algorithms

MDPI

Review

On the Role of Clustering and Visualization Techniques in Gene Microarray Data

Angelo Ciaramella and Antonino Staiano *

Dipartimento di Scienze e Tecnologie, Università di Napoli Parthenope, 80133 Naples, Italy; angelo.ciaramella@uniparthenope.it

* Correspondence: antonino.staiano@uniparthenope.it

Received: 25 May 2019; Accepted: 14 June 2019; Published: 18 June 2019

Abstract: As of today, bioinformatics is one of the most exciting fields of scientific research. There is a wide-ranging list of challenging problems to face, i.e., pairwise and multiple alignments, motif detection/discrimination/classification, phylogenetic tree reconstruction, protein secondary and tertiary structure prediction, protein function prediction, DNA microarray analysis, gene regulation/regulatory networks, just to mention a few, and an army of researchers, coming from several scientific backgrounds, focus their efforts on developing models to properly address these problems. In this paper, we aim to briefly review some of the huge amount of machine learning methods, developed in the last two decades, suited for the analysis of gene microarray data that have a strong impact on molecular biology. In particular, we focus on the wide-ranging list of data clustering and visualization techniques able to find homogeneous data groupings, and also provide the possibility to discover its connections in terms of structure, function and evolution.

Keywords: clustering; data visualization; gene expression data; data mining

1. Introduction

Since the dawn of the new millennium, all science fields have experienced an impressive increase in available data, not only in terms of quantity but also in terms of quality and sharing, thus calling for new theories, techniques and tools to enable scientists and information stakeholders to properly exploit the oceans of distributed and diverse data in knowledge extraction.

There are two primary elements to this shift: on the one side, in biology, astrophysics, social sciences, and in several other fields, traditional interactive information analysis and information visualization techniques have proven to be far insufficient to deal with data sets defined by enormous quantities and complexity (very high dimensionality, i.e., number of features). Second, the concurrent assessment of hundreds of features may reveal earlier unidentified patterns leading to a greater comprehension of the fundamental process dynamics and trends. Therefore, the field of Knowledge Discovery in Databases or KDD [1] is becoming of crucial significance not only in its traditional arena but also as an additional instrument for nearly all study areas.

Specific information mining techniques for model identification and analysis [2] are at the heart of the method. In biology, researchers have been effective in cataloging genes through DNA sequencing initiatives and can now produce large amounts of information on gene expression using microarrays.

Microarrays of gene expression, the growth of which began in the second quarter of the 1990s, have a strong effect on molecular biology. Indeed, while the capacity to determine a single gene expression is not new, the capacity to assess the expression of all DNA in an individual at once is the latest development and leads to fresh techniques of diagnosis and therapy for several kinds of illnesses. It is also becoming progressively apparent, however, that merely producing the data is not sufficient and extracting the appropriate information is far from trivial.

Over the past two decades, a great deal of jobs have concentrated on developing machine learning approaches suitable for dna information assessment [3–6].

Clustering is one of the main instruments for analyzing genetic data [7]. The nature of the problem makes scientists consider clustering as an instrument not only to find coherent and consistent groupings in data, but also to relate to one another the elements of structure, functions and development.

Many findings in experimental biology first occur as an image of an organism, cells, or microarray scans. As the amount of these outcomes accelerates, it becomes essential to automatically extract features and significance from those images.

Naïve 2*D* or 3*D* visualizations alone at the other end of the information pipeline are insufficient to explore genetical data. Biologists need visual instruments that promote the exploration of high-dimensional information based on many parameters. Data visualization is fundamental for extracting helpful information from big amounts of raw data. The human eye and brain together constitute an amazing model recognition tool, but for them to function, the data must be presented in a low-dimensional space, commonly two or three dimensions. Even a very simple association may appear very vague when the information is displayed in tabular type, but visual testing often makes it very easy to see. Visualization in bioinformatics needs a lot of research efforts for developing tools that enable to tackle the challenges of the present and future genomics and proteomics. These arguments motivate the content of the paper which provides a review of the recent developments in two specific data mining research directions, namely, clustering and visualization of data, in the perspective of genetic applications.

1.1. A Biological Introduction: Microarray Gene Expression Technology

Biologists can now acquire a a pattern of gene expression equivalent to the reaction of the organism to a specific experimental situation. The technique of microarray offers a way for simultaneous identification of the expression of several (or the whole set) of genes of an organism at any fixed time instant [8], generating models of gene activity that retain dynamic information about a cell function. This information is crucial if complicated cell relationships are to be investigated. They also have very significant applications in pharmaceutical and clinical studies, in addition to the huge science contents of microarrays in the basic study of expressions, regulations and interactions of genes. The *cDNA* microarray [9,10] oligonucleotide arrays are the two main kinds of microarray studies. Despite variations in the details of their experimental protocols, three prevalent fundamental processes are involved in both kinds of tests: [11]:

Chip manufacture: A microarray is a tiny chip (produced of chemically covered glass, nylon mesh, or silicon) on which tens of thousands of molecules (samples) of DNA are connected in set grids.

Target preparation, labeling, and hybridization: Usually, two samples of mRNA (i.e., test and control samples) are backward transcribed into cDNA (targets), marked with either fluorescent or radioactive isotopics, and then hybridized with the samples on the chip surface.

The scanning process: To read the signal intensity emitted from the marked and hybridized targtes, chips are scanned.

In general, both cDNA microarrays and oligonucleotide arrays experiments evaluate the amount of expressions for each DNA sequence by the signal intensity proportion between the test and control samples, so that the data sets generated from both techniques contain the same biological information content [12].

Microarray Experimental Data

A microarray experiment typically assesses a large number of DNA sequences (genes, cDNA clones, or expressed sequence tags [ESTs]) under multiple conditions. These conditions may be a time series during a biological process (e.g., the yeast cell cycle) or a collection of different tissue samples (e.g., normal versus cancerous tissues). From now on, we will refer to gene expression data without making a distinction among DNA sequences, which will uniformly be called "genes". In the same way, we will refer to all kinds of experimental conditions as "samples". A gene expression data set from a microarray experiment can be represented by a real-valued expression matrix $M = \{a_{ij} | 1 \leq i \leq n, 1 \leq j \leq m\}$, where the rows form the expression patterns of genes, the columns represent the expression profiles of samples, and each cell a_{ij} is the measured expression level of gene i in sample j. The original gene expression matrix obtained from a scanning process contains noise, missing values, and systematic variations arising from the experimental procedure. Data preprocessing is indispensable before any cluster analysis can be performed. Some problems of data preprocessing have themselves become interesting research topics. Those questions are beyond the scope of this paper; however, the interested reader may refer to [13,14] for an examination of the problem of missing value estimation and to [15,16] for addressing the problem of data normalization or data compression (e.g., compressive sensing) [17,18].

In literature, there is a huge number of available gene expression data sets, aimed at different biological and medical goals, the most of which are publicly available for experimenting with. In the following, a description of a brief list of usefeul data sets (see Table 1 for a summary) for assessing the suitability and the effectiveness of various clustering, visualization and, in general, machine learning techniques. The interested reader may refer to [19] for more extensive references.

- ALLAML [20] contains two classes of samples, namely ALL and AML, each of 47 and 25 samples, respectively, for an overall number of 72 samples. Each sample is formed by 7129 gene expression values.
- LEUKEMIA [20] contains in total 72 samples in two classes: acute lymphoblastic and acute myeloid corresponding to 7129 genes.
- CLL SUB 111 [21] dataset has gene expressions from high density oligonucleotide arrays containing genetically and clinically distinct subgroups of B-cell chronic lymphocytic leukemia (B-CLL). The dataset consists of 11,340 gene expression levels, 111 instances and three classes.
- GLIOMA [22] contains in total 50 samples in four classes: cancer glioblastomas, non-cancer glioblastomas, cancer oligodendrogliomas and non-cancer oligodendrogliomas, which have 14, 14, 7, 15 samples, respectively. Each sample is formed by 12,625 genes.
- LUNG [23] contains in total 203 samples in five classes, adenocarcinomas, squamous cell lung carcinomas, pulmonary carcinoids, small-cell lung carcinomas and normal lung, with 139, 21, 20, 6, 17 samples, respectively.
- LUNG DISCRETE [24] contains 73 samples in seven classes where each sample consists of 325 gene expressions.
- DLBCL [25] is a modified version of the original DLBCL dataset. It consists of 96 samples in nine classes, where each sample is defined by the expression of 4026 genes.
- CARCINOM [26] contains 174 samples in 11 classes, prostate, bladder/ureter, breast, colorectal, gastroesophagus, kidney, liver, ovary, pancreas, lung adenocarcinomas and lung squamous cell carcinoma.

Table 1. Gene expression Datasets Description.

	Size	**# of Features**	**# of Classes**
ALLAML	72	7129	2
LEUKEMIA	72	7070	2
CLL_SUB_111	111	11,340	3
GLIOMA	50	4434	4
LUNG_C	203	3312	5
LUNG_D	73	325	7
DLBCL	96	4026	9
CAR	174	9182	11

2. Clustering for Microarray Gene Expression Data

In the last few years, the scientific community has witnessed the proliferation of a large number of machine learning techniques [27] and, in particular, of clustering methods [12] for applications to bioinformatics. Each clustering technique shows some interesting features and all together share the same aspect: none of them is the "best" clustering algorithm for all type of data. Thus, it is necessary to orient oneself in this jungle of clustering methodologies. With this aim, we give an overview of "the state of the art" in the field of clustering algorithms in bioinformatics stressing the potentialities and the deficiencies of each method, and highlighting the needs that must be satisfied by novel clustering methods.

As we learn from many clustering books, the clustering activity can be synthetically described as the art of finding the groups in data. In other words, the primary thrust of clustering is to arrange a collection of data into a small number of groups (clusters) so that the elements that are similar become allocated to the same group. The elements (patterns) that are quite distant should be placed into separate categories. The "distance" function measures the level of similarity; the smaller the distance between two samples, the greater the level of their similararity. There are literally hundreds of clustering techniques well reported in the literature and presented with benchmarks thoroughly chosen. Obviously, they all can not be discussed and their main benefits and disadvantages contrasted.

Nevertheless, three particular kinds of clustering algorithms can be identified: those based on an effort to discover the appropriate partitioning in a defined number of clusters, those based on a hierarchical effort to discover clusters composition, and those based on a probabilistic template for the fundamental cluster structures. In the next three parts, we address each of these in turn.

2.1. Partitive Clustering Algorithms

In this form of clustering [28], the aim is to find a partition in a data set that groups the samples into K non-overlapping sets such that all samples belonging to the same set are as "similar" as possible, that is, given the set of N data points $S = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N\}$, with $\mathbf{t}_i \in \mathbb{R}^D$, for $i = 1, 2, \ldots, N$, the algorithm task is to find K clusters $C = \{C_1, C_2, \ldots, C_K\}$ such that each data point $\mathbf{t}_i$ is assigned to a cluster C_k. An appropriate objective function, such as minimizing the distance between each point and the centroid $\mathbf{v}_i \in \mathbb{R}^D$, of the cluster to which it is assigned, captures similarity. Very often, a sum of variance criterion of the following form

$$J = \sum_{i=1}^{K} \sum_{j=1}^{N} u_{ij}^m \|\mathbf{t}_j - \mathbf{v}_i\|^2 = \sum_{i=1}^{K} \sum_{j=1}^{N} u_{ij}^m d_{ij}^2 \tag{1}$$

is adopted, where d_{ij} is any distance measure, e.g., Euclidean distance. The partition matrix $U = [u_{ij}]$ is used to store all results of clustering (partitioning) the points into clusters. The partition matrix meets the following requirements, depending on whether we are interested in set-oriented or fuzzy set-oriented partitioning:

- for set-oriented clustering, $u_{ij} \in \{0, 1\}$, $0 < \sum_{j=1}^{N} u_{ij} < N$, for $i = 1, 2, \ldots, K$, $\sum_{i=1}^{K} u_{ij} = 1$, for $j = 1, 2, \ldots, N$,
- for fuzzy-oriented clustering, we get $u_{ij} \in [0, 1]$ with the same two requirements as stated above.

The above requirements suppose an intuitive and simple interpretation:

- each cluster is nonempty;
- each point belongs exactly to one cluster (set-oriented clustering) or might belong to more than one cluster simultaneously, i.e., the sum of its membership values for all clusters is 1 (fuzzy set-oriented clustering).

The optimization is carried out through an iterative process in which the objective is to calculate the prototypes ($\mathbf{v}_i$) and update the partition matrix (U) on the basis of the minimized objective function, J, first-order conditions. The entire process will be ended when a stopping condition is met.

K-means algorithm [29] is a typical partitive clustering technique. Given an a priori fixed value of K, K-means divides the data set into K disjoint clusters which minimize the objective function of Equation (1), where $m = 1$ and $u_{ij} \in \{0, 1\}$. K-means algorithm is simple and fast, converging in just a few iteration steps.

Unfortunately, when used as a gene-based clustering algorithm, it exhibits several shortcomings. To begin with, in a collection of gene expressions, the number of gene groupings is generally unknown beforehand. Users generally perform the algorithms several times, each time with a different value for K, and compare the clustering outcomes to find the optimum cluster number.

This comprehensive fine-tuning procedure may not be practical for a gene expression data set containing thousands of genes. Moreover, gene expression data typically involve an enormous quantity of noise; nonetheless, K-means forces each gene into a cluster that can lead the algorithm to be noise prone [30,31].

In the past, several clustering algorithms [31–33] were suggested to solve the disadvantages of the K-means algorithm. These techniques certain global parameters to regulate the performance of the resulting clusters (e.g., the maximum radius of a cluster and/or the minimum distance between clusters), therefore clustering consists of extracting only the adequate clusters from the data set. In so doing, the number of clusters can be determined automatically and samples not belonging to any adequate cluster are considered outliers. However, the quality of each cluster in gene data may differ extensively. Thus, choosing suitable global constraining parameters is often a challenging issue.

2.2. Hierarchical Clustering

The primary characteristic of hierarchical techniques is that they gradually combine points or separate superclusters. Indeed, on these grounds, we can recognize two diverse kinds of hierarchical techniques: *agglomerative* and *divisive*. In the following discussion, we focus mainly on the agglomerative strategy. The rationale is to begin with a number of clusters corresponding exactly to the cardinality of a data set, and then by merging continuously to reduce their number. Let us assume that the set of data points S consists of N samples, $S = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N\} \subset \mathbb{R}^D$. Notwithstanding their variety, any hierarchical clustering technique is conceptually the same and can be defined in a concise manner as follows: begin with N clusters by assigning each sample to a distinct cluster and continue with this original cluster setup by combining the nearest clusters. Namely, if H and T are the identified two closest clusters, consider a new cluster $\{H, T\}$ and decrease by one the total number of clusters. The whole procedure repeats until a minimal number of the clusters has been reached. The result of these steps is a partition hierarchy of the data points. Despite the ease of the strategy, defining how to compute a distance between a pair of clusters in not immediate. Among the number of alternatives to choose from, the three most common policy employed in practice are [34–36]:

- single linkage method: The closeness between H and T is calculated on the basis of the minimum distance between the samples that belong to the respective clusters.
- complete linkage method: The closeness between H and T is calculated as the distance between the most two distant samples, one from each cluster.
- average linkage method: The closeness between H and T is calculated as the average of all the distances between pairs of samples, one from each cluster. The criterion considers all possible pairs of distances between samples in the clusters, and is thus far more accurate and resistant to outliers.

The computational burden depends on the size of the data. More importantly, significant memory demands exist as the consecutive outcomes (clusters) must be maintained at each stage of the technique.

Unlike partition-based clustering, which tries to straightly decompose a data set into a collection of disjoint groups, hierarchical clustering produces a hierarchical sequence of nested groups that can be graphically depicted by a tree called a dendrogram. A dendrogram's branches not only keeps track of the clusters' development, but also shows the cluster closeness. One can get a given number of clusters by cutting the dendrogram at a desired level.

An important important application of an agglomerative strategy to microarray data are the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [37], where the authors conceived a technique of representing the clustered data set graphically. In particular, each cell of the matrix of gene expression is colored based on the measured fluorescence ratio and the matrix rows are reordered based on the structure of the hierarchical dendrogram and a rule for the node ordering. Once the clustering procedure is performed, a colored table, i.e., an image of the clusters, represents the initial gene expression matrix where big adjacent color patches reflect clusters of genes that share an analogous pattern of expressions over multiple conditions. The method described in [37] was much favored by many biologists and has become a popular tool in gene expression data analysis [25,37–39]. Although some effective variant was proposed as, for instance, in dynamic agglomerative clustering [40], standard agglomerative approaches are, however, not sufficiently robust [41], meaning that even little perturbations in the data set could significantly alter the hierarchical dendrogram structure. Furthermore, hierarchical policies have high-computational costs; indeed, the cost of a standard agglomerative hierarchical procedure is $O(N^2 logN)$ [42].

2.3. Model-Based Clustering

In this framework [43–46], a statistical basis is provided to model the cluster structure of gene expression data and it is assumed that the data come from a multivariate finite mixture model of the general form

$$f(\mathbf{t}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{t};\theta_k), \tag{2}$$

where f_k are the component distributions, π_k and θ_k are the component probabilities (the probability that an observation belongs to the kth component) and the component parameters, respectively. Roughly speaking, the general procedure is as follows: given a data set $S = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N\}$, determine how many clusters K one wants to fit to the data, choose parametric models for each of these K clusters (for example, multivariate Normal distributions in which the mean, μ_k, and the covariance, Σ_k, are the parameters), and eventually employ an EM (Expectation Maximization) [47] scheme for determining, from data, parameters and probabilities for each component. When EM convergence occurs, a sample is assigned to the component having maximum conditional probability.

Gene expression data are typically extremely related, so there may be cases where there is a strong correlation between a single gene and two distinct clusters. Thus, model-based clustering probabilistic nature is especially appropriate for gene expression data. Model-based clustering, however, assumes that your data fits a given distribution, even though, often, this assumption does not hold. Modeling gene expression data are a continuing endeavor by many scientists, and there is presently no well-established model of representation. For instance, Yeung et al. [46] studied several data transformation used in practice to assess, among a number of gene expression data sets, the level to which the data fits a multivariate Gaussian model. The authors ascertained that in all cases there is a poor fit with the Gaussina model and there is no uniform rule to suggest a transformation that would best improve this fit.

2.4. Other Approaches

Other several clustering approaches have been proposed. Some approaches are used not only for clustering but also for data projection and visualization. We can list the following approaches:

- **Self Organizing Maps**. The k-means method is a well-known centroid approach. A neural variation that allows samples to influence the location of neighboring clusters is known as the self-organizing map (SOM) or Kohonen map [48]. Such maps, usually a $2D$ rectangular grid of neurons, are particularly valuable for describing the relationships between clusters. The neurons of the neural network are all connected with their own reference vector, and each data point is mapped to the neuron with the closest reference vector. During the training steps, each data point directs the movement of the reference vectors towards the denser areas of the input vector space, so that those reference vectors are trained to fit the distributions of the input data set. When the training is complete, clusters are identified by mapping all data points to the output neurons. SOM was applied to gene expression data in [41] with good results over the k-means approach, nonetheless it requires an a priori clusters number and lattice structure of the neural network.
- **Graph-Theoretical Clustering**. Graph-theoretical clustering techniques are explicitly presented in terms of a graph, thus converting the problem of clustering a data set into such graph theoretical problems as finding minimum cut or maximal cliques in the proximity graph G. In [49], the CLuster Identification via Connectivity Kernels (CLICK) was proposed. CLICK tries to discover highly connected components in the proximity graph as clusters. In [49], the authors applied their CLICK clustering method to public gene expression data demonstrating better quality in terms of homogeneity and separation compared with other methods.

  In [50], both a theoretical algorithm and a practical heuristic called CAST (Cluster Affinity Search Technique) is presented. CAST takes as input a real, symmetric, $N \times N$ similarity matrix $Sim(Sim(i,j) \in [0,1])$ and an affinity threshold ϵ. CAST alternates between adding high affinity samples to a given cluster and removing low affinity samples from it. CAST does not require the number of clusters and is effective in handling outliers. Nevertheless, for CAST, it is difficult to determine a proper value for the global parameter ϵ.
- **Density-Based Hierarchical Approaches**. In [51], a density-based, hierarchical clustering method (DHC) was proposed in order to identify the co-expressed gene groups from gene expression data. As the name suggests, DHC combines both the the model-based and hierarchical clustering approaches. DHC is effective in detecting the co-expressed genes (which have relatively higher density) from noise (which have relatively lower density), and thus is robust in the noisy environment. However, DHC is not efficient from the computational complexity point of view and exhibits the typical difficulty to determine the appropriate value of its parameters. A different approach, called NEC, was defined in [52]. The authors argued that most of clustering algorithms proposed in the literature were based on the Euclidean metric, even though Euclidean metric is often limited and inadequate. NEC is a clustering method accomplished in two steps, where a Probabilistic Principal Surfaces approach (i.e., a density-based modelling) [2] is firstly used to find an initial rough clusterization (with a large number of clusters), and, secondly, it begins an agglomeration phase on the basis of specific non-Euclidean metric defined in terms of Fisher's and Negentropy information. The computational burden of NEC is limited due to Fisher information and Negentropy, thus the technique can efficiently and effectively be applied to gene expression data [53,54] and, with some generalization, to other kinds of data [55].
- **Biclustering**. Biclustering [56,57], also named subspace clustering, aims at finding a subgroup of genes with similar expressions belonging to a subset of samples. Rows, or genes, and columns, or samples, of a gene expression matrix are clustered simultaneously. The rationale in using biclustering is that, among the large number of genes, only a subset contributes to the target in which a researcher is interested in, while the remaining ones might mask the role of relevant genes in pursuing that target. Furthermore, it has been argued that coexpressed genes might behave independently.
- **Multi-objective evolutionary clustering algorithms**. A very recent trend in clustering gene expression data tries to overcome two main deficiencies in clustering techniques when facing with different molecular data sets, namely, (i) the impossibility to discriminate among the importance

of features during the cluster formation; indeed, different features could have, and frequently do actually have, different effects on clustering, and (ii) the lack of multiple internal evaluation functions. To this end, in [19], a multiobjective framework has been proposed in order to gain robustness when facing several and different molecular data; to this end, the authors select five diverse group validity indices as multiobjective functions simultaneously optimized, in order to properly seize multiple characteristics of the evolving agglomerations.

3. Visualization Techniques for Microarray Gene Expression Data

When coping with a large amount of data, visualization is essential for developing good models. Scientists and decision makers need visualization facilities that help explore very high-dimensional data to extract useful information from it [2]. Many algorithms for data visualization have been proposed by both neural computing and statistics communities, most of which are based on a projection of the data onto a two- or three-dimensional visualization space. We briefly review some of these advanced visualization techniques [58], and show an integrated environment, fully based on PPS, for visualizing high dimensional biomedical data in a 3*D* space. Finally, we will see how the NEC framework provides some nice visualization plots based on its dendrogram.

3.1. Visualization Methods: A Brief Review

- **Principal Component Analysis (PCA)** [59]. A well established linear projection technique is used to map data from higher to lower dimensional spaces. PCA linearly transforms data, preserving as much as possible its variance.
- **Probabilistic PCA** [60]. The lack of a generative model in PCA gives no means to interpret its error function in a principled way. Probabilistic Principal Component Analysis (PPCA) was introduced to enhance PCA, i.e., turning PCA into a generative model by using a latent variable approach. PPCA consists of a Gaussian mixture model where each component has a diagonal covariance matrix with a single variance parameter for describing the variance in each Gaussian of the mixture.
- **Mixture of PPCA** [61]. Because PCA describes only a linear data projection, it is a method that is rather restricted. Using a set of local linear models is one way around this. This is attractive since each model is easier to comprehend and generally easier to accommodate.
- **Multidimensional Scaling (MDS)** [62]. MDS maps data points from an original high dimensional space to spaces of a lower dimension, likewise several other methods, but approaching the problem differently, i.e., on the basis of the dissimilarities between data points rather than the points themselves. In particular, MDS tries to find a lower-dimensional representation of the data preserving the pairwise distances as much as possible. A variant of MDS is the so-called Sammon mapping [62,63].

PCA, PPCA and mixture of PPCA are appropriate when the data are linear or approximately piecewise linear. An alternative approach is to use global nonlinear methods such as SOM in fact, due to its ease and its several plotting choices [48,63,64], it was used for a broad number of applications.

However, SOM also does not provide any probability density function, suffering from other disadvantages that can be overcome using nonlinear, latent variable models such as the Generative Topographic Mapping (GTM) [65]. These latter models are able to effectively show the composition of simple and complex data sets, even though they are less effective when dealing with too complex data sets.

Even if it is nonlinear, a single two-or three-dimensional projection may be inadequate to catch all of the information of interest within a data set. For instance, a projection that is capable of better separating two clusters might not be the one that better reveals the structure within one of those clusters. Hierarchical models, comprising multiple 2D or 3D visualization spaces, come into play here. The idea is to display an entire data set at the top-level projection in the hierarchy, where a clustering structure is possibly revealed, whereas lower-level projections enable displaying internal structures within single clusters, i.e., an arrangement of sub-clusters not evident in higher-level projections. The interested reader may refer to [66,67] for the hierarchical versions of linear latent variable model and of GTM, respectively.

3.2. Visualization Based on Probabilistic Principal Surfaces

We focus on the visualization capabilities of PPS mentioned earlier when describing NEC clustering in Section 2.4. PPS is a spherical version of the latent variable model defined in GTM (for details on latent variable models and GTM, the reader may refer to [65,68], in particular, if the latent space is chosen to be three-dimensional, then it is possible to construct a spherical manifold uniformly arranging PPS nodes $\{\mathbf{x}_m\}_{m=1}^{M}$, on a sphere surface in the $\mathbb{R}^3$ latent space, where a set of latent basis functions are evenly located on the sphere with a lower density [69]. Indeed, the sphere gives us a continuous manifold where data can be projected, and it is especially adequate to visualize high-dimensional data due to its natural tendency for characterizing data sparsity and periphery particularly at increasing dimensionality (curse of dimensionality). This way, a decision maker is capable of (a) projecting, visualizing, rotating and manipulating his data on the surface of the sphere; (b) carrying out deeper analysis by interacting with data and discovering areas of interest on the sphere; and (c) choosing points of interest, looking at their neighbors, analogous points, and showing all related information, etc.

After a spherical PPS model is fitted to the data, the data themselves are projected into the latent space as points onto a sphere. Furthermore, some important information is obtained by drawing the data density function (i.e., through the responsibilities for each latent variable) with varying intensity on the spherical manifold [2]. The sphere will comprise areas that are denser than others and this visual information is easy to see and understand (Figure 1); interestingly, denser areas might suggest the existence of more clusters, thus requiring further investigations. Given a set of data samples $\mathbf{t}_i$, $i = 1, \ldots, N$, the latent manifold coordinates $\hat{\mathbf{x}}_i$ of each data point $\mathbf{t}_i$ are computed as

$$\hat{\mathbf{x}}_i \equiv \langle \mathbf{x} | \mathbf{t}_i \rangle = \int \mathbf{x} p(\mathbf{x}|\mathbf{t}) d\mathbf{x} = \sum_{m=1}^{M} r_{mi} \mathbf{x}_m,$$

where r_{mi} are the latent variable responsibilities, defined as

$$\begin{aligned} r_{mi} = p(\mathbf{x}_m|\mathbf{t}_i) &= \frac{p(\mathbf{t}_i|\mathbf{x}_m)P(\mathbf{x}_m)}{\sum_{m'=1}^{M} p(\mathbf{t}_i|\mathbf{x}_{m'})P(\mathbf{x}_{m'})} \\ &= \frac{p(\mathbf{t}_i|\mathbf{x}_m)}{\sum_{m'=1}^{M} p(\mathbf{t}_i|\mathbf{x}_{m'})}. \end{aligned}$$

The responsibility r_{mi} corresponds to the posterior probability that that the ith data point was generated by the mth component. Since $\|\mathbf{x}_m\| = 1$ and $\sum_m r_{mi} = 1$, for $i = 1, \ldots, N$, these coordinates lie within a unit sphere, i.e., $\|\hat{\mathbf{x}}_i\| \leq 1$.

Cluster Visualization

Another helpful plot for an information miner is cluster plotting and the data points therein. All of the plots on the spherical manifold can be viewed readily by moving the globe interactively. Once a decision maker has a general view of the number of clusters on the globe, he can use this information to discover the clusters by using agglomerative hierarchical clustering on the Gaussian centers in the data space. The points for which the corresponding centers are in the same group belong to the same cluster. The point projections into the latent space are then used to display the groupings onto the latent sphere [2,54] (see Figure 2). PPS has been used effectively in a number of microarray data, such as in a human cancer cell line (HeLa) using cDNA microarrays [53,70], for studying cell cycle in yeast [13,71].

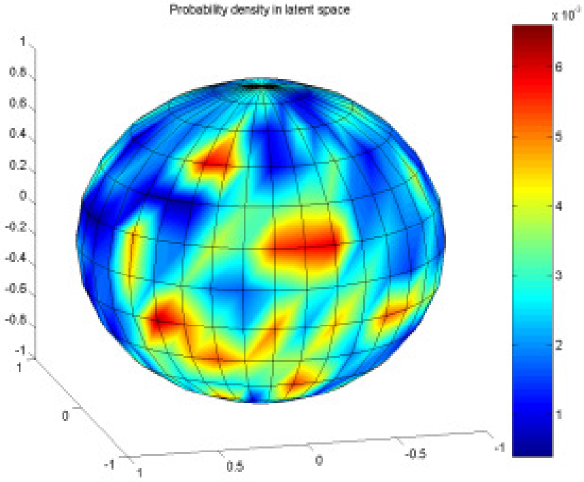

Figure 1. Latent variable responsibilities onto the spherical latent space. Note how the red areas correspond to higher density locations.

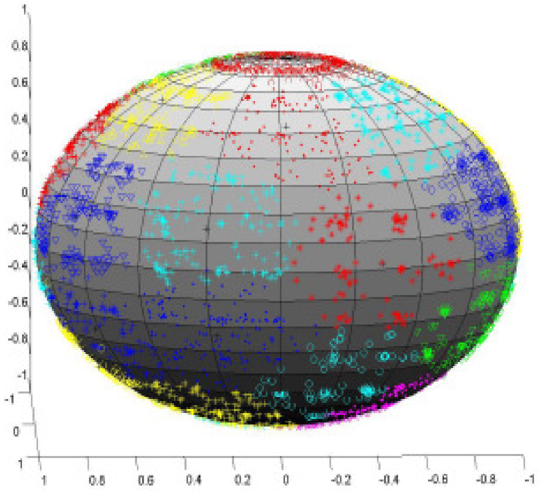

Figure 2. Visualization of resulting clusters from a hierarchical approach.

4. Conclusions

Among machine learning methods, clustering is one of the most important instruments to study gene microarray data. However, a plethora of clustering techniques also exists specifically tailored for gene microarray data, thus making it difficult for a data scientist to effectively choose the best one for the particular problem at hand. Another effective tool for studying gene microarray data consists of data visualization techniques, which permit to extract useful information from large quantities of raw data. Biologists need a visual environment that facilitates exploring high-dimensional data dependent on many parameters. Indeed, the human eye and brain together constitute an amazing model recognition tool, but, for them to function, the data must be presented in a low-dimensional space, commonly two or three dimensions. Even a very simple association may appear very vague when the information is displayed in tabular type, but visual testing often makes it very easy to see. Nonetheless, as it emerges from early discussions, molecular data are complex, and it calls for tackling several sub-problems in order to develop effective tools for data analysis, e.g., effective preprocessing techniques for filtering out noise from gene expression data, feature selection methods to find important features and reduce the very high dimensionality of molecular data. Furthermore, there is a huge effort by researchers to develop different clustering strategies to make effective their approaches in a variety of etherogeneous data sets rather than for an ad-hoc case study. Unfortunately, we are still far from discovering ”The Master Algorithm” [72] able to effectively address whatever problem at hand and to solve it, and that's why the literature on the topic is endless. However, some interesting lines could be drawn. Firstly, from an analytical point of view, one of the main difficulties in analyzing gene expression data are how to tackle the very high dimension of data sets, which leads to both the curse of dimensionality issue in addition to making it complex to properly assess the relative weight of each feature in determining the evolution of clusters. Visualization techniques suffer very high dimensions because the loss of information when compressing data to two or three dimensions. While feature selection could aid in reducing the size of parameters, it could be more beneficial to analyse in a systematical and formal way the dimensionality of data sets by means of techniques to study the intrinsic dimension of data [73], consisting of dimensionality reduction methods that aim at projecting the original data set of dimensionality N, without information loss, onto a lower M-dimensional submanifold. Since the value of M is unknown, techniques that allow knowing in advance the value of M, called intrinsic dimension estimation, might help.

Secondly, complete framework for data analysis, encompassing preprocessing, feature selection, clustering and visualization methods might be of a certain amount of help, thus an increasing number of such works, as described in [74,75] have been proposed in recent years. In more detail, in [74], a computational method that uses a spatial reference map for inferring the spatial location of cells, in the context of complex and heterogeneous tissues, is proposed, whereas, in [75], has been developed a scalable toolkit for analyzing single-cell gene expression data, comprising techniques for preprocessing, visualization, clustering, trajectory inference, differential expression testing, and simulation of gene regulatory networks.

Thus, we have stressed on the necessity of clustering and visualization approaches for gene expression data, combined in a unified framework, also providing a brief review of potential clustering approaches and visualization techniques among which a biologist might choose for performing his data analysis tasks.

Author Contributions: A.C. and A.S. equally contributed to manuscript Conceptualization, Writing, and Reviewing.

Funding: This research received no external funding.

Acknowledgments: This work was partially funded by the University of Naples Parthenope (*Sostegno alla ricerca individuale per il triennio 2016–2018* project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hand, D.; Mannila, H.; Smyth, P. *Principles of Data Mining*; The MIT Press: Cambridge, MA, USA, 2001.
2. Staiano, A.; De Vinco, L.; Ciaramella, A.; Raiconi, G.; Tagliaferri, R.; Longo, G.; Miele, G.; Amato, R.; Del Mondo, C.; Donalek, C.; et al. Probabilistic principal surfaces for yeast gene microarray data-mining. In Proceedings of the ICDM'04 Fourth IEEE International Conference on Data Mining Brighton (UK), Brighton, UK, 1–4 November 2004; pp. 202–209.
3. Calcagno, G.; Staiano, A.; Fortunato, G.; Brescia-Morra, V.; Salvatore, E.; Liguori, R.; Capone, S.; Filla, A.; Longo, G.; Sacchetti, L. A multilayer perceptron neural network-based approach for the identification of responsiveness to interferon therapy in multiple sclerosis patients. *Inf. Sci.* **2010**, *180*, 4153–4163. [CrossRef]
4. Camastra, F.; Di Taranto, M.D.; Staiano, A. Statistical and computational methods for genetic diseases: An overview. *Comput. Math. Methods Med.* **2015**, *2015*, 954598. [CrossRef] [PubMed]
5. Di Taranto, M.D.; Staiano, A.; D'Agostino, M.N.; D'Angelo, A.; Bloise, E.; Morgante, A.; Marotta, G.; Gentile, M.; Rubba, P.; Fortunato, G. Association of USF1 and APOA5 polymorphisms with familial combined hyperlipidemia in an Italian population. *Mol. Cell. Probes* **2015**, *29*, 19–24. [CrossRef] [PubMed]

6. Staiano, A.; Di Taranto, M.D.; Bloise, E.; D'Agostino, M.N.; D'Angelo, A.; Marotta, G.; Gentile, M.; Jossa, F.; Iannuzzi, A.; Rubba, P.; et al. Investigation of single nucleotide polymorphisms associated with familial combined hyperlipidemia with random forests. *Neural Nets Surround.* **2013**, *19*, 169–178.
7. Pirim, H.; Ekşioğlu1, B.; Perkins, A.; Yüceer, Ç. Clustering of High Throughput Gene Expression Data. *Comput. Oper. Res.* **2012**, *39*, 3046–3061. [CrossRef] [PubMed]
8. Heath, L.S.; Ramakrishnan, N.; Sederoff, R.R.; Whetten, R.W.; Chevone, B.I.; Struble, C.A.; Jouenne, V.Y.; Chen, D.; van Zyl, L.; Grene, R. Studying the Functional Genomics of Stress Responses in Loblolly Pine with the Expresso Microarray Experiment Management System. *Comp. Funct. Genom.* **2002**, *3*, 226–243. [CrossRef]
9. Lockhart, D.J.; Dong, H.; Byrne, M.C.; Follettie, M.T.; Gallo, M.V.; Chee, M.S.; Mittmann, M.; Wang, C.; Kobayashi, M.; Horton, H.; et al. Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays. *Nat. Biotechnol.* **1996**, *14*, 1675–1680. [CrossRef]
10. Schena, M.D.; Shalon, R.; Davis, R.; Brown, P. Quantitative Monitoring of Gene Expression Patterns with a Compolementatry DNA Microarray. *Science* **1995**, *270*, 467–470. [CrossRef]
11. Tefferi, A.; Bolander, E.; Ansell, M.; Wieben, D.; Spelsberg, C. Primer on Medical Genomics Part III: Microarray Experiments and Data Analysis. *Mayo Clin. Proc.* **2002**, *77*, 927–940. [CrossRef]
12. Jiang, D.; Tang, C.; Zhang, A. Cluster Analysis for Gene Expression Data: A Survey. *IEEE Trans. Knowl. Data Eng.* **2004**, *18*, 1370–1386. [CrossRef]
13. Amato, R.; Ciaramella, A.; Deniskina, N.; del Mondo, C.; di Bernardo, D.; Donalek, C.; Longo, G.; Mangano, G.; Miele, G.; Raiconi, G.; et al. A Multi-Step Approach to Time Series Analysis and Gene Expression Clusterings. *Bioinformatics* **2006**, *22*, 589–596. [CrossRef]
14. Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R. Missing Value Estimation Methods for Dna Microarrays. *Bioinformatics* **2019**, in press. [CrossRef]
15. Hill, A.; Brown, E.; Whitley, M.; Tucker-Kellogg, G.; Hunter, C.; Slonim, A. Evaluation of Normalization Procedures for Oligonucleotide Array Data Based on Spiked cRNA Contros. *Genome Biol.* **2001**, *2*, research0055.1–research0055.13. [CrossRef] [PubMed]
16. Schuchhardt, J.; Beule, D.; Malik, A.; Wolski, E.; Eickhoff, H.; Lehrach, H.; Herzel, H. Normalization Strategies for cDNA Microarrays. *Nucleic Acids Res.* **2000**, *28*, e47. [CrossRef] [PubMed]
17. Ciaramella, A.; Gianfico, M.; Giunta, G. Compressive sampling and adaptive dictionary learning for the packet loss recovery in audio multimedia streaming. *Multimed. Tools Appl.* **2016**, *75*, 17375–17392. [CrossRef]
18. Ciaramella, A.; Giunta, G. CPacket loss recovery in audio multimedia streaming by using compressive sensing. *IET Commun.* **2016**, *10*, 387–392. [CrossRef]
19. Li, X.; Wong, K.-C. Evolutionary Multiobjective Clustering and Its Applications to Patient Stratification. *IEEE Trans. Cybern.* **2019**, *45*, 1680–1693. [CrossRef]
20. Golub, T.R.; Slonim, D.K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.P.; Coller, H.; Loh, M.L.; Downing, J.R.; Caligiuri, M.A.; et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **1999**, *286*, 531–537. [CrossRef]
21. Haslinger, C.; Schweifer, N.; Stilgenbauer, S.; Döhner, H.; Lichter, P.; Kraut, N.; Stratowa, C.; Abseher, R. Microarray gene expression profiling of B-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and VH mutation status. *J. Clin. Oncol.* **2004**, *22*, 3937–3949. [CrossRef]
22. Nutt, C.L.; Mani, D.R.; Betensky, R.A.; Tamayo, P.; Cairncross, J.G.; Ladd, C.; Pohl, U.; Hartmann, C.; McLaughlin, M.E.; Batchelor, T.T.; et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res.* **2003**, *63*, 1602–1607.
23. Bhattacharjee, A.; Richards, W.G.; Staunton, J.; Li, C.; Monti, S.; Vasa, P.; Ladd, C.; Beheshti, J.; Bueno, R.; Gillette, M.; et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 13790–13795. [CrossRef]
24. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef] [PubMed]
25. Alizadeh, A.A.; Eisen, M.B.; Davis, R.E.; Ma, C.; Lossos, I.S.; Rosenwald, A.; Boldrick, J.C.; Sabet, H.; Tran, T.; Yu, X.; et al. Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling. *Nature* **2000**, *403*, 503–511. [CrossRef] [PubMed]

26. Su, A.I.; Welsh, J.B.; Sapinoso, L.M.; Kern, S.G.; Dimitrov, P.; Lapp, H.; Schultz, P.G.; Powell, S.M.; Moskaluk, C.A.; Frierson, H.F., Jr.; et al. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res.* **2001**, *61*, 7388–7393. [PubMed]
27. Liew, A.W.; Yan, H.; Yang, M. Pattern Recognition Techniques for the Emerging Field of Bioinformatics: A review. *Pattern Recognit.* **2005**, *38*, 2055–2073. [CrossRef]
28. Bezdek, J.C.; Keller, J.; Krisnapuram, R.; Pal, N.R. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*; Kluwer Academic Publisher: Norwell, MA, USA, 1999.
29. McQueen, J.B. Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 7 January 1966.
30. Sherlock, G. Analysis of Large-Scale Gene Expression Data. *Curr. Opin. Immunol.* **2000**, *12*, 201–205. [CrossRef]
31. Smet, F.D.; Mathys, J.; Marchal, K.; Thijs, G.; Moor, M.; Bart, D.; Moreau, A. Adaptive Quality-Based Clustering of Gene Expression Profiles. *Bioinformatics* **2002**, *18*, 735–746. [CrossRef] [PubMed]
32. Heyer, L.J.; Kruglyak, S.; Yooseph, S. Exploring Expression Data: Identification and Analysis of Coexpressed Genes. *Genome Res.* **1999**, *9*, 1106–1115. [CrossRef] [PubMed]
33. Ralf-Herwig, P.A.; Muller, C.; Bull, C.; Lehrach, H.; Brien, J.O. Large-Scale Clustering of cDNA-Fingerprinting Data. *Genome Res.* **1999**, *9*, 1093–1105. [CrossRef]
34. Dubes, R.; Jain, A. *Algorithms for Clustering Data*; Prentice Hall: Upper Saddle River, NJ, USA, 1988.
35. Duda, R.O.; Hart, P.E.; Stork, D.G. *Pattern Classification*, 2nd ed.; John Wiley & Sons Inc.: Hoboken, NJ, USA, 2001.
36. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley and Sons: Hoboken, NJ, USA, 1990.
37. Eisen, M.B.; Spellman, P.T.; Brown, P.O.; Botstein, D. Cluster Analysis and Display of Genome-Wide Expression Patterns. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 14863–14868. [CrossRef]
38. Iyer, V.R.; Eisen, M.B.; Ross, D.T.; Schuler, G.; Moore, T.; Lee, J.C.F.; Trent, J.M.; Staudt, L.M.; Hudson, J., Jr.; Boguski, M.S.; et al. The Transcriptional Program in the Response of Human Fibroblasts to Serum. *Science* **1999**, *283*, 83–87. [CrossRef] [PubMed]
39. Perou, C.M.; Jeffrey, S.S.; Rijn, M.V.D.; Rees, C.A.; Eisen, M.B.; Ross, D.T.; Pergamenschikov, A.; Williams, C.F.; Zhu, S.X.; Lee, J.C.F.; et al. Distinctive Gene Expression Patterns in Human Mammary Epithelial Cells and Breast Cancers. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 9212–9217. [CrossRef] [PubMed]
40. Liang, F.; Wang, N. Dynamic agglomerative clustering of gene expression proles. *Pattern Recognit. Lett.* **2007**, *28*, 1062–1076. [CrossRef]
41. Tamayo, P.; Solni, D.; Mesirov, J.; Zhu, Q.; Kitareewan, S.; Dmitrovsky, E.; Lander, E.S.; Golub, T.R. Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 2907–2912. [CrossRef] [PubMed]
42. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data Clustering: A Review. *ACM Comput. Surv.* **1999**, *31*, 254–323. [CrossRef]
43. Fraley, C.; Raftery, A.E. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *Comput. J.* **1998**, *41*, 578–588. [CrossRef]
44. McLachlan, G.J.; Bean, R.W.; Peel, D. A Mixture Model-Based Approach to the Clustering of Microarray Expression Data. *Bioinformatics* **2002**, *18*, 413–422. [CrossRef] [PubMed]
45. McLachlan, G.J.; Peel, D. *Finite Mixture Models*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2000.
46. Yeung, K.Y.; Fraley, C.; Murua, A.; Raftery, A.E.; Ruzz, A.L. Model-Based Clustering and Data Transformations for Gene Expression Data. *Bioinformatics* **2001**, *17*, 977–987. [CrossRef]
47. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum-Likelihood from Incomplete Data Via the EM Algorithm. *J. R. Stat. Soc.* **1977**, *39*, 1–38. [CrossRef]
48. Kohonen, T. *Self Organizing Maps*; Springer: Berlin/Heidelberg, Germany, 1995.
49. Shamir, R.; Sharan, R. Click: A Clustering Algorithm for Gene Expression Analysis. In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, La Jolla/San Diego, CA, USA, 19–23 August 2000.
50. Ben-Dor, A.; Shamir, R.; Yakhini, Z. Clustering Gene Expression Patterns. *J. Comput. Biol.* **1999**, *6*, 281–297. [CrossRef]

51. Jiang, D.; Pei, J.; Zhang, A. DHC: A Density-Based Hierarchical Clustering Method for Time-Series Gene Expression Data. In Proceedings of the Third IEEE Symposium on Bioinformatics and Bioengineering, Bethesda, MD, USA, 12 March 2003.
52. Ciaramella, A.; Staiano, A.; Tagliaferri, R.; Longo, G. NEC: A Hierarchical Agglomerative Clustering based on Fischer and Negentropy Information. In *Neural Nets*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 49–56
53. Napolitano, F.; Raiconi, G.; Tagliaferri, R.; Ciaramella, A.; Staiano, A.; Miele, G. Clustering and visualization approaches for human cell cycle gene expression data analysis. *Int. J. Approx. Reason.* **2008**, *47*, 70–84. [CrossRef]
54. Ciaramella, A.; Cocozza, S.; Iorio, F.; Miele, G.; Napolitano, F.; Pinelli, M.; Raiconi, G.; Tagliaferri, R. Interactive data analysis and clustering of genomic data. *Neural Netw.* **2008**, *21*, 368–378. [CrossRef] [PubMed]
55. Camastra, F.; Ciaramella, A.; Son, L.H.; Riccio, A.; Staiano, A. *Fuzzy Similarity-Based Hierarchical Clustering for Atmospheric Pollutants Prediction, Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2019; Volume 11291.
56. Mitra, S.; Das, R.; Banka, H.; Mukhopadhyay, S. Gene Interaction—An evolutionary biclustering approach. *Inf. Fusion* **2009**, *10*, 242–249. [CrossRef]
57. Pontes, B.; Giráldez, R.; Aguilar-Ruiz, J.S. Biclustering on expression data: A review. *J. Biomed. Informat.* **2015**, *57*, 163–180. [CrossRef] [PubMed]
58. Staiano, A.; Tagliaferri, R. Visualization of High Dimensional Scientific Data, Book of Tutorials. In Proceedings of the International Joint Conference on Neural Networks, Montreal, QC, Canada, 31 July–4 August 2005.
59. Bishop, C.M. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, UK, 1995.
60. Tipping, M.E.; Bishop, C.M. Probabilistic principal component analysis. *J. R. Stat. Soc.* **1999**, *21*, 611–622. [CrossRef]
61. Tipping, M.E.; Bishop, C.M. Mixtures of probabilistic principal component analyzers. *Neural Comput.* **1999**, *11*, 443–482. [CrossRef] [PubMed]
62. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning—Data Mining, Inference, and Prediction*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2009.
63. Vesanto, J. SOM-Based Data Visualization Methods. *Intell. Data Anal. J.* **1999**, *3*, 111–126. [CrossRef]
64. Kaski, S. Data Exploration Using Self Organizing Maps. Ph.D. Thesis, Helsinki Institute of Technology, Espoo, Finland, 1997.
65. Bishop, C.M.; Svensen, M.; Williams, C.K.I. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998**, *10*, 215–234. [CrossRef]
66. Bishop, C.M.; Tipping, M.E. A hierarchical latent variable model for data visualization. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 281–293. [CrossRef]
67. Tino, P.; Nabney, I. Hierarchical GTM: Constructing localized nonlinear projection manifolds in a principled way. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 639–656. [CrossRef]
68. Bishop, C.M. Latent variable models. In *Learning in Graphical Models*; Jordan, M.I., Ed.; MIT Press: Cambridge, MA, USA, 1999; pp. 371–403.
69. Chang, K. Nonlinear Dimensionality Reduction Using Probabilistic Principal Surfaces. Ph.D. Thesis, The University of Texas at Austin, Austin, TX, USA, 2000.
70. Whitfield, M.L.; Sherlock, G.; Saldanha, A.J.; Murray, J.I.; Ball, C.A.; Alexander, K.E.; Matese, J.C.; Perou, C.M.; Hurt, M.M.; Brown, P.O.; et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **2002**, *13*, 1977–2000. [CrossRef] [PubMed]
71. Spellman, P.T.; Sherlock, G.; Zhang, M.Q.; Iyer, V.R.; Anders, K.; Eisen, B.; Brown, P.O.; Botstein, D.; Futcher, B. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* **1998**, *9*, 3273–3297. [CrossRef] [PubMed]
72. Domingos, P. *The Master Algorithms. How the Quest for the Ultimate Learning Machine Will Remake Our World*; Basic Books; Hachette Book Group: New York, NY, USA, 2015.
73. Camastra, F.; Staiano, A. Intrinsic dimension estimation: Advances and open problems. *Inf. Sci.* **2016**, *328*, 26–41. [CrossRef]

74. Satija, R.; Farrell, J.A.; Gennert, D.; Schier, A.F.; Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **2015**, *33*, 495–502. [CrossRef] [PubMed]
75. Wolf, F.A.; Theis, P.A.F.J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **2018**, *19*, 15. [CrossRef] [PubMed]