation of complex data allowing the identification of innovative solutions otherwise unattainable with classical methods. Therefore, to investigate data with information poverty and apparently without significant correlations, methodologies orientated to the analysis of data in their entirety will be taken into account by extracting a substantial amount of information. Notably, it was decided to investigate and describe new empirical approaches adaptable to the multidimensional biomedical data currently present on the Allen Brain Atlas (ABA) [8].

Bohland et al. in [9] introduce a study to identify distinct neuroanatomical areas based on gene expression patterns. Spatial correlation has detected distributions of genes or patterns among gene expression profiles in the brain. They observed patterns in a multivariate exploration technique and data analysis was adopted on a broad set of spatially co-registered gene expression volumes derived from the Allen Brain Atlas (ABA). Coronal and sagittal volumes were registered, and the maximum correlation coefficients were evaluated. The analysis was focused on 75% of best correlation, and no analysis was performed on the lowest correlation (25% of cases); thus they present their result on 75% of higher correlation values. Following the Bohland approach, our proposed method intends to investigate 25% of lower cases. Thus a reformulation of the method was imposed to manage the lower gene correlation which can highlight distinctive neuroanatomical sites. This point of view allows work to be carried out on a small number of genes and to obtain some unique neuroanatomical sites.

The interpretation of the spatial distributions of gene expressions, present in those partitions of data that are poor in information and therefore neglected by current literary approaches, has been taken into consideration. We aim to find significant distributions when highlighting specific pathologies are hosted in the area with 25% of correlations.

The empirical methodologies are based on the identification of physiological and physical correlations of biological data into the brain regions of the adult mouse and consist of well-known methods in the area of machine learning, data sampling methods, clustering and validation methods, according to a new approach to identify new correlations in neglected areas.

The correlations among genomic distributions in the brain are of fundamental importance for the deepening of brain functions. In-depth, data analysis will be addressed on the central nervous system of the adult mouse in which it is possible to highlight thousands of genes [10]. Lein et all in [1] assert that genomic sequence information and high-throughput has allowed an expansion in the field of

global analysis. Such methodologies give a framework for analysing the relationship between brain structure and their functionality. However, in many cases, these techniques have been adopted for large brain components, thus given their significant amount of data their interpretation is not naive.

Furthermore, for more in-depth functional knowledge, it is appropriate to highlight the genomic distribution correlations in a 3D space. A set of databases are involved in literature, some of them include heterogeneous data, some other homogeneous and all of them have the aim of allowing a more robust knowledge of gene patterns and their distribution of different neuro-anatomic regions [11–18]. The Allen Brain Atlas (ABA) offers a gene expression dataset integrated by a database that allows the extraction of quantitative information and selection of genes based on spatial or anatomical localisation.

AGEA (Anatomic Gene Expression Atlas), free integrated tool in ABA discussed in [19], enables new possibilities for understanding the brain organisation in terms of genomics distributions or patterns in 2D or 3D shape, and allow the understanding of spatial correlations across expressions data for thousands of genes. In the same paper the author's argument about the tools to examine anatomical relationships: 3d dimensional visualisation, gene retrieval, and hierarchical transcription.

Fürth et all in [20] introduce an integrated framework to automatically annotate, analyse, visualise and easily share whole-brain data at cellular resolution, based on a scale-invariant, interactive mouse brain atlas. The primary aim of this framework regards the capability to provide a brain map which integrates heterogeneous data coming from neuron identity, their connectivity and functionality. In [21], Hawrylycz et al. disserted on the gene expression, and gene co-expression relationships also demonstrate that the distributions of significant neuron cells strongly reflect the functional brain variation. Thus the anatomical division along the edges are in many cases linked with the genes interlocked with synaptic transmission. Therefore, molecular distribution and its spatial neocortex have linearly dependent topographic distribution. Recent dissemination of morphological items allows the assertion that new metrics based on the shape of patterns can give a new approach to discovering cluster association [22]. Morphometric similarity introduced by Seidlitz et al. in [23] presented a robust method to understand how human cortical networks reinforce the different individual features and not only pathologies. A set of free resources for analysis of genome sequencing are available on the shelf, but they require a lot of computational time and memory for their run. BWA, GATK, and snpEff are only some of the standard tools that can be used to align and detect variants from a single genome. However, simultaneous analysis of many genomes is significantly accelerated by the use of supercomputers which is welcomed by the geneticist's community. Many contributions for genomic analysis are orientated to the study of the entire genome and require the alignment and comparison of raw sequence data. Thus some of these cases require sequencing methods with a parallel approach for simultaneous analysis of multiple genomes. Puckelwartz et al. in [24] engaged the problem of sequencing with the adoption of a supercomputer to improve both the speed-up and results for an increased proper sequence. Our proposed method does not use a parallel or distributed system, and it needs only standard elaboration systems to find genetic distributions for evaluating the morphological shape. This is due to the use of the classified genes that are hosted in heterogeneous dataset inside the ABA. The Human Genome Project is an example of the heterogeneous dataset to identify all nucleotides in human chromosomes. These include wide-ranging data domains, accessibility through many levels of abstraction. In [25] the authors support and explore a discussion on the advantages for defining a database for the Human Genome Project designed in a heterogeneous way. Its main features are the high-level data model, the representation of notions as objects of a relational model, the metadata as an expression of a working database, the tables to populate the database; and its implementation based on a conventional relational database.

Artificial Neural Network (ANN), in the last decade, has had a significant consideration both in the different area of image analysis and computer vision and in the field of classification of high-dimensional data and feature selection extraction. Despite seeming to be a panacea for many problems present on these items, Aziz et al. in [26] study the performance of Independent Component Analysis as an alternative optimisation technique, they show experimental results in which the proposed methodologies gives more accurate classification rate for ANN classifier.

Representation, visualisation and imaging are only a few words to identify the possibility to highlight the depiction of heterogeneous and multiple information concisely. A number of exhaustive studies have been carried out, during the last few years, allowing the vices and virtues of the use of concise visualisation of a large quantity of data to be focused on. In the last decade, some members of an image processing group have orientated their attention on the bioimage informatics in which novel image processing methods twisted to data mining and visualisation approaches, and search and management data increase the focus on the application rather than the theoretical aspect. The underlying methodologies are hosted inside the classical field of image analysis and computer vision: feature extraction, segmentation, registration. Peng in [27] summarises with a brief overview of the available bioimage databases, analysis tools and other resources.

In visual analytic tools, emerging topics and new developments allow defining techniques to synthesise information and to mine features from extensive often-contradictory data. Therefore, visual representations and the technologies to understand visually complex information allow to navigate into and then they offer the best interaction user-data [28].

The Allen Brain Atlas (ABA) was designed and implemented by Allen Institute for Brain Science, and in 2006 at great demand, it was included on web distribution [29]. A set of materials can be extracted from its databases. However, the DB for human and mouse represent the most relevant heterogeneous DBs that allows integrating data coming from different sources [30]. During the last years, ABA has collected growing resources, integrating a wide range of neuroanatomical data, such as two-dimensional data structures (MRI, cytology, ...) and 3-D representations with appropriate reconstructions and gene expression data, too. Moreover, search tools and data viewers are only two, of the tools hosted in the ABA, that can be highlighted from the set of services in the open website ABA [2, 31]. The section 2 introduces the dataset and preprocessing task in which algorithms will be applied, data analysis will be argued in the section 3, section 4 reports the description of the method, and finally evaluation metrics and experiments with conclusions will be shown in 5 and 6 respectively.

2 Materials

We performed a rigorous analysis of the distribution of genome in the mouse brain. We first mapped every distribution on the different mouse areas and analysed their different variants. The method then encoded the variants into the list of elements and created feature vectors for each case and control sample.

Bohland et al. [9] have analysed the ABA data in which a high Pearson correlation rate was identified(75%) between sagittal and coronal sections. The proposed method intends to investigate the genomes in which a low Pearson correlation is present (25%), the total of 4104, genes to distinguish physiological areas of the mouse brain. The set of data was discarded in [9], thus in fig. 1 the distribution of the correlation values for coronal-sagittal sections is shown. The study focuses on the search for physiological relations between the coronal and sagittal sections of the genes and possible interactions with the brain regions.
In detail, figure 2 depicts the gene's distribution in every district with a 25% correlation between sagittal and coronal sections.

The data used in this work are gene expressions resulting from the digitisation of the result of the in situ hybridisation process (ISH) and were collected by the ABA dataset (http://mouse.brain-map.org). Given the high quality of the AGEA interface (http://mouse.brain-map.org/agea), a user navigates on a genomic map, based on the gene transcriptions of the brain and interprets these results in the

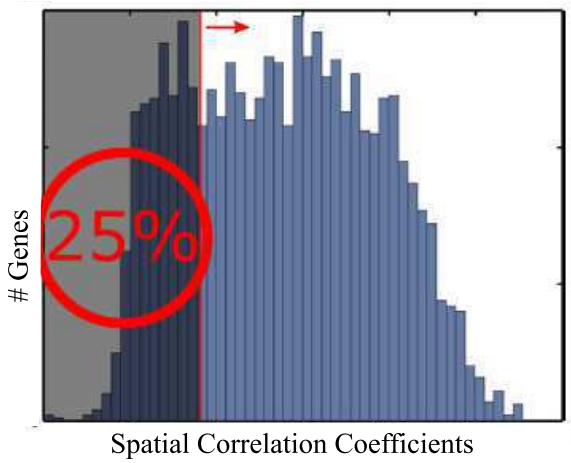

Fig. 1: 25% Histogram of the Pearson correlation coefficients, the values represent the correlation between coronal and sagittal sections of the same gene.

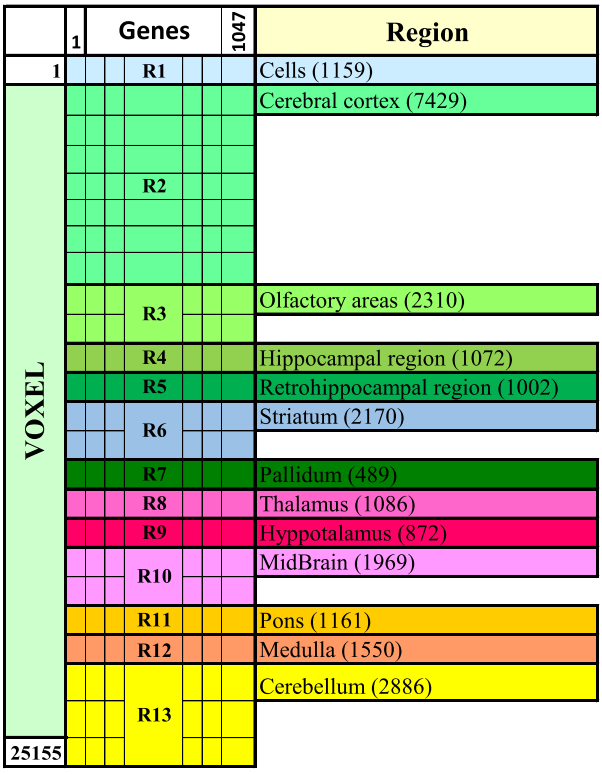

Fig. 2: Dataset at 25% correlation.

context of a reference atlas generated from histological stains of the Nissl type, called the Allen Reference Atlas (ARA).

The brain of the adult mouse has been partitioned into 49742 virtual voxels with 200 μm size. In correspondence to each voxel, the values of gene expressions have been associated for all the genes [32, 33], and the gene expressions are shown in a voxel array for genes. For each voxel v, the expression $E(v, g)$ of the gene $g \in G$ is the weighted average of the intensity I in scale grays evaluated on each p pixel present on the V voxel.

$$E(v,g) = \frac{\sum_{p \in V} M(p) I(p)}{\sum_{p \in V} 1},$$

$$\text{where } M(p) = \begin{cases} 1 & \text{if the gene } g \text{ has an expression on the voxel } p, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The matrix voxel $E(v, g)$, has size 49742×4104 and is stored in the ExpEnergy.mat data, such file has been included in the toolbox 'Brain Gene Expression Analysis to Matlab toolbox for the analysis of the brain-wide gene-expression data [34]. *The Brain Gene Expression Analysis toolbox* [29] has been adopted for the visualisation of data. It is a Matlab toolbox that includes computational techniques for quantitative analysis of gene expression data hosted in the section adult mouse brain of the Allen atlas. Therefore, to extract information from the low correlation genes and to manage multidimensional data, a Matlab library has been designed and developed for the best integration of the whole system. Such modules include data visualisation functions, with a spatial resolution of $200\mu m$. They are useful both for the comparison of gene expressions and classical neuroanatomy and for a statistical study of the co-expression networks of gene groups and grouping of voxel genes.

Preprocessing

In many cases, the pre-processing phase is fundamental for a correct analysis of data; therefore it is suitable to identify the most appropriate methods to improve data that will be processed in the following phases. To this end, some methodologies were evaluated and, after careful analysis, two of them gave the best results. Hence they are adopted (Z-score normalisation, Principal Component Analysis).

2.0.1 Normalization: Mean, and standard deviation are two useful parameters in different areas of data analysis. Due to the dimensionality or resolution of data, some cases highlight false positive results. Therefore, a normalisation phase, as a preprocessing activity, is undoubtedly necessary to produce homogeneous data. The normalisation process was performed with the Z-score method. Given a dataset M and for each $g_m \in M$ its mean μ and standard deviation σ, the set of values contained in M may be normalised with the following normalisation ratio Z_m:

$$Z_m = \frac{g_m - \mu}{\sigma}$$

It may be shown that a series of values normalised with the Z-score has an average equal to zero and a standard deviation similar to one. Therefore, such normalised data represent an independent unit of measurement that can be used to compare values with different units of measurement.

2.0.2 Principal Component Analysis: The Principal Component Analysis (PCA) is a useful methodology to identify representative models in the data and express the data in such a way as to highlight the similarities and the differences between the elements [35, 36]. Given the dataset D with size $n \times m$, in which n measurements have $m-$dimension, the method may be formalized:

1. calculate the average of each $m-$dimensional vector;
2. evaluate the covariance matrix of the whole data series;
3. find the corresponding eigenvectors and eigenvalues;
4. sort the eigenvectors in descending order with respect to the corresponding eigenvalues and choose the first k eigenvectors related to the data of maximum variance. This is done to form a new matrix W whose size is $n \times k$ composed by chosen eigenvectors and inserted in the ordered column.
5. apply the matrix W to the dataset elements, in order to transform the elements into a new $k-$dimensional subspace

$$y = W^T \times x \quad \forall\, x$$

where x is a vector $m \times 1$ representing an element, and y a vector $k \times 1$ in the new space obtained by applying the transform W on the x genes.

3 Data Analysis

Clustering techniques or supervised and non-supervised methods are commonly used as tools for information extraction, and they are used in studies on genomic expressions and microarray. The goal is to reduce data by grouping similar data into the same subset or cluster [37]. In this way, the objects belonging to the same cluster have similar characteristics [10, 35]. Clustering is a process which allows a partitioning of a set of data into a small number of subsets. Formally, given a positive value k and a dataset of n elements with $m-$dimension in which each element is named x_{ij} $i = 1, \ldots, n$ e $j = 1, \ldots, m$, clustering methods partition n elements in k cluster or subset in a supervised way or not. Clustering methods need two fundamental elements: a metric or degree of similarity, subdivided into metrics or semi-metrics, and the mode of its use on clusters [38, 39].

Hierarchical Clustering

Agglomerative and divisive approaches are introduced in the hierarchical cluster literature. The methodology proposed in this paper considers the first method in which a bottom-up analysis is envisaged. Following the standard formalisation of agglomerative clusters, the method considers as an initial step many clusters equal to the number of elements, and also a succession of iterative steps. The task of sequential steps aggregates, depending on a static criterion, the various clusters that satisfy it, so after a finite number of steps, a single cluster containing all the elements will be detected. Given a set of n vectors, they can be clustered with the support of distance or similarity matrix $n \times n$, using the following steps:

1. clusters have when initially assigning each element to one cluster;
2. calculate the coupled similarity for all clusters;
3. consider the two most similar clusters to merge them to reduce the number of clusters by one;
4. calculate the distances between the new cluster and the old clusters, updating the distance matrix;
5. repeat steps 2 and three until getting a single cluster with n vectors.

In the proposed methodology the distance adopted is the Euclidean metric among the vectors; furthermore, step 3 can be identified with three cluster strategies: *single-linkage*, *complete-linkage* and *average-linkage*.

The aggregation of two clusters with *single-linkage* approach is obtained by calculating the minimum distances between an element of a cluster and each element of all the other clusters. The *complete-linkage* method considers the maximum distance between all elements of the pair of clusters. The last approach, *average-linkage*, to aggregate clusters, consider the average distance between all the pairs of vectors of two distinct clusters.

Partitional Clustering

The method based on partitional clustering requires two constraints: to define a priori the number of clusters in which the initial set has to be subdivided and the relative distance. The method establishes the belonging of an element to a sub-set according to the distance that such an element has with the representative element of the cluster, named centroid. The K-means is a partition clustering algorithm that has the purpose of associating each element to a cluster by determining the positions of centroids $\mu_i,\ i = 1 \ldots k$ for each C_i cluster. Moreover, it minimises the distance between the element and the centroids (cluster representatives) and also maximises the distance between the clusters [40].

$$\bigcup_{i=1}^{k} C_i = X \quad C_i \cap C_j = \emptyset\ \forall i \neq j \quad \emptyset \subset C_i \subset X\ \forall i$$

The K-means, to assign an element to the cluster, solves the following expression:

$$\arg\min_{c} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i) = \arg\min_{c} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

where C_i is the set of points that belong to the cluster i, and c is the partition at the i^{th} iteraction.

Given the adaptability of the K-means method to heterogeneous datasets, there are numerous metrics proposed in the literature. Minkowski metrics, correlation coefficients and Hamming distance are just some of the possible adoptable metrics in the clustering process involved in K-means. The identification of the right metric, to obtain a convergent process at the global minimum, can be extremely burdensome and in many cases, with non-trivial distributions, the method can be classified as $NP - hard$.

4 Description of the method

Let E be the matrix with dimension 49742×4104 that contains the elements of the dataset. The 2-dimensional matrix has rows to identify a voxel and columns for a gene. Thus the intersection between a voxel and a gene represents the gene expression hosted in each voxel of the mouse brain.

The goal is to find similarities between physiological and physical links of the $E_{75} \subset E$ dataset of (the most correlated genes between sagittal and coronal sections) and the complementary counterpart $E_{25} \subset E$ which has less-correlated gene expressions.

The gene expression data-set has two representations. The first is represented by the matrix E, it is named 'fine' and includes all gene expressions. The second one is more synthetic, it is called 'big12' and is referred to a matrix F with a size of 25155×4104, it includes voxels and an equal number of genes. The subset taken in consideration in this article will be 'big12', with the $F_{75} \subset F$ dataset which include more gene expression correlation than the $F_{25} \subset F$.

All the voxels of the 'big12' dataset belong, neuroanatomically and disjointly, to the 13 regions of the brain: Cells, Cerebral Cortex, Olfactory areas, Hippocampus region, Retrohippocampal region, Striatum, Pallidum, Thalamus, Hypothalamus, Midbrain, Pons, Medulla, Cerebellum (see Figure 3).

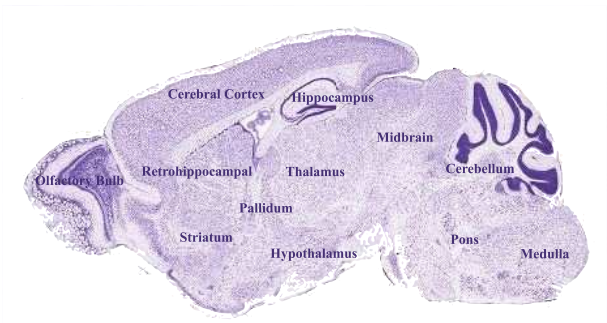

Fig. 3: Regions of the brain of an adult mouse ABA atlas (Kcnab3-RP_050927_01_A02-sagittal: *gene=Kcnab3, ProbeType=RNA, ProbeOrient.=Antisense, Plane=sagittal, Treatments=ISH*).

After a meticulous study of the dataset, a preprocessing phase on the dataset F has been necessary. Thus a normalisation task and a

5×5 Gaussian filter with a global and local approach were included in the proposed methodology.

The local approach performs the normalisation process for each gene expression (column in the E matrix), and for each subset of the 13 regions (see figure 4). The Z-score normalisation was applied, and then a task eliminated all the data residing on the Gaussian distribution queues $[\mu - 3\sigma, \mu 3\sigma]$. The other approach applies the z-score normalisation and the Gaussian filter on a whole column (see figure 5).

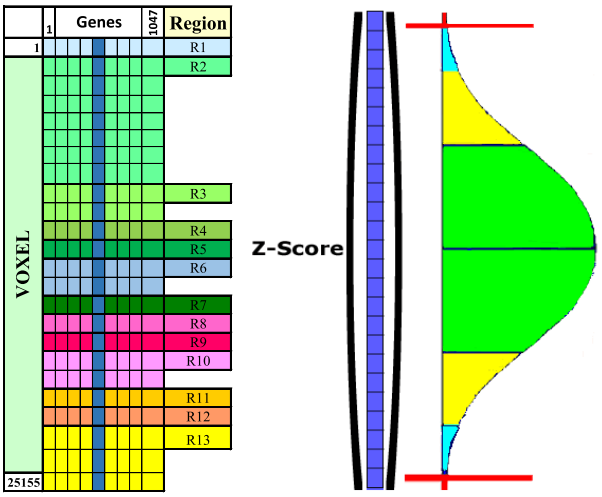

Fig. 4: Local Zscore.

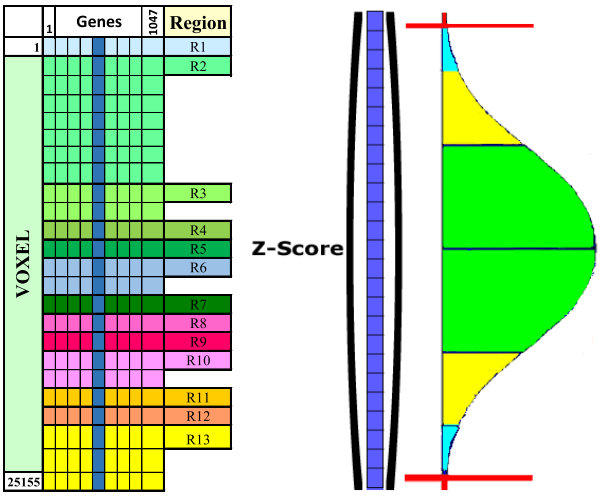

Fig. 5: Global Zscore.

Following the preprocessing phase, two matrix were extracted that represent both the region and the genes expression. Let G be the array of characteristics of size 1047×26 defined as:

$$G = [\overline{G}_{max}, \overline{G}_{dens}] \in \mathbb{M}(1047, 13, 2)$$

$$\text{with } \eta_i^j \in \overline{G}_{max}, \delta_i^j \in \overline{G}_{dens},$$

where $i = 1, \ldots, 1047$, $j = 1, \ldots, 13$, η_i^j is maximum normalized value of the gene i-th in the j-th region and δ_i^j is the density, given by the ratio between the not-null values of the voxel number of the i-th gene and the voxel number of the j-th region.

$$\eta_i^j = \max_{k=i_j}^{i_j|R_i|} g_i(k) \qquad \delta_i^j = \frac{|g_i(k) \neq 0|}{|R_j|}$$

where $k = i_j, \ldots, i_j \, |R_j|$

The creation of this matrix binds the genes and the regions according to the maximum gene expressions and their densities. Figure 6 shows the G characteristics matrix in which PCA will extract the right selection of features. The next phase should consist of hierarchical and partitioning clustering techniques, but a resizing action was assumed to select the minimum number of information needed for the cluster analysis. PCA method was applied to the matrix G which underwent a re-sized PR_2 matrix (see figure 7) before proceeding with cluster analysis.

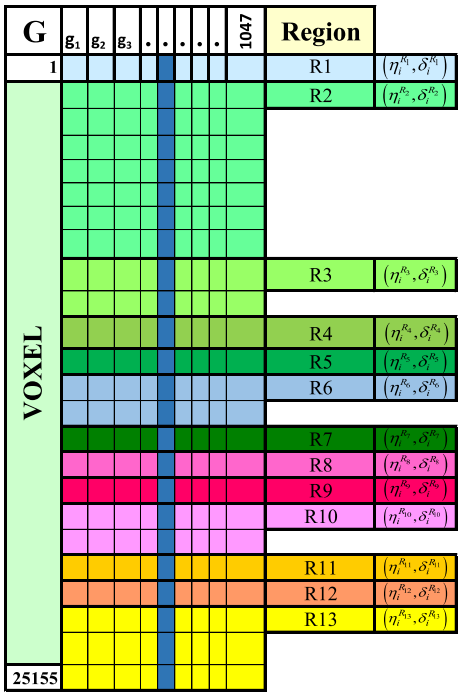

Fig. 6: The G characteristics matrix.

This paper introduces a hierarchical clustering technique with Euclidean metric and complete-linkage as aggregation criterion. Dendrogram shape can represent the results. Whereas, the partition method was executed both with the K-means algorithm and with the c-mean fuzzy cluster.

In all of the different modalities, establishing the number of clusters K a priori is essential. In our case, $K \in [13, 60]$ and the regions of the brain are $R = 13$. In all two cases (hierarchical, K-means) a set of 1047 genes was partitioned in K clusters. Therefore, the genes contained in each cluster can be considered as generators of the respective regions.

Among the different analyses implemented to evaluate data, only the two methods that best highlighted the results obtained would be described.

MAX2 Method

Let $PR_2 \in \mathbb{M}(1047, 2)$ be the projection matrix after the PCA transforms the matrix G with only two selected columns (the first two eigenvectors of the new space). The i-th row vector of PR_2 is defined as: (x_{Gi}, y_{Gi}), where i indexes the i-th gene expression.

Therefore, by fixing the number of clusters K and the clustering algorithm, the matrix PR_2 will have the following partition:

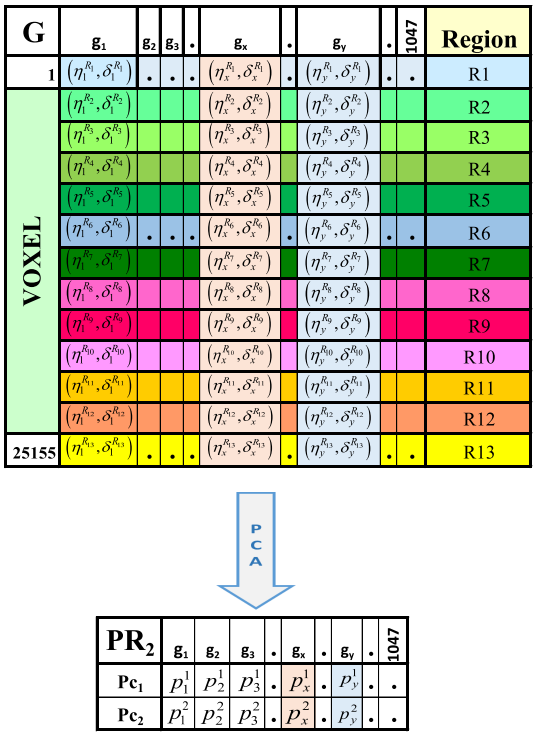

Fig. 7: Application of PCA at the G matrix.

$$P_{PR_2} = \{C_1, C_2, \ldots, C_K\}$$

C_i includes all of the row-vectors of PR_2 related to the i-th cluster. In the next step, we intend to compare the obtained clusters with the neuroanatomic regions of mouse brain. Let $G_{max} \in \mathbb{M}(1047, 13)$ be the matrix which contains only the maximum values (η_i^j).

The method focuses on the rows of the matrix G_{max} corresponding to the genes included in the C_i cluster. LetÃćÂĂÃŹs consider the i-th row in G_{max} for each gene extracted from C_i, and store the position of the maximum value in the range $[1, 13]$. Such position indexes the region where the gene has the maximum gene expression, thus a region will be associated with each gene.

However, we can associate to each cluster the region with the highest occurrence among all the positions associated with the elements of the cluster itself. To summarise, starting from the partitioning of the genes in disjointed clusters, we can say that it will be possible to associate a more common region to each cluster.

NearGene Method

The main difference with the MAX2 Method is the rule of comparison between clusters found and neuroanatomical regions. To have a correct comparison the NearGene method and Max2 we are forced to borrow the same matrix adopted by MAX2.
Let $PR_2 \in \mathbb{M}(1047, 2)$ be the projection matrix, $P_{PR_2} = \{C_1, C_2, \ldots, C_K\}$ the partial cluster of the PR_2 matrix and $G_{max} \in \mathbb{M}(1047, 13)$ the matrix in which maximum values are stored.

For each row of the matrix G_{max}, a $[0, 1]$ normalization phase was applied and then their percentage was calculated. Moreover, for each element $(x_{g_h}^i, y_{g_h}^i)$ of the cluster C_i with $h \in \{c_1, c_2, \ldots, |C_i|\}$, the centroid $centroid(C_i) = (\bar{x}^i, \bar{y}^i)$ of the cluster will be evaluated. Finally, the row $h \in PR_2$ with the following feature:

$$\arg\min_h \|(x_{g_h}^i, y_{g_h}^i), (\bar{x}^i, \bar{y}^i)\|$$

is identified.
The index of the h-th gene allows the region of the maximum value in the h-th row of the matrix G_{max} to be associated to the cluster C_i; thus, as in the MAX2 method, to the C_i cluster will be assigned the index of the maximum region.
Starting from the previously selected elements, the following steps of the method tune its actions on the rows of the G_{max} matrix attendant to the genes included in the C_i cluster. Therefore, for each gene found in C_i, the method looks for the i-th row in G_{max}, and then the index (position) of the maximum value will be taken into account because it indicates the region in which the gene assumes its maximum expression. It is clear that each gene of every cluster will be assigned a neuroanatomical region, and in the same way, it is possible to assign to each cluster C_i the region with the maximum occurrence chosen from among all components (genes) of the cluster.
Therefore, it is simple to affirm that starting from a genes partition into disjointed clusters; a region with high occurrence value will be assigned to every cluster.

5 Metrics for data evaluation

To assess the correct assignment of gene expressions to clusters with the use of the less correlated data set F_{25} two indicators are described which globally represent the degree of information and their goodness: Entropy and purity.

Therefore, we intend to evaluate the degree of similarity [41] between gene expression present in the regions and in the clusters starting from a CM (confusion matrix) in which rows give information on the clusters, otherwise the columns the real neuroanatomic region.

Let r be the number of the regions (13), and let $CM \in \mathbb{M}(r, r)$ be the confusion matrix. The components of CM are (mc_i^j) and represent the number of elements (genes) that are assigned to the i-th cluster but they belong to the real region j with $i, j = 1, \ldots, r$.
Two metrics have been included in the methodology: **entropy** and **purity** which indicates how different regions are distributed inside the same cluster [42].

Given a particular cluster c_i of size g_i, the entropy of this cluster will be defined as follows:

$$E(c_i) = -\frac{1}{\log |R|} \sum_{h=1}^{|R|} \frac{mc_i^h}{mc_i} \log \frac{mc_i^h}{mc_i}$$

Where $|R|$ is the number of regions in the dataset, and mc_i^h is the number of elements of the h-th region that was assigned to the i-th cluster. The entropy of the solution is obtained as a sum of the individual cluster entropies in accordance to the cluster size:

$$E = \sum_{i=1}^{K} \frac{mc_i}{mc} E(c_i)$$

The values of entropy range between 0 and 1, thus the value 0 will be obtained for a perfect match, namely the elements of the class coincide with the elements of the region; while the value 1 will be representative of lousy clustering.

The other validation index has been taken into consideration: **purity** or **goodness** of the cluster C_i, and its value is evaluated by:

$$P(C_i) = \frac{1}{mc_i} \max_i mc_i^j$$

It indicates the ratio between the maximum number of elements in a cluster associated with a region and the cardinality of the cluster. Therefore, to evaluate the goodness of the whole clustering process a

weighted sum on the cardinalities of every cluster has been included in the methodology:

$$P = \sum_{i=1}^{k} \frac{mc_i}{mc} P(C_i)$$

where k is the number of clusters.
Like for the entropy, the $[0, 1]$ interval is also the range for the goodness, but for the best distribution of cluster the value of goodness needs to be near the 1 value.

6 Results and Conclusions

Concerning the definition of the methodology described in 4 and the brief description of the algorithms used for clustering in 3, an exhaustive analysis will be reported. All of the possible combinations will be tested, and a selection of plots will be reported to highlight the quality of methodology. We note that the methodology was applied in both dataset F_{25} and F_{75} and this to evaluate the results among DBs. Global or local normalisation and hierarchical or partitional clustering can be combined in four different analysis as follow:

1. global normalization and hierarchical clustering;
2. global normalization and partitional clustering;
3. local normalization and hierarchical clustering;
4. local normalization and partitional clustering.

The images in Figures 8 and 9 show the clustering results for 13 regions, in which, for the F_{25} dataset, is applied to the array $G_{dens} \in \mathbb{M}(1047, 13)$ where the components are given by the densities (δ_i^j), $i = 1, \ldots, 1047$ and $j = 1, \ldots, 13$. We recall that the density δ_i^j is the ratio between the number of voxels of non-null values of the i-th gene and the voxel number of the j-th region in the respective dataset. The circles, representing the clusters, have their centre on the cluster's centre of gravity and as a radius the maximum distance between the centre and all the other elements of the cluster.

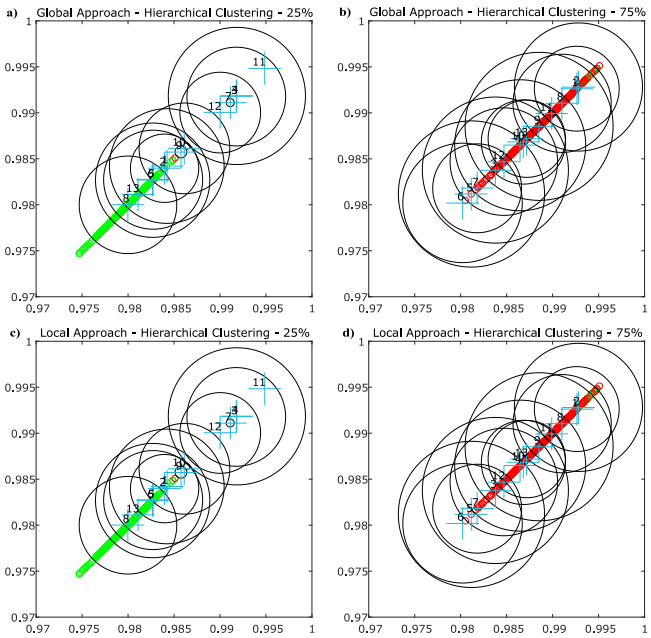

Fig. 8: The clustering results for 13 regions on the F_{25} and F_{75} dataset where a global normalization and hierarchical clustering a-b), and local normalization and hierarchical clustering c-d) approach were applied to the matrix G_{dens}. x-axis and y-axis represent the density δ_i^j of the i-th gene and of the j-th cluster respectively.

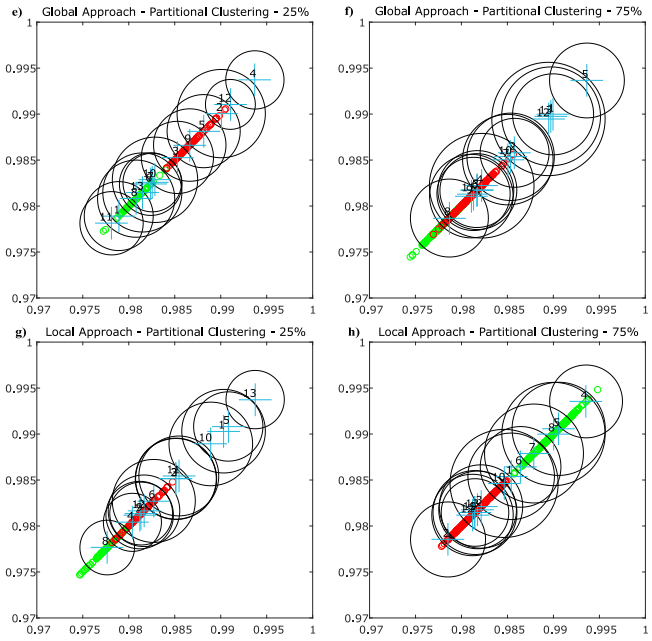

Fig. 9: The clustering results for 13 regions on the F_{25} and F_{75} dataset where a global normalization and partitional clustering e-f), and local normalization and partitional clustering g-h) approach were applied to the matrix G_{dens}. x-axis and y-axis represent the density δ_i^j of the i-th gene and of the j-th cluster respectively.

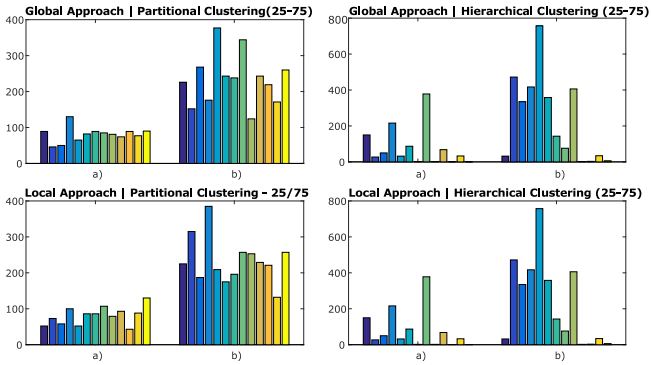

Fig. 10: Plots of gene expression distribution after the cluster analysis on the matrix G_{dens}: a) dataset 25%, b) dataset 75%.

On the other hand, Figure 10 depicts the different distributions of the genes partitioning in the 13 clusters concerning the genes of the different datasets: $F_{25}(a)$ and $F_{75}(b)$, in accordance to the four proposed approaches.

In the plots highlighted in Figures 11), 12) and 13) the same analysis, as the previous ones, is performed but applied to the characteristics matrix G_{max} given by the values of maximum gene expression η_i^j of genes compared to brain regions.

The results shown in the figures 14, 15 and 16 were obtained using the array of characteristics G which contains both the maximum values and the densities.

All the previously clustering representations highlight a substantial overlap of the data represented in the two-dimensional space. On the other hand, if before clustering steps the G matrix is transformed, namely G is projected on the two-dimensional space of maximum variance with the use of PCA. Thus, there is a slight overlap of the data as shown by the plots of figures 17, 18 and 19 that depict the effect of PCA on the datasets. Such plots put on the axis of the abscissa and of the ordinate the main components associated to the i-th gene and the j-th cluster.

All the above results dictate decisions on which path to undertake. It is easy to note that the use of PCA on the characteristic matrix G

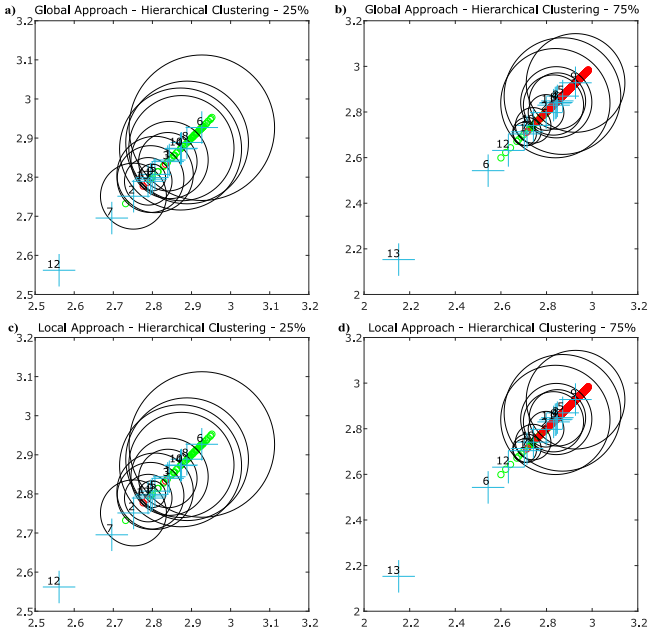

Fig. 11: The clustering results for 13 regions on the F_{25} and F_{75} dataset where a global normalization and hierarchical clustering a-b), and local normalization and hierarchical clustering c-d) approach were applied to the matrix G_{max}. x-axis and y-axis represent the density η_i^j of the i-th gene and of the j-th cluster respectively.

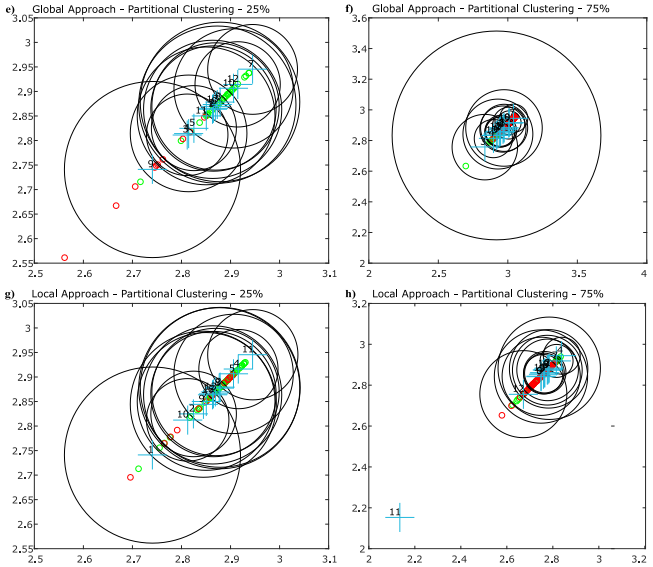

Fig. 12: The clustering results for 13 regions on the F_{25} and F_{75} dataset where a global normalization and partitional clustering e-f), and local normalization and partitional clustering g-h) approach were applied to the matrix G_{max}. x-axis and y-axis represent the density η_i^j of the i-th gene and of the j-th cluster respectively.

shows a correct classification, but a unique methodology path has to be considered to extracting the best results. Therefore, the four methods were approached with MAX2 and NearGene algorithms to assess the best results. In the following, a set of plots are reported to consider the behaviour of the different approaches to varying the number of clusters between 13 and 60. The coordinates of the points show the number of clusters used by the clustering on the abscissa, and on the ordinate the number of distinct regions identified by the clusters.

From the plots in figure 20 and 23 it is possible to highlight that the *NearGene* method is quite superior to the *MAX2* method because

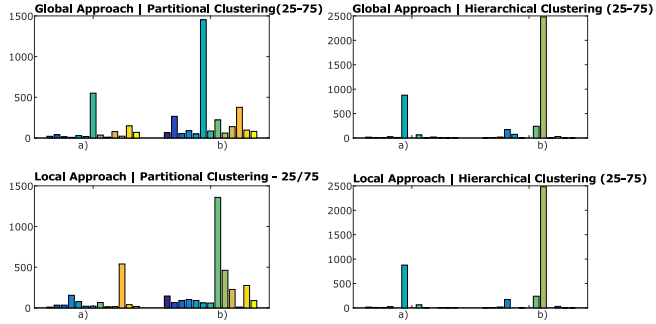

Fig. 13: Plots of gene expression distribution after the cluster analysis on the matrix G_{max}: a) dataset 25%, b) dataset 75%.

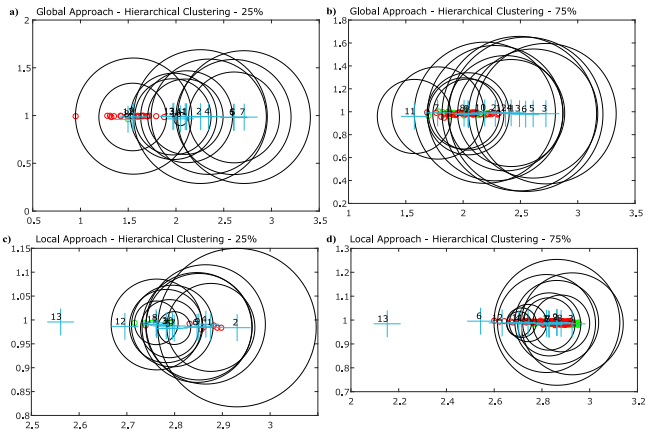

Fig. 14: The clustering results for 13 regions on the F_{25} and F_{75} dataset where a global normalization and hierarchical clustering a-b), and local normalization and hierarchical clustering c-d) approach were applied to the whole matrix G in which G_{max} and G_{dens} were considered.

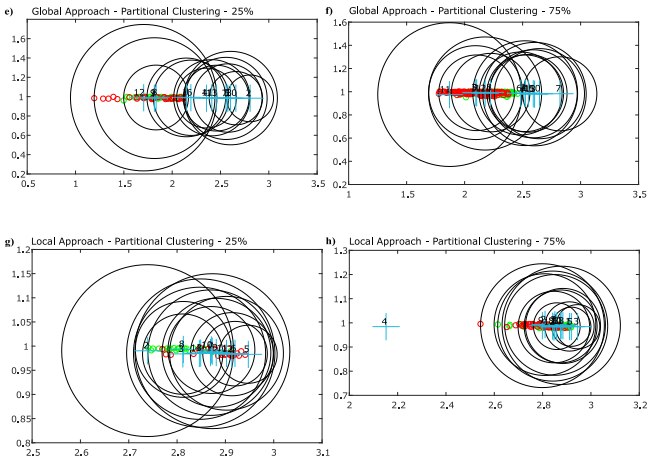

Fig. 15: The clustering results for 13 regions on the F_{25} and F_{75} dataset where a global normalization and partitional clustering e-f), and local normalization and partitional clustering g-h) approach were applied to the whole matrix G in which G_{max} and G_{dens} were considered.

it can identify all 13 brain regions with a relatively small number of clusters. Regarding the results of the *NearGene*, it has to be noted that on the dataset F_{25} (see figure 22) the 13 regions are identified with a fewer number of clusters, compared to the analysis performed on the F_{75} dataset.

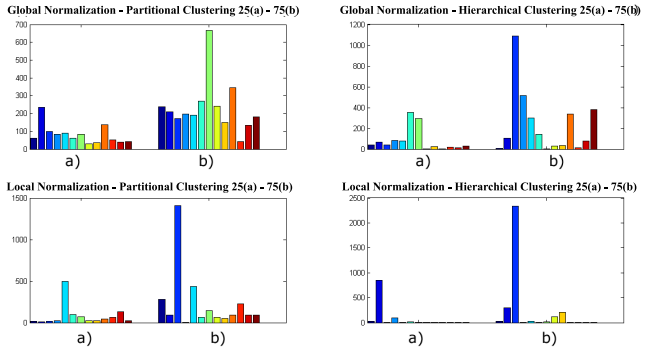

Fig. 16: Plots of gene expression distribution after the cluster analysis on the whole matrix G in which G_{max} and G_{dens} were considered: a) dataset 25%, b) dataset 75%.

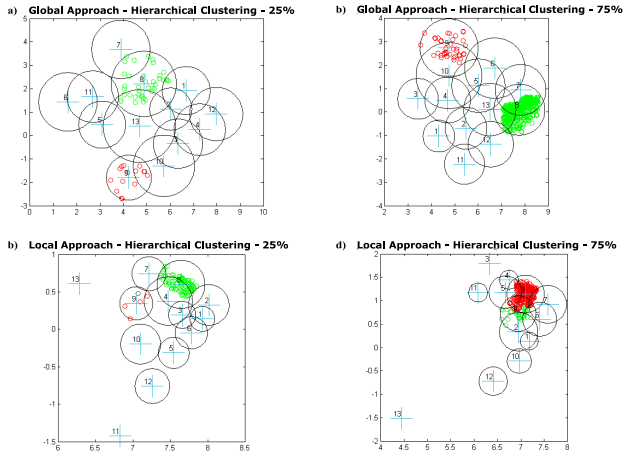

Fig. 17: The results for 13 regions on the F_{25} and F_{75} dataset where a global normalization and hierarchical clustering a-b), and local normalization and hierarchical clustering c-d) approach were applied to the whole matrix G after the PCA method was imposed.

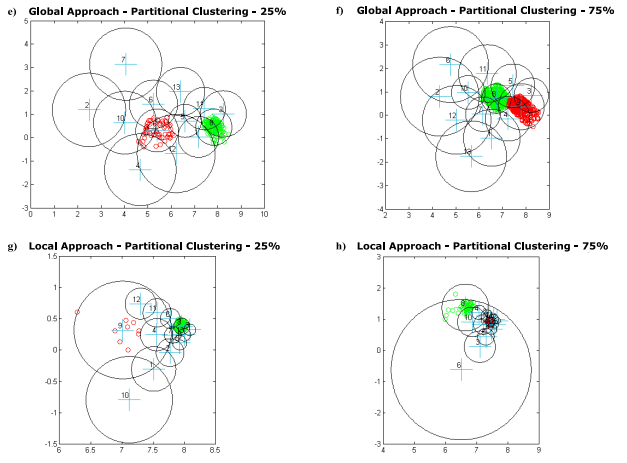

Fig. 18: The results for 13 regions on the F_{25} and F_{75} dataset where a global normalization and partitional clustering e-f), and local normalization and partitional clustering g-h) approach were applied to the whole matrix G after the PCA method was imposed.

The mean of entropy values and also the average of purity depicted in figure 24 show better performance for the global approach and moreover, for both types of clustering, the desired physiological finding is reached first.

The figure 25 contains two Heatmaps, representing the intersection of the cluster voxels with those of the regions 25.(A) and vice versa 25.(B). Highlighted in figure 25, cluster 5 physically locates the retro-hippocampal region; this is evident from the fact that the pixel is highlighted in white; therefore, the cluster/region correlation is very high. Knowing the genes that represent cluster 5, which have been identified only as a function of gene expression and not the spatial position, can be physiologically representative concerning the retro-hippocampal region.

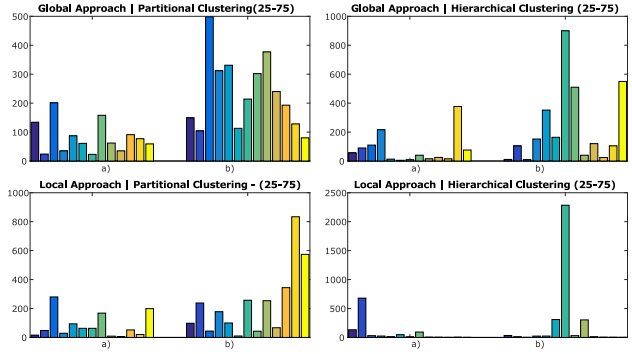

Fig. 19: Plots of gene expression distribution after the cluster analysis on the whole matrix G after a PCA method was considered: a) dataset 25%, b) dataset 75%.

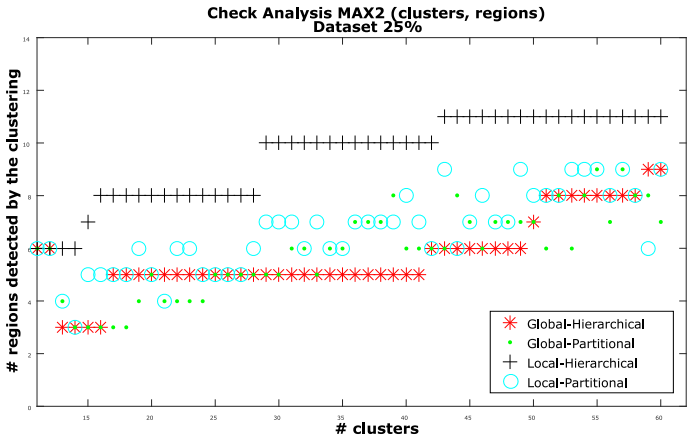

Fig. 20: The MAX2 plot reports the number of unique regions versus the number of clusters on the F_{25} dataset.

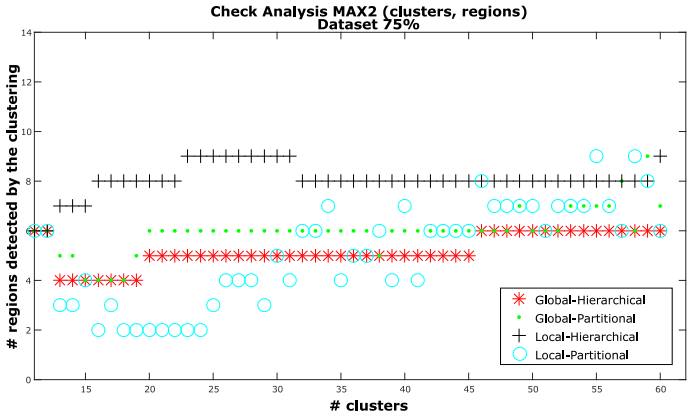

Fig. 21: The MAX2 plot reports the number of unique regions versus the number of clusters on the F_{75} dataset.

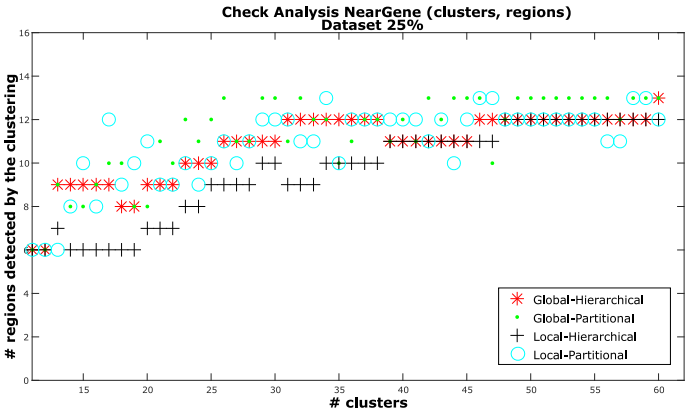

Fig. 22: The NearGene plot reports the number of unique regions versus the number of clusters on the F_{25} dataset.

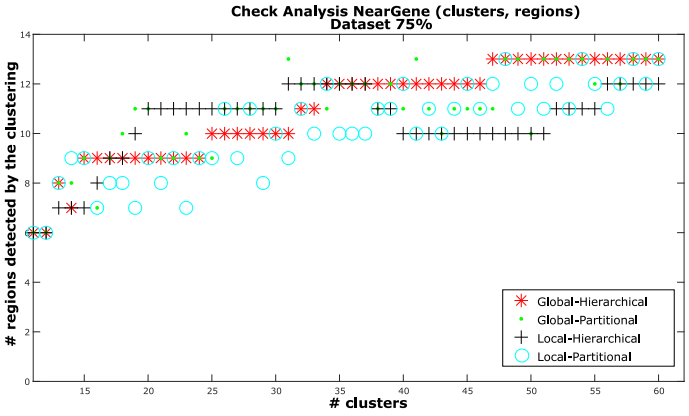

Fig. 23: The NearGene plot reports the number of unique regions versus the number of clusters on the F_{75} dataset.

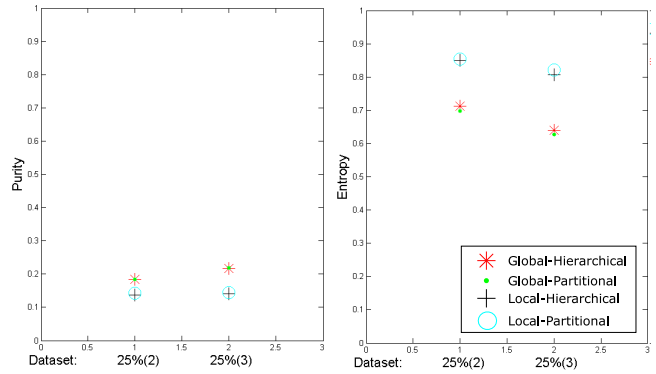

Fig. 24: Average of Purity and Entropy index for dataset F_{25}

Therefore, to the best of our knowledge, we can reasonably state that the obtained results support the purpose of this study. Since by analysing the F_{25} dataset of genes, which are considered the least significant, it is possible to extract equivalent information to the obtained results from the analyses performed on the F_{75} dataset.

7 Acknowledgments

Two grants have supported this contribution: *ICT Tools for the diagnosis of Autoimmune diseases in the Mediterranean Area*, INTERREG 2017/2020 V-A Italia-Malta asse 1.1;

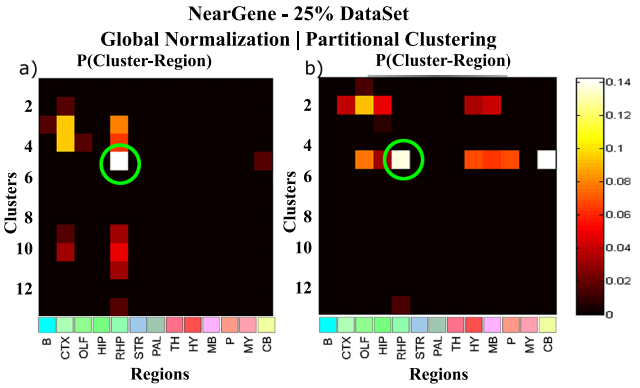

Fig. 25: HeatMap: global - partitional cluster

Mediterranean Center For Human Health Advanced Biotechnologies (MED-CHHAB), P.O.FESR 2007/2013 Asse 1, CUP-B71D11000140007.

8 References

1 Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007;445(7124):168–176.

2 Goldowitz D. Allen Reference Atlas. A Digital Color Brain Atlas of the C57BL/6J Male Mouse - by H. W. Dong. Genes, Brain and Behavior. 2010;9(1):128–128.

3 Ng L, Pathak S, Kuan C, Lau C, w Dong H, Sodt A, et al. Neuroinformatics for Genome-Wide 3-D Gene Expression Mapping in the Mouse Brain. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2007;4(3):382–393.

4 Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270(5235):467–470.

5 Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(12):5463–5467.

6 Shendure J, Ji H. Next-generation DNA sequencing. Nature Biotechnology. 2008;26(10):1135–1145.

7 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;20(9):1297–1303.

8 Hawrylycz M, Ng L, Feng D, Sunkin S, Szafer A, Dang C. The allen brain atlas. Springer; 2014.

9 Bohland JW, Bokil H, Pathak SD, Lee CK, Ng L, Lau C, et al. Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy. Methods. 2010;50(2):105–112.

10 Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America. 1998;95(25):14863–14868.

11 Visel A, Thaller C, Eichele G. GenePaint.org: An atlas of gene expression patterns in the mouse embryo. Nucleic Acids Research. 2004;32(DATABASE ISS.):D552–D556.

12 Gong S, Doughty M, Harbaugh CR, Cummins A, Hatten ME, Heintz N, et al. Targeting Cre recombinase to specific neuron populations with bacterial artificial chromosome constructs. Journal of Neuroscience. 2007;27(37):9817–9823.

13 Evans AC, Janke AL, Collins DL, Baillet S. Brain templates and atlases. NeuroImage. 2012;62(2):911–922.

14 Jones AR, Overly CC, Sunkin SM. The allen brain atlas: 5 years and beyond. Nature Reviews Neuroscience. 2009;10(11):821–828.

15 Carson JP, Thaller C, Eichele G. A transcriptome atlas of the mouse brain at cellular resolution. Current Opinion in Neurobiology. 2002;12(5):562–565.

16 Boline J, Lee EF, Toga AW. Digital atlases as a framework for data sharing. Frontiers in Neuroscience. 2009;2(JUL):100–106.

17 Toga AW, Thompson PM. Brain Atlases of Normal and Diseased Populations. International Review of Neurobiology. 2005;66:1–54.

18 Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT, Baldarelli RM, et al. MGD: The mouse genome database. Nucleic Acids Research. 2003;31(1):193–195.

19 Ng L, Bernard A, Lau C, Overly CC, Dong HW, Kuan C, et al. An anatomic gene expression atlas of the adult mouse brain. Nature Neuroscience. 2009;12(3):356–362.

20 Fürth D, VaissiÃÍre T, Tzortzi O, Xuan Y, MÃd'rtin A, Lazaridis I, et al. An interactive framework for whole-brain maps at cellular resolution. Nature Neuroscience. 2018;21(1):139–153.

21 Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature. 2012;489(7416):391–399.

22 Ballaró B, Florena AM, Franco V, Tegolo D, Tripodo C, Valenti C. An automated image analysis methodology for classifying megakaryocytes in chronic myeloproliferative disorders. Medical Image Analysis. 2008;12(6):703 – 712.

23 Seidlitz J, VÃąÅąa F, Shinn M, Romero-Garcia R, Whitaker KJ, VÃl'rtes PE, et al. Morphometric Similarity Networks Detect Microscale Cortical Organization and Predict Inter-Individual Cognitive Variation. Neuron. 2018;97(1):231–247.e7.

24 Puckelwartz MJ, Pesce LL, Nelakuditi V, Dellefave-Castillo L, Golbus JR, Day SM, et al. Supercomputing for the parallelization of whole genome analysis. Bioinformatics. 2014;30(11):1508–1513.

25 Sargent R, Fuhrman D, Critchlow T, Sera TD, Mecklenburg R, Lindstrom G, et al. The design and implementation of a database for human genome research. In: Proceedings of 8th International Conference on Scientific and Statistical Data Base Management; 1996. p. 220–225.

26 Aziz R, Verma CK, Jha M, Srivastava N. Artificial neural network classification of microarray data using new hybrid gene selection method. International Journal of Data Mining and Bioinformatics. 2017;17(1):42–65.

27 Peng H. Bioimage informatics: A new area of engineering biology. Bioinformatics. 2008;24(17):1827–1836.

28 Thomas JJ, Cook KA. A visual analytics agenda. IEEE Computer Graphics and Applications. 2006;26(1):10–13.

29 BRAIN ARCHITECTURE WEB PORTAL. BAP, editor. brain-toolbox. Cold Spring Harbor Laboratory; 2013. last update October 15,2013. http://brainarchitecture.org/allen-atlas-brain-toolbox.

30 Grange P, Hawrylycz M, Mitra PP. Computational neuroanatomy and co-expression of genes in the adult mouse brain, analysis tools for the Allen Brain Atlas. Quantitative Biology. 2013;1(1):91–100.

31 Sunkin SM, Ng L, Lau C, Dolbeare T, Gilbert TL, Thompson CL, et al. Allen Brain Atlas: An integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Research. 2013;41(D1):D996–D1008.

32 W DH. Allen Reference Atlas. A Digital Color BrainAtlas of the C57BL/6J Male Mouse. Genes, Brain and Behavior. 2010;9:128.

33 Fakhry A, Ji S. High-resolution prediction of mouse brain connectivity using gene expression patterns. Methods. 2015;73:71–78.

34 Grange P, Bohland JW, Hawrylycz M, Mitra PP. Brain Gene Expression Analysis: a MATLAB toolbox for the analysis of brain-wide gene-expression data. ArXiv e-prints. 2012;.

35 Divya, Altaf I. Cluster analysis using gene expression data. In: 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT); 2017. p. 1–6.

36 Wang YX, Gao YL, Liu JX, Kong XZ, Li HJ. Robust Principal Component Analysis Regularized by Truncated Nuclear Norm for Identifying Differentially Expressed Genes. IEEE Transactions on Nanobioscience. 2017;16(6):447–454.

37 Noto T, Barnagian D, Castro JB. Genome-scale investigation of olfactory system spatial heterogeneity. PLoS ONE. 2017;12(5).

38 Jain AK, Dubes RC. Algorithms for clustering data. Prentice-Hall Advanced Reference Series. Prentice Hall PTR; 1988.

39 Huang X, Zhang L, Wang B, Li F, Zhang Z. Feature clustering based support vector machine recursive feature elimination for gene selection. Applied Intelligence. 2018;48(3):594–607.

40 Lupaşcu CA, Tegolo D. Automatic unsupervised segmentation of retinal vessels using self-organizing maps and K-means clustering. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2011;6685 LNBI:263–274.

41 Pluim JPW, Maintz JBA, Viergever MA. Mutual-information-based registration of medical images: A survey. IEEE Transactions on Medical Imaging. 2003;22(8):986–1004.

42 Zhang J, Pei Y, Fletcher GHL, Pechenizkiy M. Structural measures of clustering quality on graph samples. In: Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM'16; 2016. p. 345–348.