June 2012

Facultad de Informática
U.P.M

Author: ***Antonio Gracia Berná***
Supervisors: ***Víctor Robles Forcada***
Santiago González Tortosa

[MEDVIR: 3D VISUAL INTERFACE APPLIED TO GENE PROFILE ANALYSIS]

MASTER'S THESIS

MASTER IN ADVANCED COMPUTING FOR SCIENCE AND ENGINEERING (CACI)

MedVir framework is an intuitive and new mechanism based on the visualization of multidimensional and medical data. The data are visualized in a tridimensional environment by means of a powerful dimensionality reduction process. After this, the expert can interact with the visualization. It has been devised in order to make easier to the expert the possibility of drawing conclusions in a fast way.

To my parents Antonio and Silvia,

thank you for giving my life and

patiently raising me.

To Cristina, thank you for supporting

me.

I love you

Summary

The origins for this work arise in response to the increasing need for biologists and doctors to obtain tools for visual analysis of data. When dealing with multidimensional data, such as medical data, the traditional data mining techniques can be a tedious and complex task, even to some medical experts. Therefore, it is necessary to develop useful visualization techniques that can complement the expert's criterion, and at the same time visually stimulate and make easier the process of obtaining knowledge from a dataset. Thus, the process of interpretation and understanding of the data can be greatly enriched.

Multidimensionality is inherent to any medical data, requiring a time-consuming effort to get a clinical useful outcome. Unfortunately, both clinicians and biologists are not trained in managing more than four dimensions. Specifically, we were aimed to design a 3D visual interface for gene profile analysis easy in order to be used both by medical and biologist experts.

In this way, a new analysis method is proposed: MedVir. This is a simple and intuitive analysis mechanism based on the visualization of any multidimensional medical data in a three dimensional space that allows interaction with experts in order to collaborate and enrich this representation. In other words, MedVir makes a powerful reduction in data dimensionality in order to represent the original information into a three dimensional environment. The experts can interact with the data and draw conclusions in a visual and quickly way.

Contents

1 - Introduction

Diseases such as cancer have become a great social problem to which we must seek a solution as soon as possible. In the struggle to find the solution, biologists, doctors and computer experts are working together in interdisciplinary projects, creating a new field called Computational Biology (CB) [1]. One of the many types of CB research [2] is the gene profile discovery in diseases using data analysis techniques. The work, here presented, is focused on this field.

Besides that, visualization techniques are also being used by the scientific community to understand and obtain different conclusions about a particular dataset in an easy way. Nevertheless, these techniques are not frequently used to analyze very huge data volumes in life sciences field, particularly in genomics, due to the high complexity of the data ('*curse of dimensionality*' [3]). There exist approaches that highlight the most important features of the data, and they make possible the construction of virtual reality spaces to visually understand the intrinsic nature of the data [4]. The benefits of representing n-dimensional data in tridimensional spaces are very well-known [3]. Nowadays, these kinds of representations are carried out by means of dimensionality reduction and transformation of the data, and making use of a strong component of interaction methods.

Initially, the development of a tool for making easier and faster the visualization of multidimensional medical datasets in a tridimensional environment was required. This mechanism should allow the acquisition of valid and implicitly underlying knowledge in data, for example different patterns, relationships between attributes or instances, outliers or trends in data.

Thus, a new analysis approach is presented: MedVir. This is a simple and intuitive analysis mechanism based on the visualization of any medical data (in the present case DNA Microarray data, e.g. gene profiles, patients, clinical data, etc.) in a three dimensional space that allows interactions with experts in order to collaborate and enrich this representation. In other words, MedVir makes a powerful reduction in data dimensionality, through an Evolutionary Optimization technique, in order to represent all the information in three dimensions. After this, the expert could be able to understand multidimensional biological data in an easier way and interact with them in a way as never seen before. Therefore, the multidimensional intrinsic nature of data will be presented in a tridimensional environment, while a minimum loss of information during the dimensionality reduction process is achieved. This will allow that the possible patterns, trends, outliers or relationships that originally exist in the data will be preserved in the tridimensional space in order to be analyzed by the biologist expert.

This study directly focuses on DNA Microarray data. A lot of research into this topic has tried to find the gene profile (based on D gene expressions) that helps to diagnose the disease. After finding the gene profile, the researchers do not usually analyze the relationship between these gene expressions at the different stages of a disease, or in different patients. However, it is crucial to understand the disease and its possible manifestations in different patients in order to generate personalized treatments. Of course, depending on the number of gene expressions of the profile, patients and other data, this analysis could require more effort and considerable time on the part of the experts. For this reason, an analysis mechanism that reduces both time and effort is needed.

At the same time, the use of data mining techniques for the gene profile discovery of diseases is becoming usual. These techniques do not usually analyze the relationships between genes in depth, depending on the different variety of manifestations of the disease (related to patients). This kind of analysis takes a considerable amount of time and is not always the focus of the research. However, it is crucial in order to generate personalized treatments to fight the disease.

MedVir can be seen as a mechanism for validating a particular gene profile analysis, in order to be used by the medical and biologist experts. As said above, MedVir could be considered as a dimensionality reduction approach. In this way, MedVir is compared with different dimensionality reduction state-of-the-art algorithms.

1.1 - Document structure

The work is organized as follows: the following section presents literature of Dimensionality Reduction algorithms. Here, a *formal definition* of the problem, the different *taxonomies* in Dimensionality Reduction, *methods* and different *measures* to assess the quality in sense of geometry preservation, are presented. In **Section 3**, the MedVir framework is described in detail. Next, in **Section 4**, MedVir is assessed when applying it to DNA Microarray data. In order to evaluate it properly, a comparison of the results between MedVir and different state-of-the-art algorithms (in terms of geometry preservation capability) is carried out. **Section 5** attempts to perform a data analysis when representing DNA Microarray data using MedVir. So, several conclusions about a particular dataset will be extracted. **Section 6** formulates final conclusions and futures lines of research in MedVir, and **Section 7** presents the bibliography.

2 - State of the Art

By analyzing the Dimensionality Reduction (DR) word in the literature, it shows two general approaches for carrying out a DR [5]:

Feature Extraction: Transforming the existing features into a different set of attributes.
Feature Selection: Selecting a subset of the existing features without a transformation [6-8].

Since the MedVir algorithm is based on the first one, all of this work focuses on Feature Extraction (FE). There are currently two canonical ways of dealing with the data when carrying out a DR process. The first one does so in a linear (Linear Dimensionality Reduction or LDR), whiles the second one is in a nonlinear way (Non Linear Dimensionality Reduction or NLDR).

LDR handles datasets containing linear dependencies. However, they are not powerful enough to deal with complex datasets. The behavior of many datasets, such as a DNA Microarray, could not be explained by means of LDR because maybe it contains essential multiple nonlinear relationships between attributes that cannot simply be interpreted by using linear models [3]. This suggests the design of other techniques (NLDR methods) in order to highlight the true underlying structure of the data. These methods assume that data are generated according to a nonlinear model.

The following subsections present the definition of a dimensionality reduction process, a taxonomy of DR-FE algorithms based on feature extraction approaches, some technical details about the most known DR-FE algorithms and assessment measures to compare the algorithms in sense of geometry preservation.

2.1 - Definition of the problem

The problem of (nonlinear) dimensionality reduction can be defined as follows. Assume we have a dataset represented in a *n x D* matrix **X** consisting of *n* data vectors x_i (i ϵ {1, 2,…, n}) with dimensionality *D*. Assume further that this dataset has intrinsic dimensionality *d* (where *d* < *D*, and often *d* << *D*). Here, in mathematical terms, intrinsic dimensionality means that the points in dataset **X** are lying on or near a *manifold* [9] with dimensionality d that is embedded in the *D*-dimensional space.

Dimensionality reduction techniques transform dataset **X** with dimensionality *D* into a new dataset **Y** with dimensionality *d*, while retaining the geometry of the data as much as possible. In general, neither the geometry of the data manifold, nor the intrinsic dimensionality *d* of the dataset **X** are known. Therefore, dimensionality reduction is an ill-posed problem that can only be solved by assuming certain properties of the data (such as its intrinsic dimensionality).

2.2 - Taxonomy in DR-FE

Different taxonomies or classification models in DR-FE techniques have been proposed. Laurens van der Maaten et al. [10] carried out a thorough comparative review (figure 1) of the most important linear DR technique (PCA), and twelve front-ranked nonlinear DR techniques. They divided the DR techniques into two criteria.

First of all, they defined the convex and non-convex intrinsic nature of the techniques. Convex techniques optimize an objective function that does not contain any local optima (i.e., the solution space is convex [11]), whereas non-convex techniques optimize objective functions that do contain local optima. The second division criterion is related to full or sparse spectral techniques. The first one carries out an eigen-decomposition of a full matrix that captures the covariance between dimensions or the pairwise similarities between datapoints. The other case solves a

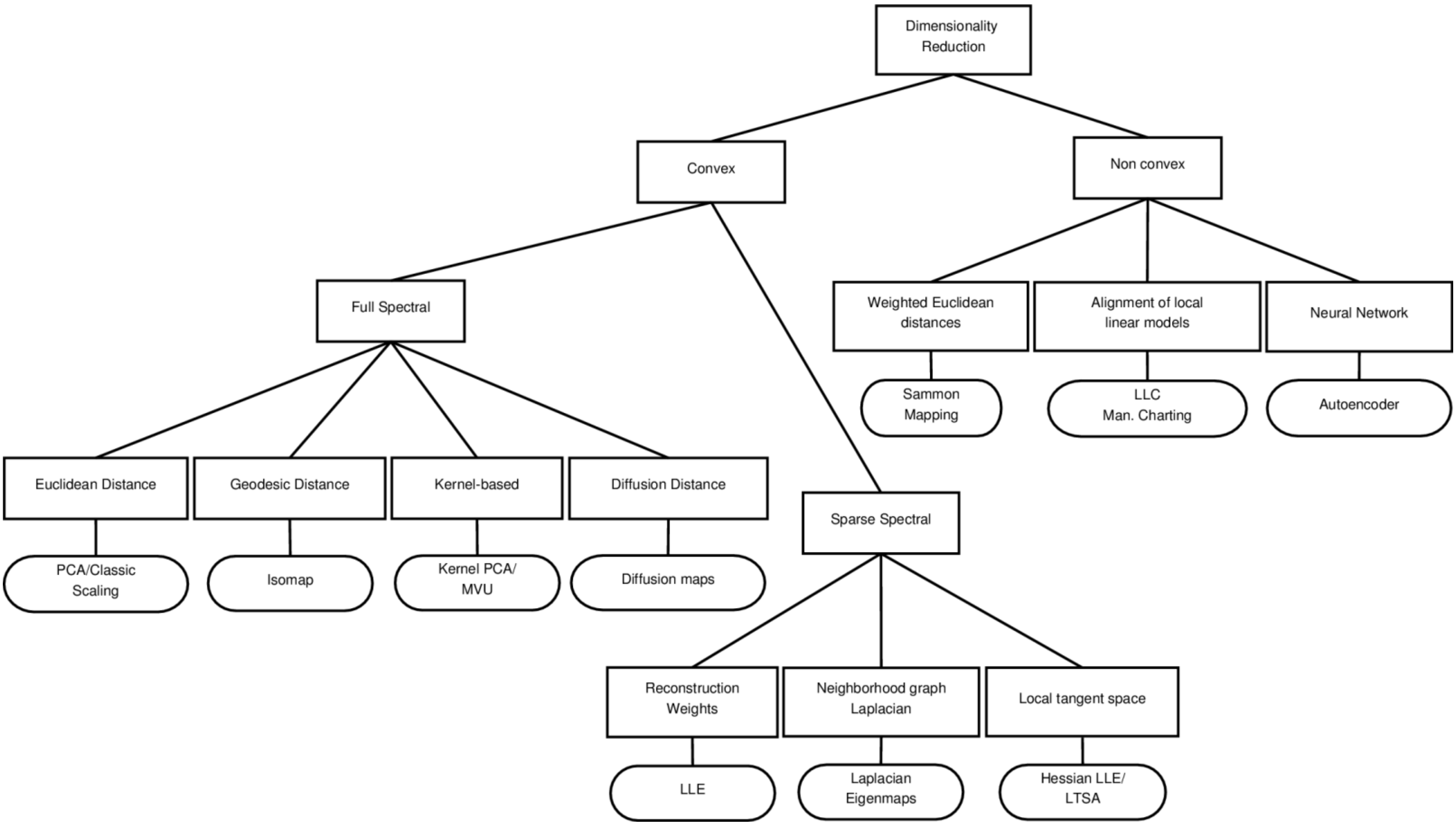


Figure 1. Laurens van der Maaten's Taxonomy.

sparse (generalized) eigenproblem.

On the other hand, John A. Lee et al. proposed a different taxonomy of DR techniques [3] in accordance with procedures that reduce the features or dimensionality of the data by preserving the global shape of the geometry, or by preserving the local properties and neighborhood information of the data [12]. That is distance and topology preservation, respectively. The first one, the principle of distance preservation (DP), is very intuitively both simple to understand and easy to compute. The other one, topology preservation (TP), appears to be more powerful and elegant than DP but is also more complex to implement. DP and TP are described in detail in the following subsections.

2.2.1 - Distance preservation

From the point of view of an ideal case, the preservation of the pairwise distances measured in a dataset ensures that the low-dimensional embedding inherits the main geometric properties of data, like the global shape. However, in nonlinear cases distances cannot be perfectly preserved. To explain this, it is necessary to make use of the concept of mathematical *manifold*. As said in [9], a topological manifold M is a topological space that is locally Euclidean, meaning that around every point of M there is a neighborhood that is topologically the same as the open unit ball in $\Re^d$.

The underlying idea of DP is based on the fact that, theoretically, any manifold can be described by pairwise distances. That is, if a complete mapping between the pairwise distances in the high dimensional space and low dimensional space is carried out, the DR process will take place successfully. Intuitively, if far points are kept far, and if close points remain close, the manifold will be very similar in both dimensional spaces.

DP methods can be divided, as considered by Lee et al. [3], into three groups:

- Spatial distances as Euclidean (L = 2) or Manhattan (L = 1), are well known because of the intuitive and natural way everybody measure distances in a Euclidean space. Algorithms with this kind of distance are MDS [13], Sammon non-linear mapping [14] and Curvilinear component analysis (CCA) [15].

- Geodesic distances - graph distances. There exists an axiomatic basis for non-linear DR based on the fact that it is possible to get a deep insight into the data by means of an *unfolding* data process. Taking this into consideration, the use of geodesic distances is obvious and justified. The geodesic distance between two points is defined as the distance along the mathematical manifold where the data points are embedded. The geodesic distance can be partially approximated by constructing a neighborhood graph, and considering the distances between the points as paths in the graph (figure 2). This kind of distance was originally conceived to deal with some shortcomings in the spatial metrics (figure 3). Examples of algorithms using this distance are Isomap [12, 16], Geodesic nonlinear mapping (GNLM) [17-19] and Curvilinear distance analysis (CDA) [19, 20].

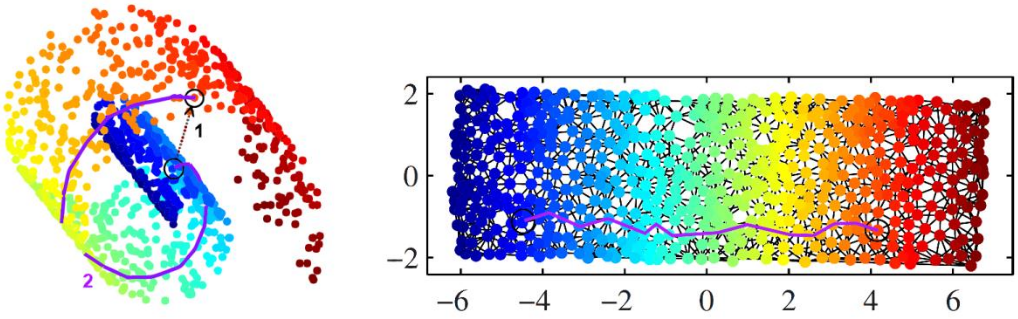


Figure 2. Geodesic distance between two points. This dataset consists of a list of 3-dimensional points. It is, a two-dimensional manifold embedded into a three-dimensional space.

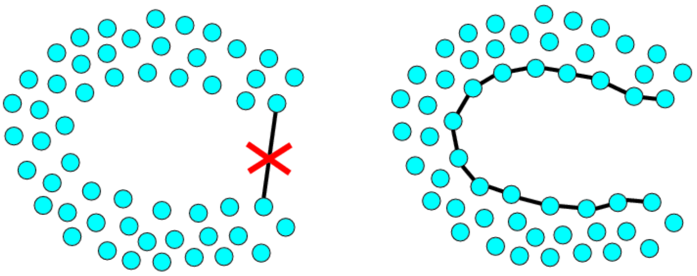


Figure 3. Appearance of short circuit phenomenon. Left-hand figure: when performing an unfolding process, the appearance of short circuit induced by the Euclidean distance is likely. Right-hand figure: the benefits of the geodesic distance. The two points are not neighbors, because they are far away in accordance with the geodesic distance.

- Other distances. There are also Non Linear DR (NLDR) methods that rely on less geometrically intuitive ideas. These techniques are characterized by the use of other distances. For instance, Kernel PCA (KPCA) [21, 22], which is closely related to the spectral methods. In these cases, the methods directly stem from mathematical considerations as regards the kernel functions.

2.2.2 - Topology preservation

Techniques that reduce the dimensionality of the data by preserving the topology of the data rather than their pairwise distances are also called local preservation approaches. These techniques help to overcome the drawback of using the principle of DP: the manifold could be constrained with distance conditions and, in many situations, the embedding of a manifold requires some flexibility because some subregions must be locally stretched or shrunk to embed them into other dimensional spaces.

What defines a manifold is its local topological information, i.e., the neighborhood relationships between the subregions of the manifold. In most cases, a manifold can be entirely characterized by giving relative or comparative proximities: a first region is close to a second one but far from a third one.

Most of these techniques work with a discrete mapping model, and the topology is also defined in a discrete way. Such discrete representation of the topology is called a lattice [23], i.e, a set of points regularly and homogeneously spaced on a graph.

Topology preservation (TP) techniques can be divided into two classes according to the kind of topology they use. The first one deals with methods relying on a *predefined lattice*, i.e, the lattice is fixed in advance and cannot change after the DR process has begun. Self-Organizing Maps (SOM's) [24] and Generative Topographic Mapping (GTM) [25] are well-known as predefined lattice methods. The second group contains methods working with a *data-driven lattice*. This concept means that the shape of the lattice can be modified or is entirely built while the methods are running. Locally linear embedding (LLE), Laplacian eigenmaps (LE) and Isotop [26] lie on this category.

2.3 - Methods

Once the two different taxonomies have been presented, this section describes six of the most currently used DR algorithms in order to be fully compared in terms of geometry preservation, within the framework presented here. The most popular LDR algorithm (PCA) and five NLDR algorithms have been selected for carrying out this study. Besides that, the original Star Coordinates algorithm (used in the implementation of MedVir) is presented.

2.3.1 - Principal Components Analysis

Principal Components Analysis (PCA) [27, 28] carries out a DR process by embedding the data into a lower dimensionality linear subspace.

This technique attempts to build a low dimensional representation of the data that describes as much of the variance in the data as possible. That is, it finds a linear basis of reduced dimensionality for the data.

After this process, the amount of variance in the data is maximal. Mathematically, PCA builds a new coordinate system by selecting those d axes $w_1,\ldots,w_d$ $\epsilon\ \Re^D$, which best explain the variance in the data:

$$w_l = \arg max_{\|w\|=1}\ var(X_w) = argmax_{\|w\|=1}\, w^T Cw \qquad 1)$$

PCA searches a linear mapping w that maximizes the cost function trace $(w^T Cw)$, where C is the sample covariance matrix of the data X. It can be shown that this linear mapping is made up of the d principal eigenvectors of the sample covariance matrix of the zero-mean data. $w_1,\ldots,w_d$ are chosen in the same way, but orthogonal to each other ($C\ \epsilon\ R^{DxD}$ denotes the covariance matrix of the data X). Thus, the principal components $p_i = Xw_i$ explain most of the variance in the data. Before mapping the data, the samples in X were centered by subtracting their mean. Since PCA only considers the variance between the samples it works best if those features, that are relevant for class labeling, account for a large part of the variance. The covariance matrix grows rapidly for high-dimensional input data. In order to overcome this situation, the covariance matrix is substituted by the matrix of squared Euclidean distances.

$$D_E = \frac{1}{N} XX^T (D_E \,\epsilon\, R^{NxN}) \qquad 2)$$

2.3.2 - Kernel PCA

Kernel PCA (KPCA) [21, 22] is a non-linear extension of PCA using a technique called the kernel method. It is equivalent to mapping the data onto a very high dimensional space (up to infinite), namely, Reproducing the Kernel Hilbert Space (RKHS), and applying the same optimization technique as PCA in the RKHS.

The changes brought about by Isomap to metric MDS were motivated by geometrical consideration, but KPCA extends the algebraical features of MDS to non-linear manifolds, without regard to their geometrical meaning. Because of the non-linear mapping process, the distance preservation is not an objective of KPCA, although PCA offers distance preservation in the RKHS.

2.3.3 - Locally Linear Embedding

Locally Linear Embedding (LLE) [29, 30] tries to preserve the local properties of the data from a different point of view. In LLE, the local properties of the data manifold (represented as X_i) are constructed by mapping the datapoints as a linear combination of their k nearest neighbors. In the low-dimensional representation of the data, LLE attempts to retain the reconstruction weights in the linear combinations as best as possible.

$$\overrightarrow{X_\iota} = \sum_{j=1}^{k} W_{ij}\,\overrightarrow{X_J} \qquad 3)$$

Weights W_{ij} are computed by minimizing the constrained least-squares problem. The embedding vectors $\overrightarrow{Y_\iota}$ are reconstructed by W_{ij}, by minimizing Eq. 5.

$$E(W) = \sum_{i=1}^{N} \left|\overrightarrow{X_\iota} - \sum_{j=1}^{k} W_{ij}\overrightarrow{X_J}\right|^2 \qquad 4)$$

$$\Phi(Y) = \sum_{i=1}^{N} \left|\overrightarrow{Y_\iota} - \sum_{j=1}^{k} W_{ij}\overrightarrow{Y_J}\right|^2 \qquad 5)$$

Although W_{ij} and $\overrightarrow{Y_\iota}$ are computed by methods in linear algebra, the constraint that points are only reconstructed from neighbors can result in highly nonlinear embeddings.

2.3.4 - Laplacian Eigenmaps

The Laplacian Eigenmaps algorithm (LE) is similar to LLE in the sense that it finds a low-dimensional data representation by preserving the local properties of the manifold [31, 32]. LE can be included in sparse spectral techniques.

This algorithm attempts to compute a low-dimensional representation of the data in which the distances between a datapoint and its k nearest neighbors are minimized. The distance in the low-dimensional data representation between a datapoint and its first nearest neighbor contributes more to the cost function than the distance between the datapoint and its second nearest neighbor. Using spectral graph theory, the minimization of the cost function is defined (eq. 6) as an eigenproblem, where Wij values are from the Gaussian kernel function,

$$E(Y) = \sum_{ij} \left|\overrightarrow{Y_\iota} - \overrightarrow{Y_J}\right|^2 W_{ij} \qquad 6)$$

and for neighboring y_i, y_j (W(i,j) = 0 otherwise), the distances between the low space representations are minimized and the nearby samples x_i,x_j are highly weighted, and thus brought closed together.

2.3.5 - Difussion Maps

The Diffusion Maps (DM) algorithm [33, 34] is based on diffusion processes for finding meaningful geometric descriptions of data sets. In this technique, a graph is built from the samples on the manifold where the diffusion distance describes the connectivity on the graph between every two points. This distance is characterized by the probability of transition between them. DM captures the intrinsic natural parameters that generate the data, which usually lie on a lower dimension.

The assumption is that the data lie on a non-linear manifold. The data are transformed using a Gaussian kernel function (7). This kernel is used for the construction of Markov Random Walk (MRW) matrix. The diffusion distances in the original space are mapped into Euclidean distances in the new diffusion space. Because of the diffusion distance between two points is obtained from all of the possible paths in the graph, DM is robust to noise.

$$W(i,j) = \exp(-\frac{\|x_i - x_j\|^2}{\sigma^2}) \qquad 7)$$

2.3.6 - Isomap

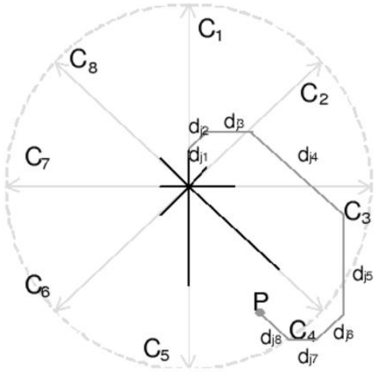
Isomap [12, 16] is one of the simplest non-linear DR methods that use the graph distance (based on geodesic distance). Isomap uses graph distances instead of Euclidean ones in the algebraical procedure of metric MDS. It is important to remember that the non-linear capabilities of Isomap are exclusively contributed by the graph distance.

In Isomap, the geodesic distances between the datapoints x_i (i = 1, 2,…,n) are calculated by constructing a neighborhood graph *G*. Every datapoint x_i is connected to its *k* nearest neighbors x_{ij} in the dataset *X*.

The shortest path between two points in the graph can easily be computed using Dijkstra's or Floyd's algorithm [35]. The geodesic distances between all datapoints in *X* are computed, making up a pairwise geodesic distance matrix. The low-dimensional representations y_i of the datapoints x_i in the low-dimensional space *Y* are computed by applying MDS in the resulting pairwise geodesic distance matrix. A significant weakness of the Isomap algorithm is its topological instability [36].

2.3.7 - Star Coordinates algorithm

The original Star Coordinates algorithm [43] constructs a low dimensional space composed of a linear combination of the attributes. It works as follows: first, it considers the attributes of the dataset as coordinate axes. Then it arranges the coordinate axes onto a flat (two-dimensional) surface forming equidistant angles between axes. The mapping of an n-dimensional point to a two-dimensional Cartesian coordinate is computed by means of the sum of all unit vectors of every coordinate, multiplied by the data value of that coordinate. Figure 4 illustrates an example of the final position of a data point in an 8-dimensional dataset. In this framework a 3D extension is used, so it can be described as in the following equation:



$$P_j(x,y,z) = \begin{cases} o_x + \sum_{i=1}^{n} u_{xi}(d_{ji} - min_i) \\ o_y + \sum_{i=1}^{n} u_{yi}(d_{ji} - min_i) \\ o_z + \sum_{i=1}^{n} u_{zi}(d_{ji} - min_i) \end{cases} \qquad 8)$$

Figure 4. Process of obtaining the final position of a data point for an 8-dimensional dataset (two-dimensional space).

Where d_{ji} is the *j*-th data with the *i*-th value, mini is the minimum value of the scaled values in every coordinate, u_{xi} and u_{yi} are unit vectors in the direction of every coordinate, and o_x,o_y,o_z is the origin of the coordinate system.

2.4 - Quality assessment measures

There are many different quality assessment criteria for evaluating the performance of the DR algorithms. Current approaches focus on evaluating the local-neighborhood-preservation and the global-structure-holding performance of DR methods. By taking both properties into consideration, their intrinsic capability can be more faithfully reflected, and hence more a rational measure for the proper selection in real-life applications can be offered. Three local-neighborhood-preservation criteria and one global-structure-holding criterion have been selected in order to evaluate the different DR algorithms.

2.4.1 - Local-neighborhood-preservation criteria

In this section the three local quality assessment criteria for NLDR were reviewed: LCMC [37], T&C [38, 39] and the MRRE [3, 40, 41]. By using the criteria below, the local-neighborhood-preservation performance of the DR methods can be effectively assessed.

If the original dataset is denoted by $X = \{x_i\}^l_{i=1}$, and the corresponding representational set computed by the DR method used is as $Y = \{y_i\}^l_{i=1}$. Then, the LCMC can be defined as:

$$Q_k = 1 - \frac{1}{lk} \sum_{i=1}^{l} (|\psi_k^x(i) \cap \psi_k^y(i)| - \frac{k^2}{l-1}) \qquad 9)$$

Where k is the pre-specified neighborhood size, $\psi_k^x(i)$ is the index set of x_i's k-NN and $\psi_k^y(i)$ is the index set of y_i's k-NN. The k-NN value represents the k nearest neighbors of a datum. If the overlap between two k-NN neighboring sets of the original and representational sets is computing, the LCMC gives a general measure for the local faithfulness of the calculated embeddings. The interval of $Q_L \in [0,1]$, whose values next to 1 mean a high neighborhood overlap between the two dimensional spaces, and next to 0 values the opposite of it.

The second local measure (T&C criterion) involves two evaluations, the trustworthiness and the continuity measure, defined, respectively, as:

$$M_T = 1 - \frac{2}{lk(2l-3k-1)} \sum_{i=1}^{l} \sum_{j \epsilon U_k(i) \notin V_k(i)} (r(i,j) - k) \qquad 10)$$

$$M_C = 1 - \frac{2}{lk(2l-3k-1)} \sum_{i=1}^{l} \sum_{j \epsilon V_k(i) \notin U_k(i)} (\hat{r}(i,j) - k) \qquad 11)$$

where k is the size of the neighborhood, r(i,j) and $\hat{r}$(i,j) are the rank of x_j and y_j in the ordering according to the distance from $x_i(y_i)$ in the original (representational) space. $U_k(i)$ and $V_k(i)$ are the set of those data samples that are in k-NN of $x_i(y_i)$ in the representational (original) space.

As regards the meaning of M_T and M_C, the first one measures the degree of trustworthiness that data points originally farther away enter the neighborhood of a sample in the embeddings. The latter evaluates the degree of continuity that data points that are originally in the neighborhood are pushed farther away in data representations. Therefore, the T&C measure is defined as:

$$Q_T = \alpha M_T + (1 - \alpha) M_C \qquad 12)$$

where α ϵ [0,1] is the compromise parameter. The trade-off between the two terms, tunable by a parameter α, governs the trade-off between trustworthiness and continuity. A properly selected α value, can reflect the consistency between the local neighborhoods of the original data and the corresponding ones in the embeddings calculated by the NLDR method. The interval of Q_T ϵ [0, 1] means that the higher values represent a good trustworthiness and continuity preservation.

The MRRE criterion is based on a very similar principle to that of the T&C, but it includes two elements defined as

$$W_T = 1 - \frac{1}{H_k} \sum_{i=1}^{N} \sum_{j \epsilon U_k(i)} \frac{|r(i,j) - \hat{r}(i,j)|}{r(i,j)} \qquad 13)$$

$$W_C = 1 - \frac{1}{H_k} \sum_{i=1}^{N} \sum_{j \epsilon V_k(i)} \frac{|r(i,j) - \hat{r}(i,j)|}{r(i,j)} \qquad 14)$$

where *k* is the size of the neighborhood and eq. 15 is the normalizing factor. The MRRE criterion is eq. 16 where β ϵ [0, 1] is the compromise parameter. The main difference between the MRRE and the T&C is that the first one considers all of the k-NN samples in the representational (original) space, and the latter focuses on the k-NN of the samples in the representational (original) space but not in the original (representational) space.

Although we are talking about subtle differences between them, they are important enough to be considered. The interval of Q_M ϵ [0, 1], whose values next to 0 will indicate a small rank error in the final embedding, is result of the error-based nature of MRRE.

$$H_k = Nl \sum_{i=1}^{k} (l - 2i + 1)/i \qquad 15)$$

$$Q_M = \beta W_T + (1 - \beta) W_C \qquad 16)$$

2.4.2 - Global-structure-holding criterion

The global geometry preservation (GGP) relies on the assumption that the global geometry of a mathematical n-dimensional manifold can be successfully mapped onto a lower dimensional manifold, by preserving the original distances between the high-dimensional points. Thus, theoretically the final embedding will retain, as much as possible, the original geometry of the n-dimensional manifold in which the data are embedded.

According to this concept, Sammon error [14] measure is used in order to compare the DR algorithms, in terms of global-structure preservation. Examples of error measures frequently used for structure preservation are [42] Stress (eq.17), Sammon (eq.18) and Quadratic error (eq.19).

$$S_{stress} = \sqrt{\frac{\sum_{i<j}(\delta_{ij}^2 - \zeta_{ij}^2)^2}{\sum_{i<j}\delta_{ij}^4}} \qquad 17)$$

$$Sammon = \frac{1}{\sum_{i<j}\delta_{ij}} \frac{\sum_{i<j}(\delta_{ij} - \zeta_{ij})^2}{\delta_{ij}} \qquad 18)$$

$$Quadratic = \sum_{i<j}(\delta_{ij} - \zeta_{ij})^2 \qquad 19)$$

where the distance between *i*th and *j*th objects in the original space are denoted by δ_{ij}, and the distance between their projections by ζ_{ij} . Sammon error must be minimized by carrying out a gradient descent, or by other means, usually involving iterative methods.

3 - Proposed algorithm

The MedVir framework is based on a previously developed framework [44], and its aim is to make easier the formulation of conclusions about multidimensional and biological data. Here, a new approach of acquiring knowledge in a visual and quick way is presented.

In other words, the objective of MedVir is to make a tool available for the expert in order to visualize and interact with biological data. To achieve this aim, a pipeline based on the following stages is considered (figure 5).

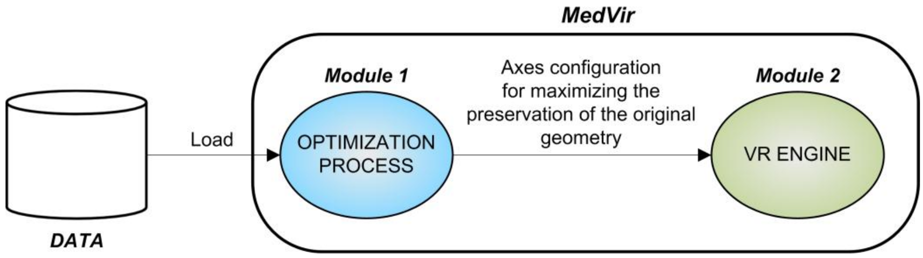

MedVir
Module 1
Axes configuration for maximizing the preservation of the original geometry
Module 2
Load
OPTIMIZATION PROCESS
VR ENGINE
DATA

Figure 5. Pipeline of the framework. First, data are loaded into Module 1. After the optimization process a set of axes is obtained. Next, data are visually represented by means of the Module 2(VR engine). Finally, the expert interprets the visualization.

The proposed framework is divided into two main modules. Module I carries out an optimization process (OP). Here, a search algorithm attempts to find a tridimensional embedding of the data which best preserves its intrinsic geometry. After the OP, the previous embedding in represented into a tridimensional environment. Thus, the Module II deals with the data visualization. In order to represent the geometry of the data in 3D, an extension of the original Star Coordinates algorithm (SC) is used [43]. This can be considered as a feature extraction process, as it transforms the previously selected features into three dimensions. Finally, the resulting 3D points are presented to the expert in a 3D and visually tangible way. In order to reinforce and stimulate the knowledge acquisition, the representation is complemented with different visual elements that make easier the interaction with the data.

At the end of the pipeline, the expert would be able to obtain an interpretation of the model in a very short time. However, a consideration must be highlighted: the final conclusions will be closely related to the knowledge and acquired experience of the expert in the data domain.

Next, the two main modules of the framework are explained in detail.

3.1 - Module 1. Optimization process

The OP is carried out by an evolutionary algorithm, particularly the Differential Evolution (DE) algorithm [45]. DE algorithm is characterized by performing a more exhaustive search than a generic evolutionary algorithm [46].

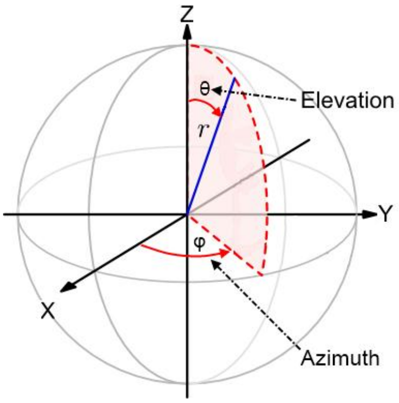
As other evolutionary algorithms, it scans the population and carries out mutation and crossover operations on the individuals for a number of generations. Nowadays, most of the DR algorithms are based on deterministic approaches, therefore it expresses the need to complement the existing techniques with a purely stochastic approach. The idea is to find one of the best possible solutions, just by means of a search through the tridimensional space of solutions.

3.1.1 - Codification of the individuals

The initial population of individuals is generated randomly, and they are represented by using the SC algorithm [43]. That is, the N attributes of the dataset are represented through *N* spatial or axis vectors. So, each individual in the population represents different axes configuration in the 3D Euclidean space. This way, the initial population is built with *K* individuals, each one representing an axes configuration. Each individual needs to be evolved in relation to a fitness function to achieve a population consisting of improved individuals. It is, *K* different axes configurations that generate the best data embeddings. Finally, the best one will be selected.

As regards the representation of the individuals, note that each axis corresponds to each one of the attributes in the input data, and it has three components normalized in spherical coordinates (θ, φ, r) where r is the radius, φ is the azimuthal angle (φ ϵ [0, 2π]) and θ corresponds to the elevation angle (θ ϵ [0, π]).

In fact, each individual is represented by Cartesian coordinates in order to evaluate the distances. In the following formula, the transformation between spherical and Cartesian coordinates is shown (See Eq. 20 and Fig. 6).



$$\begin{cases} x = rsin(\theta)\cos(\phi) \\ y = rsin(\theta)\sin(\phi) \\ z = rcos(\theta) \end{cases} \quad 20)$$

Figure 6. Relationship between spherical and cartesian coordinates.

3.1.2 - Scheme of the optimization process

The overall scheme of the OP implemented in MedVir can be seen in figure 7. First, the data are taken as an input of the optimization module. After this, the DE algorithm carries out an optimization task in order to find an axes configuration in the tridimensional space that best preserves the original geometry of the data.

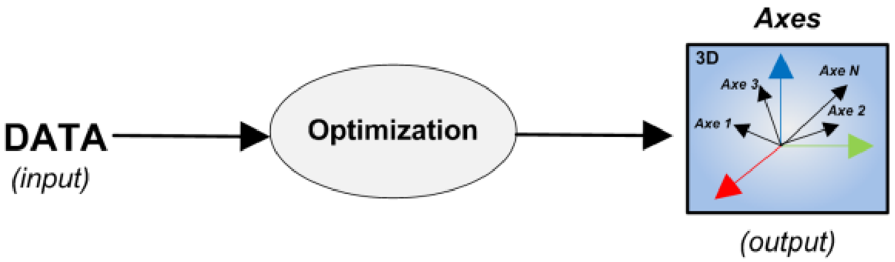


Figure 7. Optimization process - Module I. Data are used as the input of the OP. DE algorithm generates a set of axes as the output of the system.

In figure 8, the OP can be appreciated in greater detail. As a first step, this module has to calculate pair to pair distances of all the instances in the initial data. This is stored in a matrix which is considered as the target distance matrix, and it is squared. In other words, it represents the distance between the instance *i*-th in the rows and the instance *j*-th in the columns. The target distance matrix will contain absolutely all the information that it is expected to be conserved.

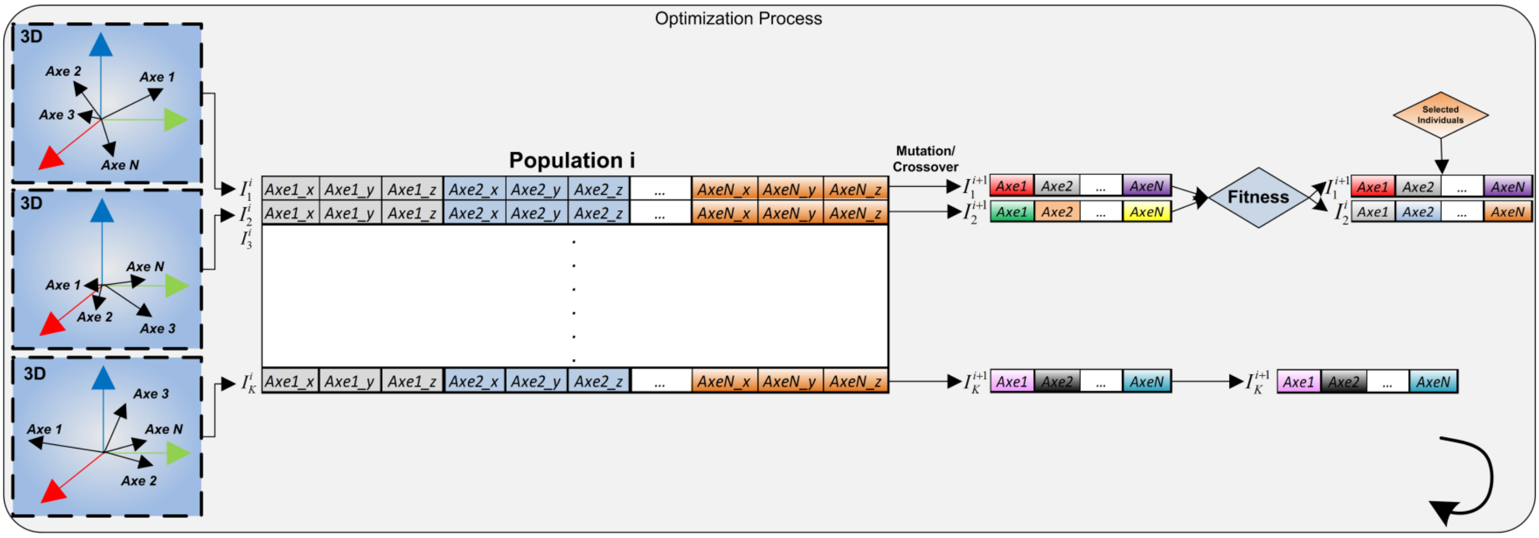


Figure 8. Optimization process in detail.

It is important to highlight the procedure of selecting the individuals in the population. For calculating the fitness value of I^i_k (initial individual) and I^{i+1}_k (new modified individual) and comparing them, a representation of the data is necessary with the resulting axes contained in both individuals by means of the SC algorithm. The SC algorithm will generate a new list of 3D datapoints for each individual. Then, a *generated distance matrix* for each one is computed. This matrix represents the distance between all of the instances of the new data using these new optimized axes. Now, the *generated distance matrix* for I^{i}_{k} and I^{i+1}_{k} can be fully compared to the original *target distance matrix*. The first comparison between the *generated distance matrix* of I^{i}_{k} and the original *target distance matrix* will generate a fitness value for I^{i}_{k} individual. The second comparison between the *generated distance matrix* of I^{i+1}_{k} and the original *target distance matrix* generates the fitness value for I^{i+1}_{k} individual. Of the two individuals, the one that has a higher fitness will be selected to continue in the evolution process.

During the OP the spatial positions of the axes will vary incrementally to obtain different tridimensional data embeddings. Finally, the individual with the best fitness value will be selected from the resulting population.

This individual represents the optimum axes configuration that, used as the input parameters to module II, allows a tridimensional representation of the points that fully approximates the original geometry of the n-dimensional data to be obtained. In other words, the selected individual will retain, as much as possible, the shape of the data that it had before the DR process.

To sum up, several aspects must be considered. First, in order to avoid heavy-computational costs, the *Euclidean distance* (L = 2) has been used in the calculation of the distance matrices. When using *Geodesic distances*, the calculation of distance graphs implies a huge increase in terms of computation time. Thus, it was decided to use the Euclidean distance as distance metrics in order to perform the distance measurements. Secondly, in relation to the crossover stage, the *exponential* operator has been considered. Finally, the stopping criterion for the DE algorithm is based on a number of generations, a fixed value of 2000 generations for each algorithm execution.

3.1.3 - Fitness

The MedVir algorithm allows the optimization using both local and global geometry preservation criteria.

Therefore, two different fitness functions have been implemented in order to solve the DR problem. The first one is directed towards improving the GGP, and uses the Sammon Error as quality assessment criterion. The second one deals with the improvement in the local topology or neighborhood preservation of the n-dimensional data, after the DR process. This function is evaluated by LCMC quality assessment criterion.

3.2 - Module 2. VR engine

Once the optimization module has produced results, the aim is just to obtain a 3D representation of the biological data. The input data of the visualization module consists of the set of optimized axes that will make a successful 3D embedding of the intrinsic geometric structure of the n-dimensional manifold possible (Fig. 9).

The Unity3D [47] visualization tool is used. Unity3D is a game engine designed for the creation of multiple 3D powerful interactive contents. Thus, the SC algorithm takes the output data of the optimization module as input data, and generates a 3D representation of the original n-dimensional dataset according to the optimized axes.

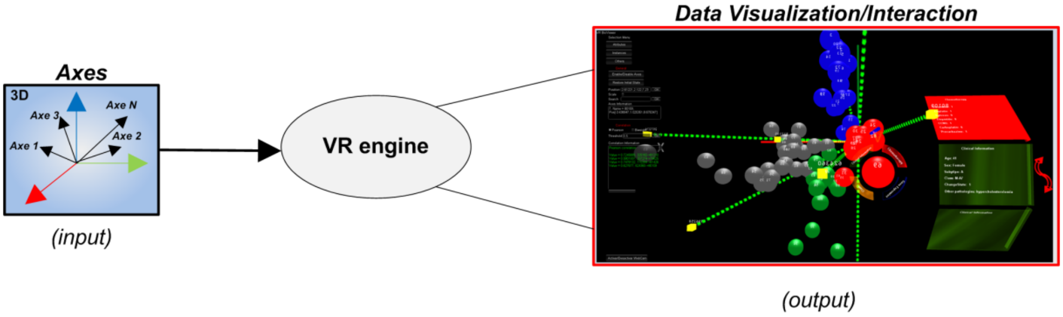


Figure 9. VR engine - Module II. The optimum axes configuration works as the input of the VR engine. By means of the SC algorithm, a 3D representation of the data will be made. Finally, the expert interprets the visualization.

The background of this dimensionality transformation is the SC algorithm [43]. Thus, once the data has been represented in 3D, the experts can interact with the elements of the visualization in real time. The aim is to find relationships, patterns or trends originally contained in the data before and after the execution the DR process in a visual and quickly way. These patterns can be expressed, for example, through the separation of the classes, clustering or relationships between the different attributes of the dataset. For example, the way of interacting with the coordinate axes could provide the medical experts valuable information. There might be many observations that the expert could be interested in. Thus, in order to increase the acquisition of knowledge from medical data, the possibility of interacting in real time with the data has been included.

The basic idea is that he can select several elements simply through simple actions. After selecting an element of the representation, for instance a coordinate axis, the expert can move that axis onto the tridimensional space and observe how the position of the points is rearranged in real time. This could inform the expert on the behavior of a certain dataset when he brings together, separates or even deletes an axis. Another option is to delete or add a particular attribute of the study. Thus, how the data are rearranged after cancelling or adding the contribution of that feature in the data can be visualized. It is, however, usually the case that after deleting a group of attributes, the final data representation does not vary. This could suggest a low level of importance of those attributes. Multiple attribute selection provides the possibility of analyzing the data variation of a set of features respect to the rest, in such a way that by selecting attributes identified as important, the behavior of the dataset could be modeled. However, he might be interested in observing how the samples are *clustered*. It is, by playing with the coordinate axes, scaling or rotating, one could see how the spheres move in and out of clusters. Maybe new clusters could be discovered.

For instance, a clear separation between classes of a dataset could be achieved, by moving the different coordinate axes. The expert is also able to observe the Biweight correlation [48] of the features. By means of a filtering process using a threshold value, the correlation values can be visualized in the form of color intensities. From here, the expert will be able to interpret the model according to his criteria and experience.

Furthermore, by selecting one of the patients (represented as spheres), just by clicking, all of the information is presented in virtual 3D panels (right-hand image). Each panel contains a different type of information such as, clinical or chemotherapy information on the patient (in the case of DNA Microarray Data). In this way, the expert can navigate around all the information on any specific patient with just one mouse click.

4 - Goodness of MedVir applied to DNA Microarray Data.

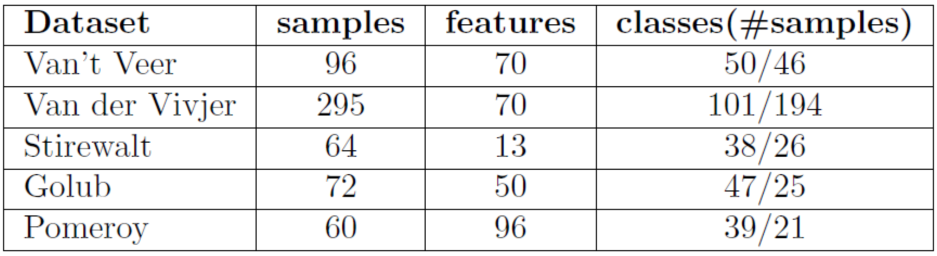
Once MedVir is presented, it has been evaluated and compared with the DR state of art algorithms. Particularly, MedVir has been studied when using real world data, such as DNA Microarray Data.

The evaluation mechanism is to analyze the geometry's preservation, once Dimensionality Reduction is carried out. That is, when the algorithm transforms the N multidimensional data into three dimensions, the geometry in 3D has to be as similar as possible to the geometry in N dimensions. For that, the quality assessment criteria, presented before, are used.

Thus, in this section the MedVir framework is evaluated and compared with the DR state-of-the-art algorithms, presented in the Methods section. In addition, representation of data with MedVir is presented in order to evaluate the visual qualities.

4.1 - Datasets

Five datasets have been used in this work: Two on Breast Cancer ([49] and [50]), two on Leukemia ([51] and [52]) and one on Medulloblastoma [53] (Table 1).



Dataset	samples	features	classes(#samples)
Van't Veer	96	70	50/46
Van der Vivjer	295	70	101/194
Stirewalt	64	13	38/26
Golub	72	50	47/25
Pomeroy	60	96	39/21

Table 1. Summary of the gene expressions used for testing the dimension reduction techniques.

The first one, the Van't Veer dataset [49], whose results have been approved by the FDA (Food and Drug Administration) were applied in a genomic profiling test called MammaPrint, that predicts whether the patients will suffer a relapse in breast cancer. The data are divided into two groups, learning (78 patients, 34 with a poor prognosis) and validation instances (19 patients, 12 patients with a poor prognosis). DNA microarray analysis [54-57] was used to determine the mRNA expression levels of approximately 24,500 genes for each patient. The microarray data are filtered to the 70 Van't Veer selected gene expression (accepted by the FDA as breast cancer biomarkers).

The second dataset, the Van der Vivjer dataset [50], is related to the previous research. It consists of tumors from a series of 295 consecutive women with breast cancer (194 with a poor prognosis) from the fresh frozen tissue bank of the Netherlands Cancer Institute. Again, the same 70 genes where selected [50] in this research.

The third dataset, from Stirewalt [51] is based on a study of 64 patients obtained from the Fred Hutchinson Cancer Research Center (FHCRC) or purchased from commercially available vendors. Analyses were also carried out using 38 normal samples and 26 with Acute Myeloid Leukemia (AML). In this study, 13 gene expressions were selected as a gene profile.

The other dataset on Leukemia is from Golub's research [52]. The data consists of 47 patients with acute lymphoblastic leukemia (ALL) and 25 patients with acute myeloid leukemia (AML). Each patient contains information of 7,129 genes. The Golub’s research reduced the number of gene expressions to 50, compiling a gene profile in order to differentiate the type of leukemia.

The data based on [53] Medulloblastomas has been used in several lines of research of great impact [53,58,59]. The data set has 60 samples containing 39 medulloblastoma survivors and 21 treatment failures. DNA microarray analysis was used to determine the mRNA expression levels of 5,920 known genes and 897 expressed sequence tags. Finally, 96 gene expressions were selected by [60] and used for our research.

4.2 - Experiments

In order to analyze the quality of MedVir, it has been compared, using the 5 gene profiles previously presented, with other different linear and nonlinear DR algorithms presented in the literature: Principal Components Analysis (PCA) [27, 28], Kernel PCA (KPCA) [21, 22], Locally Linear Embedding (LLE) [29, 30], The Laplacian Eigenmaps (LE) [31, 32], Diffusion Maps (DM) [33, 34] and Isomap [12, 16].

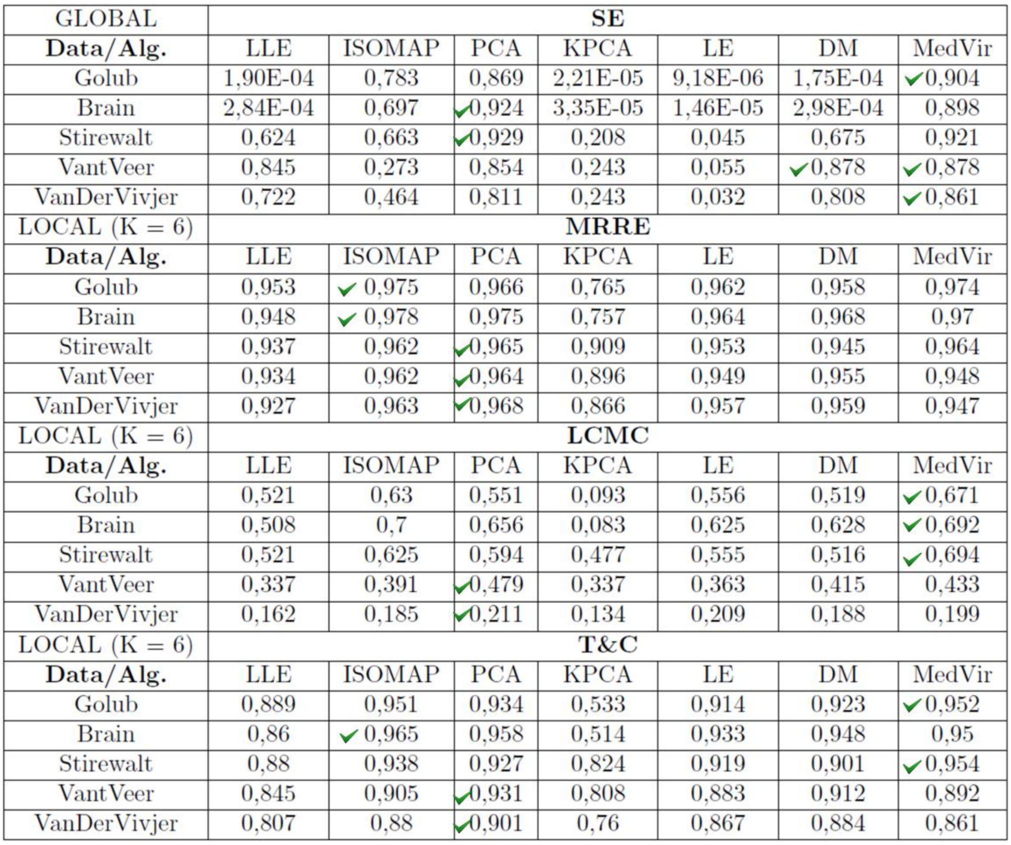
Although there is no standard framework to compare the DR algorithms, there are many different quality assessment criteria for evaluating the performance of the DR algorithms. Current approaches focus on evaluating the local-neighborhood-preservation and the global-structure-holding performance. By taking both properties into consideration, their intrinsic capability can be more faithfully reflected, and hence more a rational measure for the proper selection in real-life applications can be offered.

In this case, to compare the algorithms four geometry quality evaluation measures were selected: Sammon error [14], LCMC [37], T&C [38, 39] and MRRE [3,40,41]. All the four measures were represented in the same range [0, 1], where 0 represents the non-preservation of geometry and 1 represents a perfect preservation of geometry. In case of Sammon Error and MRRE, values were modified from the original measure (1 - *measure*). Values obtained from these measures were analyzed using two-sample Wilcoxon statistical test [61].

As said above, the MedVir algorithm allows an optimization process by using local and global parameters. In this work, two different optimizations were made: the first one with a local measure (LCMC) and the other with the global measure (SE). Both MedVir optimizations have to be compared with the other DR algorithms. In each comparison, because of the stochastic nature of MedVir, 10 executions of the algorithm were carried out, using the mean value to compare them. In this case, it was not necessary to present the standard deviation because they are less than 0.001 (low variability).

4.2.1 - MedVir optimized by LCMC

Firstly, MedVir optimized by LCMC compared with the other algorithms, were presented in the following tables. As we know, LCMC depends on the number of neighbors used. In this case, 6, 10 and 16 neighbors were used. Results using 6, 10 and 16 neighbors are presented in tables 2, 3 and 4.

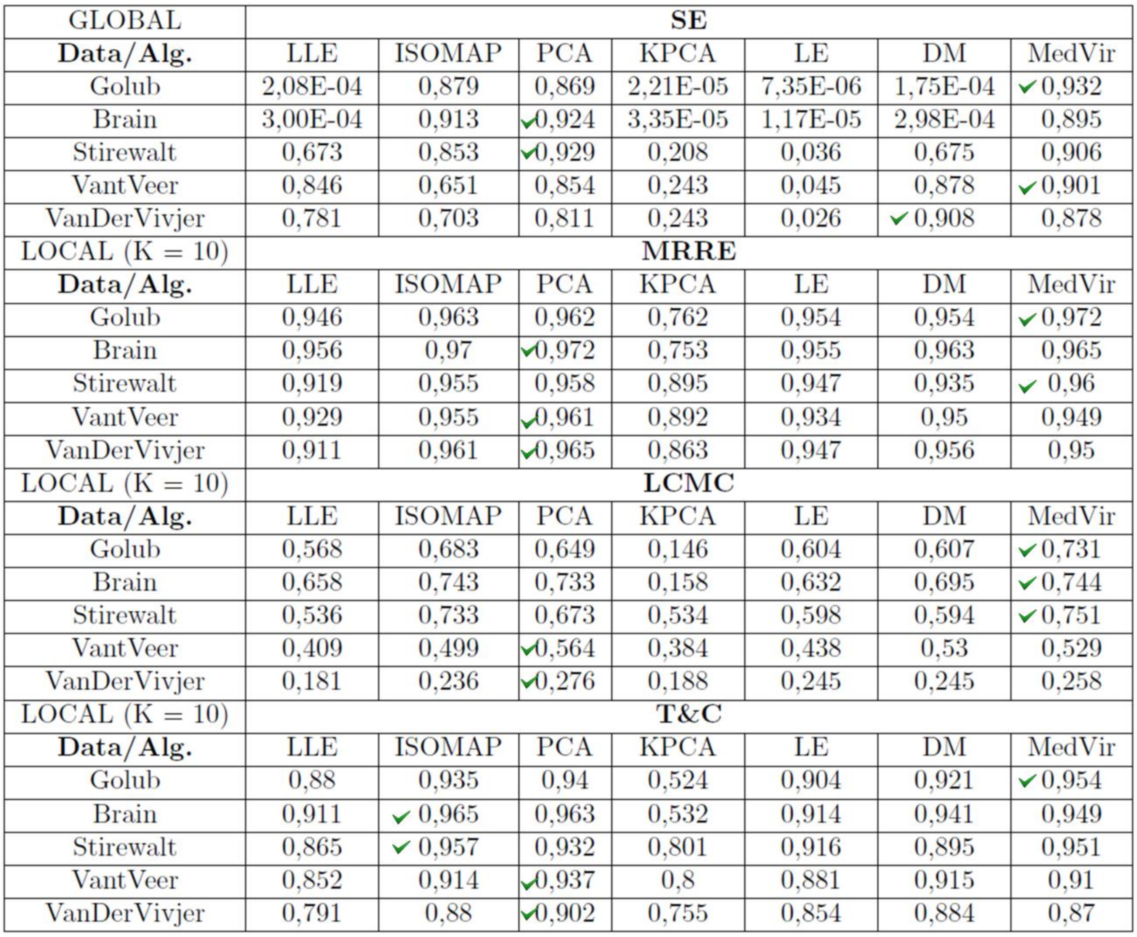


GLOBAL	SE						
Data/Alg.	LLE	ISOMAP	PCA	KPCA	LE	DM	MedVir
Golub	1,90E-04	0,783	0,869	2,21E-05	9,18E-06	1,75E-04	✓0,904
Brain	2,84E-04	0,697	✓0,924	3,35E-05	1,46E-05	2,98E-04	0,898
Stirewalt	0,624	0,663	✓0,929	0,208	0,045	0,675	0,921
VantVeer	0,845	0,273	0,854	0,243	0,055	✓0,878	✓0,878
VanDerVivjer	0,722	0,464	0,811	0,243	0,032	0,808	✓0,861
LOCAL (K = 6)	**MRRE**						
Data/Alg.	LLE	ISOMAP	PCA	KPCA	LE	DM	MedVir
Golub	0,953	✓ 0,975	0,966	0,765	0,962	0,958	0,974
Brain	0,948	✓ 0,978	0,975	0,757	0,964	0,968	0,97
Stirewalt	0,937	0,962	✓0,965	0,909	0,953	0,945	0,964
VantVeer	0,934	0,962	✓0,964	0,896	0,949	0,955	0,948
VanDerVivjer	0,927	0,963	✓0,968	0,866	0,957	0,959	0,947
LOCAL (K = 6)	**LCMC**						
Data/Alg.	LLE	ISOMAP	PCA	KPCA	LE	DM	MedVir
Golub	0,521	0,63	0,551	0,093	0,556	0,519	✓0,671
Brain	0,508	0,7	0,656	0,083	0,625	0,628	✓0,692
Stirewalt	0,521	0,625	0,594	0,477	0,555	0,516	✓0,694
VantVeer	0,337	0,391	✓0,479	0,337	0,363	0,415	0,433
VanDerVivjer	0,162	0,185	✓0,211	0,134	0,209	0,188	0,199
LOCAL (K = 6)	**T&C**						
Data/Alg.	LLE	ISOMAP	PCA	KPCA	LE	DM	MedVir
Golub	0,889	0,951	0,934	0,533	0,914	0,923	✓0,952
Brain	0,86	✓ 0,965	0,958	0,514	0,933	0,948	0,95
Stirewalt	0,88	0,938	0,927	0,824	0,919	0,901	✓0,954
VantVeer	0,845	0,905	✓0,931	0,808	0,883	0,912	0,892
VanDerVivjer	0,807	0,88	✓0,901	0,76	0,867	0,884	0,861

Table 2. MedVir optimized by LCMC and K = 6.

By analyzing these results it is possible to conclude that Isomap and LE are algorithms based only on local optima while LLE and DM are locally and globally optimal only when the number of instances in the data is high. PCA is a good algorithm in terms of local and global preservation. MedVir optimized by a local criteria, also optimizes the global criterion. Using the Wilcox statistical tests to compare the LCMC value of MedVir and the other algorithms (table 5), MedVir is better in all the cases (p - *value* < 0.001) except PCA in the Vant Veer dataset, Isomap in Brain dataset, and PCA and LE in the Van der Vivjer dataset.

Another conclusion drawn on seeing the results is that MRRE and T&C values cannot be evaluated and compared because there is no significant variability between the values.



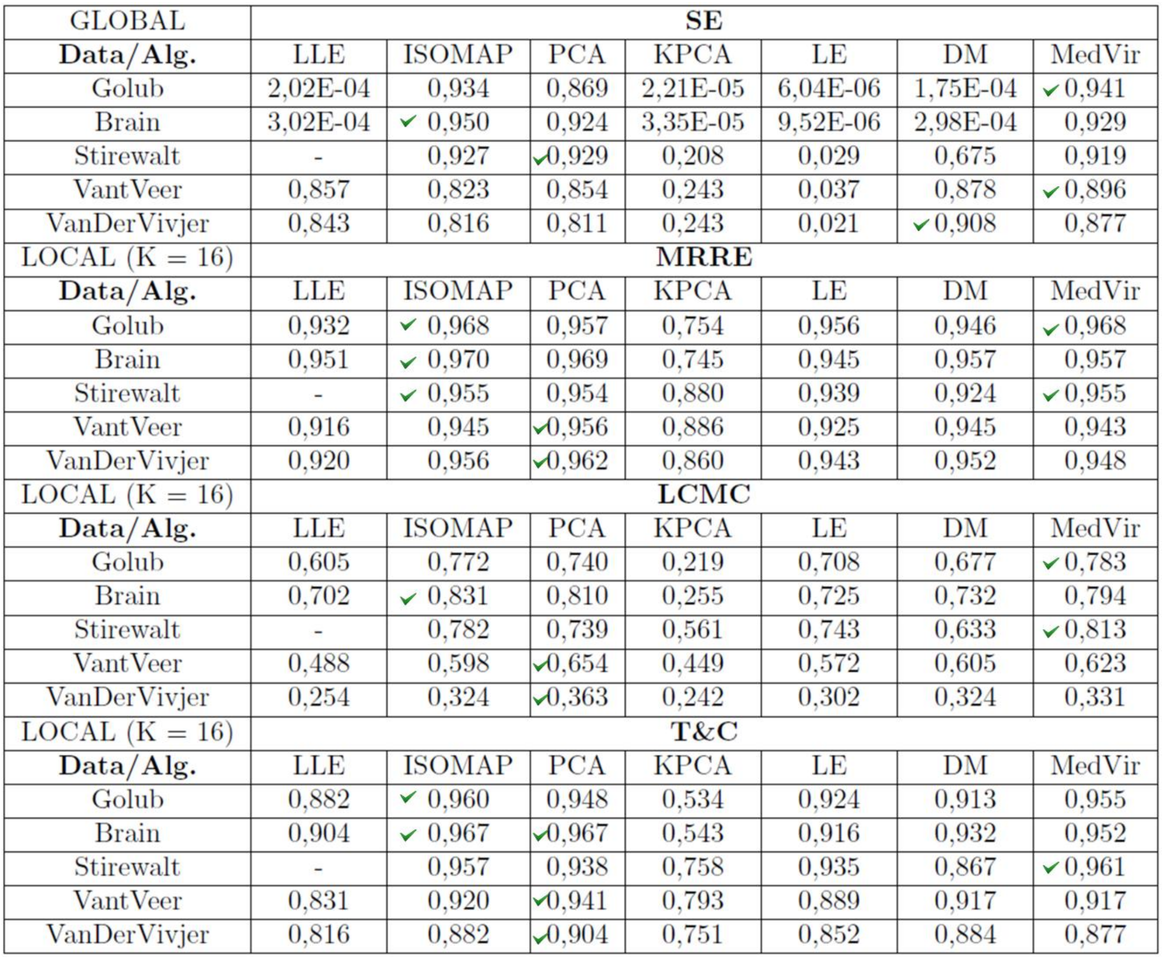
GLOBAL	SE						
Data/Alg.	LLE	ISOMAP	PCA	KPCA	LE	DM	MedVir
Golub	2,08E-04	0,879	0,869	2,21E-05	7,35E-06	1,75E-04	✓0,932
Brain	3,00E-04	0,913	✓0,924	3,35E-05	1,17E-05	2,98E-04	0,895
Stirewalt	0,673	0,853	✓0,929	0,208	0,036	0,675	0,906
VantVeer	0,846	0,651	0,854	0,243	0,045	0,878	✓0,901
VanDerVivjer	0,781	0,703	0,811	0,243	0,026	✓0,908	0,878
LOCAL (K = 10)	**MRRE**						
Data/Alg.	LLE	ISOMAP	PCA	KPCA	LE	DM	MedVir
Golub	0,946	0,963	0,962	0,762	0,954	0,954	✓0,972
Brain	0,956	0,97	✓0,972	0,753	0,955	0,963	0,965
Stirewalt	0,919	0,955	0,958	0,895	0,947	0,935	✓ 0,96
VantVeer	0,929	0,955	✓0,961	0,892	0,934	0,95	0,949
VanDerVivjer	0,911	0,961	✓0,965	0,863	0,947	0,956	0,95
LOCAL (K = 10)	**LCMC**						
Data/Alg.	LLE	ISOMAP	PCA	KPCA	LE	DM	MedVir
Golub	0,568	0,683	0,649	0,146	0,604	0,607	✓0,731
Brain	0,658	0,743	0,733	0,158	0,632	0,695	✓0,744
Stirewalt	0,536	0,733	0,673	0,534	0,598	0,594	✓0,751
VantVeer	0,409	0,499	✓0,564	0,384	0,438	0,53	0,529
VanDerVivjer	0,181	0,236	✓0,276	0,188	0,245	0,245	0,258
LOCAL (K = 10)	**T&C**						
Data/Alg.	LLE	ISOMAP	PCA	KPCA	LE	DM	MedVir
Golub	0,88	0,935	0,94	0,524	0,904	0,921	✓0,954
Brain	0,911	✓ 0,965	0,963	0,532	0,914	0,941	0,949
Stirewalt	0,865	✓ 0,957	0,932	0,801	0,916	0,895	0,951
VantVeer	0,852	0,914	✓0,937	0,8	0,881	0,915	0,91
VanDerVivjer	0,791	0,88	✓0,902	0,755	0,854	0,884	0,87

Table 3. MedVir optimized by LCMC and K = 10.

The results using 10 neighbors (K = 10) are presented in table 3. There we can see that the local values of all of the algorithms were improved. It is normal because the greater the number of nearest neighbors around, the lower margin of error the algorithm will have. As the number of neighbors increases, the SE measure will also increase (so the Sammon error decreases).

In general, there is the same tendency in the results as with K = 6. Using the Wilcox statistical tests to compare the LCMC value of MedVir and the other algorithms (table 5), MedVir is better in all the cases (p - *value* < 0.001) except PCA and DM in the Vant Veer dataset, Isomap in Brain dataset, and PCA in the Van der Vivjer dataset.

Finally, the results using 16 neighbors (K = 16) are presented in table 4. Again, the local parameters were improved and, therefore, the global parameters also improved. In the case of the LLE algorithm with the Stirewalt dataset, there was no way to executing the algorithm and obtaining the results with 16 neighbors.



GLOBAL	**SE**						
Data/Alg.	LLE	ISOMAP	PCA	KPCA	LE	DM	MedVir
Golub	2,02E-04	0,934	0,869	2,21E-05	6,04E-06	1,75E-04	✓0,941
Brain	3,02E-04	✓ 0,950	0,924	3,35E-05	9,52E-06	2,98E-04	0,929
Stirewalt	-	0,927	✓0,929	0,208	0,029	0,675	0,919
VantVeer	0,857	0,823	0,854	0,243	0,037	0,878	✓0,896
VanDerVivjer	0,843	0,816	0,811	0,243	0,021	✓0,908	0,877
LOCAL (K = 16)	**MRRE**						
Data/Alg.	LLE	ISOMAP	PCA	KPCA	LE	DM	MedVir
Golub	0,932	✓ 0,968	0,957	0,754	0,956	0,946	✓0,968
Brain	0,951	✓ 0,970	0,969	0,745	0,945	0,957	0,957
Stirewalt	-	✓ 0,955	0,954	0,880	0,939	0,924	✓0,955
VantVeer	0,916	0,945	✓0,956	0,886	0,925	0,945	0,943
VanDerVivjer	0,920	0,956	✓0,962	0,860	0,943	0,952	0,948
LOCAL (K = 16)	**LCMC**						
Data/Alg.	LLE	ISOMAP	PCA	KPCA	LE	DM	MedVir
Golub	0,605	0,772	0,740	0,219	0,708	0,677	✓0,783
Brain	0,702	✓ 0,831	0,810	0,255	0,725	0,732	0,794
Stirewalt	-	0,782	0,739	0,561	0,743	0,633	✓0,813
VantVeer	0,488	0,598	✓0,654	0,449	0,572	0,605	0,623
VanDerVivjer	0,254	0,324	✓0,363	0,242	0,302	0,324	0,331
LOCAL (K = 16)	**T&C**						
Data/Alg.	LLE	ISOMAP	PCA	KPCA	LE	DM	MedVir
Golub	0,882	✓ 0,960	0,948	0,534	0,924	0,913	0,955
Brain	0,904	✓ 0,967	✓0,967	0,543	0,916	0,932	0,952
Stirewalt	-	0,957	0,938	0,758	0,935	0,867	✓0,961
VantVeer	0,831	0,920	✓0,941	0,793	0,889	0,917	0,917
VanDerVivjer	0,816	0,882	✓0,904	0,751	0,852	0,884	0,877

Table 4. MedVir optimized by LCMC and K = 16.

Using the Wilcox statistical tests to compare the LCMC value of MedVir and the other algorithms (table 5), MedVir is better in all of the cases (p - *value* < 0.001) except PCA in the Vant Veer dataset and Isomap and PCA in the Brain dataset.

In general terms, PCA and Isomap algorithms work successfully in preservation when the number of neighbors is high. On the other hand, KPCA is an algorithm that does not preserve the geometry. MedVir is better than the rest of the DR algorithms in most of the cases, except PCA with datasets with a large number of instances, and Isomap with the Brain cancer dataset.

MedVir optimized by a local parameter is not only strong in local preservation, but is also strong in global preservation. By analyzing the results of the Sammon error when K = 16 and using the Wilcox statistical test to compare them (table 6), we demonstrate that MedVir is almost better (p - *value* < 0.001) than the other DR algorithms.

	LLE			**ISOMAP**		
MedVir VS.	k=6	k=10	k=16	k=6	k=10	k=16
Golub	0,00003102	0,0000317	0,00003124	0,00003102	0,0000317	3,12E-05
Brain	0,00002925	0,00003057	0,00002946	0,9997	0,3447	1
Stirewalt	0,0000299	0,0000317	0*	0,0000299	0,0000317	0,00003147
VantVeer	0,00003193	0,00003124	0,0000317	0,00003193	0,00003124	0,0000317
VanDerVivjer	0,0000317	0,00003193	0,0000317	0,0000317	0,00003193	0,0001149
	PCA			**KPCA**		
MedVir VS.	k=6	k=10	k=16	k=6	k=10	k=16
Golub	0,00003102	0,0000317	0,00003124	0,00003102	0,0000317	0,00003124
Brain	0,00002925	0,00003057	1	0,00002925	0,00003057	0,00002946
Stirewalt	0,0000299	0,0000317	0,00003147	0,0000299	0,0000317	0,00003147
VantVeer	1	1	1	0,00003193	0,00003124	0,0000317
VanDerVivjer	1	1	1	0,0000317	0,00003193	0,0000317
	LE			**DM**		
MedVir VS.	k=6	k=10	k=16	k=6	k=10	k=16
Golub	0,00003102	0,0000317	0,00003124	0,00003102	0,0000317	0,00003124
Brain	0,00002925	0,00003057	0,00002946	0,00002925	0,00003057	0,00002946
Stirewalt	0,0000299	0,0000317	0,00003147	0,0000299	0,0000317	0,00003147
VantVeer	0,00003193	0,00003124	0,0000317	0,0001156	0,8021	0,0000317
VanDerVivjer	1	0,00003193	0,0000317	0,0000317	0,00003193	0,0000317

Table 5. Wilcoxon test values comparing MedVir optimized by LCMC with the other algorithm.

MedVir VS.	**LLE**	**ISOMAP**	**PCA**	**KPCA**	**LE**	**DM**
Golub	3,19E-05	0,0576	3,19E-05	3,19E-05	3,19E-05	3,19E-05
Brain	3,19E-05	0,9931	0,0008	3,19E-05	3,19E-05	3,19E-05
Stirewalt	0*	8,59E-04	0,5161	3,19E-05	3,19E-05	3,19E-05
VantVeer	3,19E-05	3,19E-05	3,19E-05	3,19E-05	3,19E-05	0,0007
VanDerVivjer	3,19E-05	3,19E-05	3,19E-05	3,19E-05	3,19E-05	1,0000

Table 6. Wilcoxon test of SE values when MedVir were optimized by LCMC.

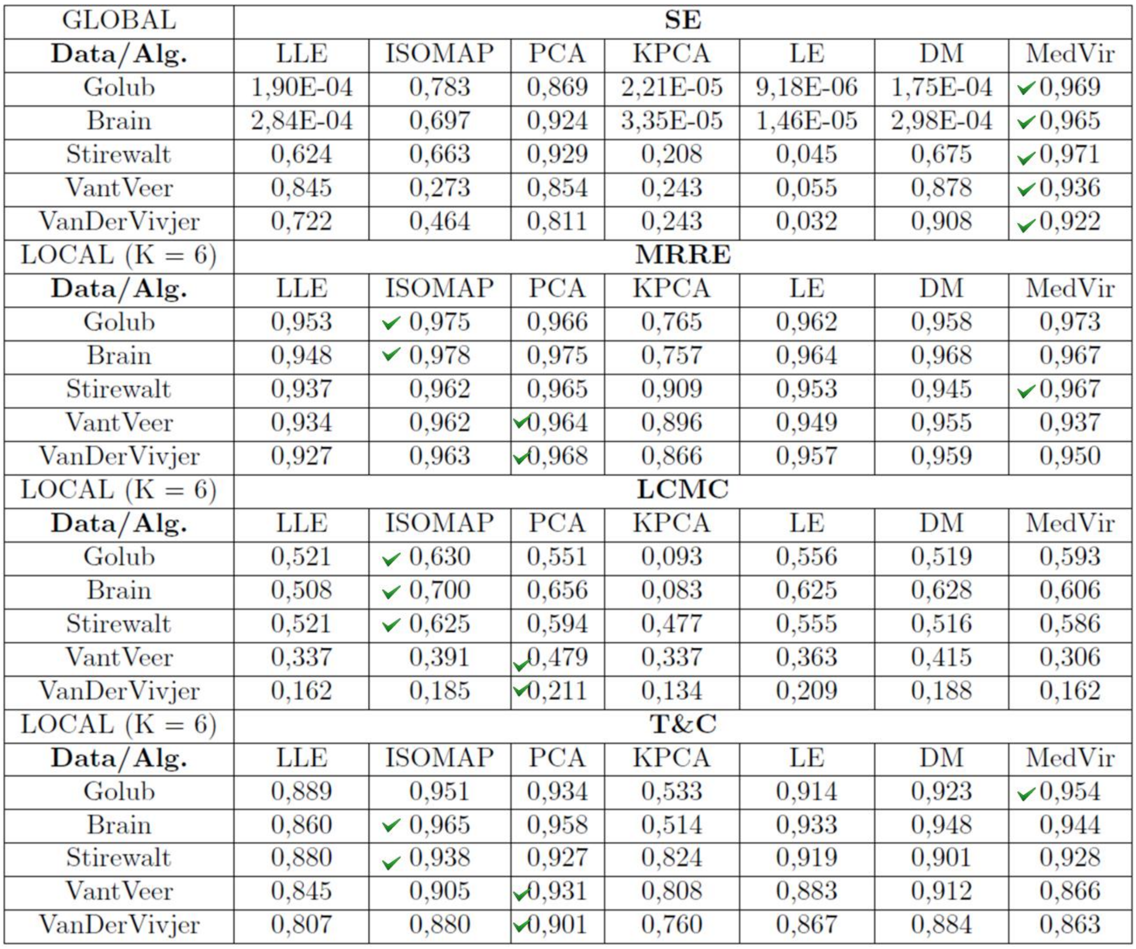
4.2.2 - MedVir optimized by SE

To optimize MedVir using the Sammon error measure, it is not necessary to indicate the K value. However, in order to calculate the local measures, it is necessary. Thus, the most restrictive K is selected (K = 6), in order to compare with the local optimization.

The results of the comparison are presented in table 7. There, the results obtained in the SE measures are better than those obtained in the local optimization. However, the local values obtained here are worse than those obtained in local optimization. The local values obtained in MedVir are within the mean of the values of the other DR algorithms.

On the other hand, the results of the MRRE and T&C measures do not vary much in respect to those obtained in the local optimization. Thus, we conclude that these two measures are not useful in validating the geometry preservation.

Finally, using the Wilcox statistical tests to compare the MedVir SE values and the other algorithms, in this case MedVir is superior to all the algorithms in all the cases (p - *value* < 0.001).



GLOBAL	SE						
Data/Alg.	LLE	ISOMAP	PCA	KPCA	LE	DM	MedVir
Golub	1,90E-04	0,783	0,869	2,21E-05	9,18E-06	1,75E-04	✓0,969
Brain	2,84E-04	0,697	0,924	3,35E-05	1,46E-05	2,98E-04	✓0,965
Stirewalt	0,624	0,663	0,929	0,208	0,045	0,675	✓0,971
VantVeer	0,845	0,273	0,854	0,243	0,055	0,878	✓0,936
VanDerVivjer	0,722	0,464	0,811	0,243	0,032	0,908	✓0,922
LOCAL (K = 6)	**MRRE**						
Data/Alg.	LLE	ISOMAP	PCA	KPCA	LE	DM	MedVir
Golub	0,953	✓ 0,975	0,966	0,765	0,962	0,958	0,973
Brain	0,948	✓ 0,978	0,975	0,757	0,964	0,968	0,967
Stirewalt	0,937	0,962	0,965	0,909	0,953	0,945	✓0,967
VantVeer	0,934	0,962	✓0,964	0,896	0,949	0,955	0,937
VanDerVivjer	0,927	0,963	✓0,968	0,866	0,957	0,959	0,950
LOCAL (K = 6)	**LCMC**						
Data/Alg.	LLE	ISOMAP	PCA	KPCA	LE	DM	MedVir
Golub	0,521	✓ 0,630	0,551	0,093	0,556	0,519	0,593
Brain	0,508	✓ 0,700	0,656	0,083	0,625	0,628	0,606
Stirewalt	0,521	✓ 0,625	0,594	0,477	0,555	0,516	0,586
VantVeer	0,337	0,391	✓0,479	0,337	0,363	0,415	0,306
VanDerVivjer	0,162	0,185	✓0,211	0,134	0,209	0,188	0,162
LOCAL (K = 6)	**T&C**						
Data/Alg.	LLE	ISOMAP	PCA	KPCA	LE	DM	MedVir
Golub	0,889	0,951	0,934	0,533	0,914	0,923	✓0,954
Brain	0,860	✓ 0,965	0,958	0,514	0,933	0,948	0,944
Stirewalt	0,880	✓ 0,938	0,927	0,824	0,919	0,901	0,928
VantVeer	0,845	0,905	✓0,931	0,808	0,883	0,912	0,866
VanDerVivjer	0,807	0,880	✓0,901	0,760	0,867	0,884	0,863

Table 7. MedVir optimized by the Sammon Error.

5 - Data Analysis

This section attempts to confirm that the knowledge, based on the visualization, which MedVir framework provides is totally valid. Thus, it is intended to accept the MedVir's hypothesis about the data as correct. In other words, a mechanism for validating MedVir results is needed. Besides that, several conclusions are extracted.

First of all, the matter is not as trivial as it might seem a priori. We must bear in mind that when performing a DR process over a multidimensional dataset, a minimal loss of information (a DR process always involves a loss of information) occurs. This loss of information could be minimally causing distortions in the representation of the data and placing the data points in wrong spatial positions, thus causing confusion to the expert that analyzes the data and leading to invalid hypotheses.

Although MedVir preserves to the maximum the geometry of the data and reduces very significantly this loss of information, the experience tells us it is very appropriate to carry out further analysis.

The use of Distance Matrices [63] in the field of DNA Microarray Data is widely accepted by the scientific community. Therefore, a good way to know if the visual representations obtained for each dataset correspond to the real nature of each one would be to compare these visualizations with the distance matrix *M* of each dataset.

A *distance matrix* shows visual similarities between the gene expression profiles of patients. Thus, the distance between two patients with very similar gene expression profiles will be close to zero (represented in blue tones in the distance matrix), while the distance between two patients with disparities in gene expression profiles will be much higher (red and yellow tones), (*M* is shown in figure 10 - left image). The distance metric used in the computation of Distance Matrix is L = 2, the *Euclidean* distance.

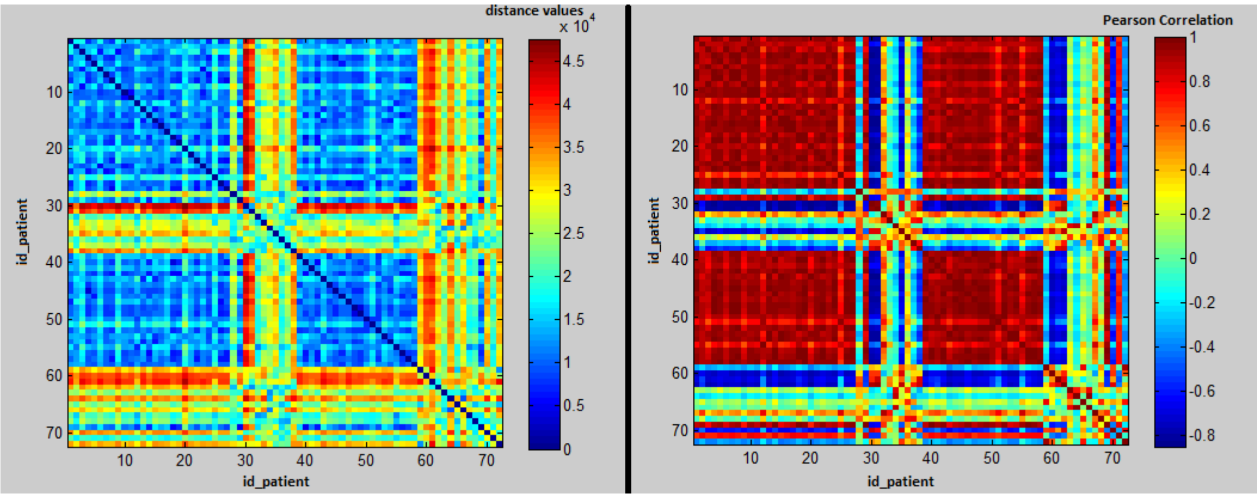
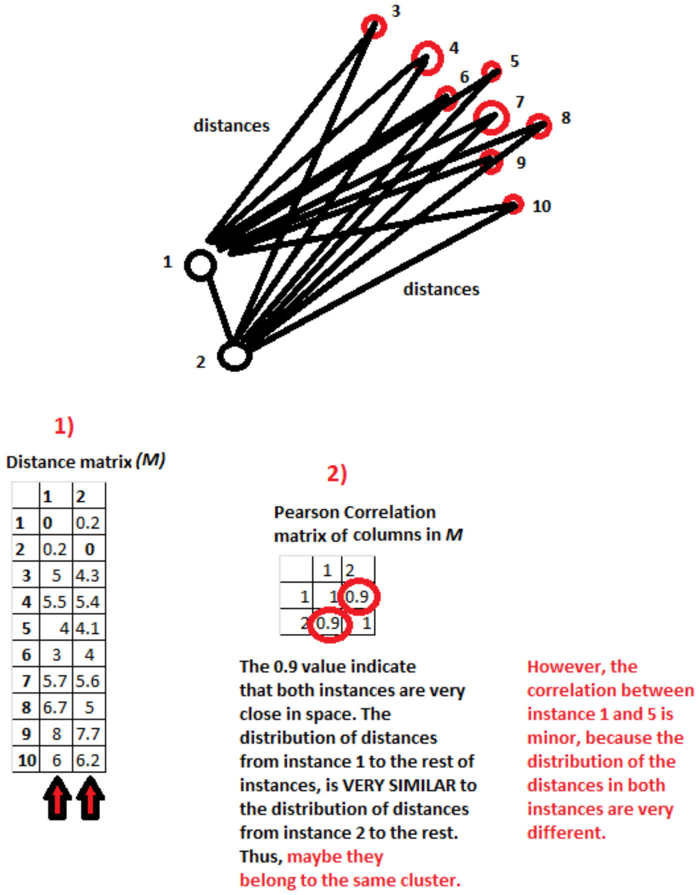


Figure 10. Matrices. Left: Distance matrix between patients of a study (blue tones indicate a high similarity between gene expression profiles of each patient, while red and yellow tones a low similarity). Right: Pearson Correlation Matrix of the Distance matrix (red tones indicate a high Pearson Correlation between distance distributions, while blue tones a low correlation).

By using M, a new matrix called *Pearson correlations matrix* (*P*) of the distance matrix *M* could also be computed. This new matrix (*P*) will tend to highlight these differences between patients. Moreover, now the values appear standardized between 1 and -1. *P* shows the degree of similarity in the distribution of all distances from one patient to the rest of patients. That is, imagine that two patients (p_1 and p_2) have very similar gene expression profiles (and therefore they are located very close in space), so it is expected that the distribution in the distances calculated from p_1 to other patients in the study is very similar to the distribution in the distances calculated p_2 to the remaining patients (Figure *11*).



Distance matrix *(M)*

	1	2
1	0	0.2
2	0.2	0
3	5	4.3
4	5.5	5.4
5	4	4.1
6	3	4
7	5.7	5.6
8	6.7	5
9	8	7.7
10	6	6.2

Pearson Correlation matrix of columns in *M*

	1	2
1	1	0.9
2	0.9	1

The 0.9 value indicate that both instances are very close in space. The distribution of distances from instance 1 to the rest of instances, is VERY SIMILAR to the distribution of distances from instance 2 to the rest. Thus, maybe they belong to the same cluster.

However, the correlation between instance 1 and 5 is minor, because the distribution of the distances in both instances are very different.

Figure 11. Graphical meaning of Pearson Correlation Matrix of M. Circles labeled with numbers represent the instances. The black lines symbolize the distances. 1) Distance matrix of instances 1 and 2 to the rest of instances. 2) Pearson Correlation matrix of *M*.

Consequently, patients with similar gene expression profiles will show a high correlation between the distributions of their distances to other patients. It is an approach similar to the distance matrix, but it tends to further highlight the differences and discriminate the differences between patients. It is a different way of showing similitudes between patients, based on Pearson correlation matrices.

5.1 - Stirewalt's dataset

Figure 12 shows the tridimensional visualization of the Stirewalt's supervised dataset using MedVir. First, the colored spheres refer to the dataset samples (patients). The coordinate axes represent each of the gene expressions included in the study (13 features). Each coordinate axis is accompanied by the name of the corresponding *id_code* gene. By taking into account the nature of the possible classes or labels of the dataset, a blue color represents normal tissue samples while the red one represents cancer samples identified as type AML (Acute Myeloid Leukemia). A complete discrimination between the two classes is appreciated, which suggests that the gene profile selection conducted by the author is quite valid [51].

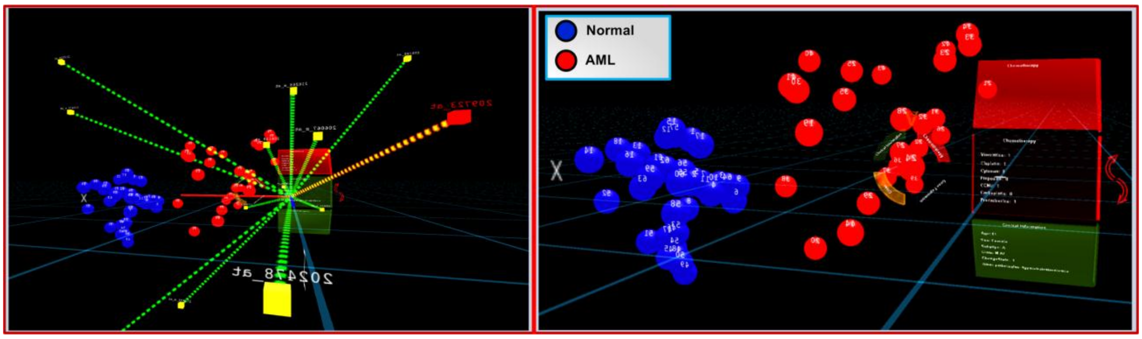


Figure 12. Visualization of the Stirewalt's dataset. The blue spheres represent normal tissue samples. Nevertheless, red spheres represent AML samples. The dotted green lines represent coordinate axes (genes). A separation of classes is clearly visible. Finally, the interface shows the different clinical information of the sample, if it exists.

Drawing conclusions… in order to validate the visualization in MedVir, we will use both $M_{\_stirew}$ and $P_{\_stirew}$ to confirm several observations. For example, the previous visualization shows a clear discrimination between the two classes of the study. Moreover, the patients labeled as *normal* (blue spheres) are spatially more compact than the *AML* ones (maybe the differences between *AML* patients, in terms of gene expression profiles, are greater than *normal* patients). Let's see these two observations by using $M_{\_stirew}$ and $P_{\_stirew}$.

First, $M_{\_stirew}$ and $P_{\_stirew}$ produce quite consistent results in accordance with the visualization obtained after running MedVir (figure 13). If we look at the main diagonal of both matrices, it is clear that patient number 1 to 18, and 45 to 64 are grouped showing very similar color intensity (black circles in figure 13). Therefore, these patients may show a pattern in their gene expression profiles, suggesting that they belong to the same typology, as MedVir previously indicated. Now, let's check the original dataset labels assigned to these patients and we found that the label is the same type for absolutely all of these patients (label 1, normal tissue). Besides that, when visualizing the dataset with MedVir, note that these patients are grouped uniquely and well differentiated from the other patients (blue spheres). Hence, in this case MedVir has correctly represented the original nature of the data.

On the other hand, the difference in density appreciated in MedVir in AML patients can be also studied by $M_{\_stirew}$ and $P_{\_stirew}$. Maybe, this difference in density could indicate the existence of sub-typologies inside AML type. In order to confirm the effectiveness of MedVir in this sense, one could make a further study by inspecting by $M_{\_stirew}$ and $P_{\_stirew}$.

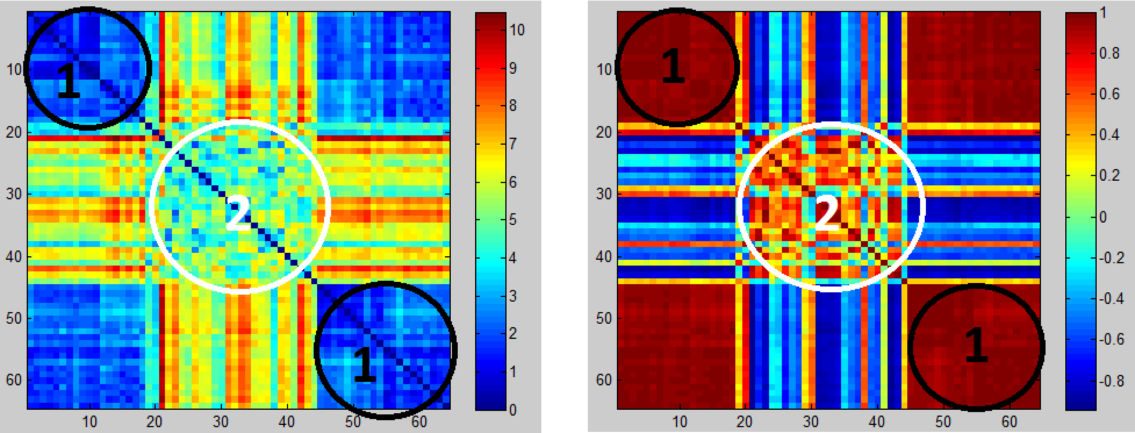


Figure 13. Distance and Pearson Correlation matrix obtained in Matlab. Left: *M* matrix of Stirewalt's dataset ($M_{\_stirew}$) . Right: *P* matrix of Stirewalt's dataset ($P_{\_stirew}$). According to the labels in the original dataset, the two different classes (1: normal, and 2: AML patients) have been marked in the figure. Black circles indicate *normal* patients (1 to 18, and 45 to 64). White circles indicate *AML* patients (19-44).

$M_{\_stirew}$ and $P_{\_stirew}$ indicate that patients labeled as *AML* (19-44) does not show a distinguishable clear and homogenous color group between them (as occurs in *normal* patients). This fact could indicate the existence of sub-typologies inside *AML* type because of the difference between the gene expression profiles of these patients. This difference in density identified by MedVir has been successfully demonstrated by $M_{\_stirew}$ and $P_{\_stirew}$. Thus, MedVir does not fail when representing the data geometry.

MedVir captures the original intrinsic geometric structure of the data and translates it to a tridimensional space.

5.2 - Van't Veer's dataset

Figure 14 presents the Van't Veer dataset in three dimensions using MedVir. The red color symbolizes non-relapse patients while the blue represents those patients who suffered a breast cancer relapse. Here, the class separation is not completely obvious. There are a small number of patients that overlap between them. This fact confirms the 83% value of classification accuracy achieved by the author [49].

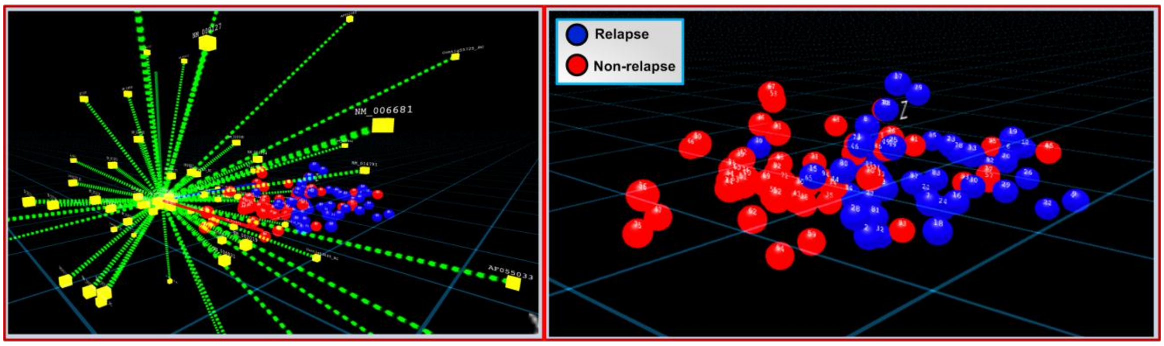


Figure 14. The Van't Veer dataset. Representation of the breast cancer disease, by using the Laura Van't Veer feature selection. The red spheres symbolize non-relapse patients in breast cancer. The blue spheres represent patients who do suffer a breast cancer relapse. Left: Visualization with the coordinate axes. Right: Representation without coordinate axes.

In order to improve the current result, MedVir allows experts to modify the representation of the data. In this case, the shortest axes have minimal influence in the global representation of the data. Thus, if Laura Van't Veer could have used this tool, she would have seen that certain gene expressions do not provide information for the separation of the classes. So she probably would have erased these expressions from the gene profile.

Figure 15 shows an example of this idea, where 15 gene expressions have been removed and the representation of data (geometry preservation) is more or less the same.

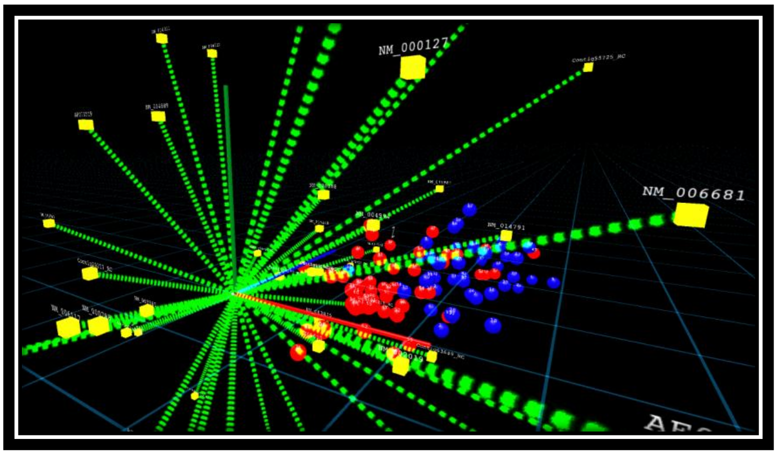


Figure 15. The Van't Veer dataset. Example of representation of breast cancer, modified by an expert biologist.

Drawing conclusions… when representing the dataset by means of MedVir, we are struck by the fact that some patients who are labeled as non-relapse (in red) are overlapped with relapse patients (in blue). From now to the end, these patients will be called 'red Intruders'. *[Let us briefly recall the nature of this dataset. When measuring the gene expression profiles for each patient (at time t = 0 of the chemotherapy treatment, once all the patients have contracted the disease) is not carried out a labeling process, but this was done within 5 years after this moment, thus checking whether the patient had relapsed or not in breast cancer. That is when the labeling process of patients was carried out]*.

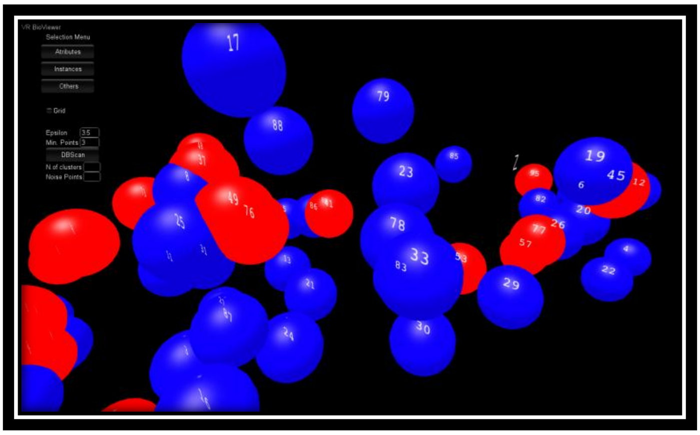


Figure 16. The 'red Intruders'. They are patients labeled as *non-relapse* and they have similar gene expression profile to *relapse* patients.

In order to confirm the veracity of this observation in MedVir, we use the $M_{\_veer}$ and $P_{\_veer.}$ One possible reason to justify the 'red Intruders' (patients number 41, 45, 48, 49, 53, 57, 76, 77 and 95) could be the following: patients whose gene expression profiles were taken in t = 0, from the beginning had very similar profiles to patients who eventually relapsed in breast cancer disease (this is discussed in the next paragraph). However, what differentiated from one to another in relapse or not relapse in the disease was the chemotherapy treatment they received. Therefore, the chemotherapy effect toke effect on the 'red Intruders' and managed to save them from a relapse of the disease.

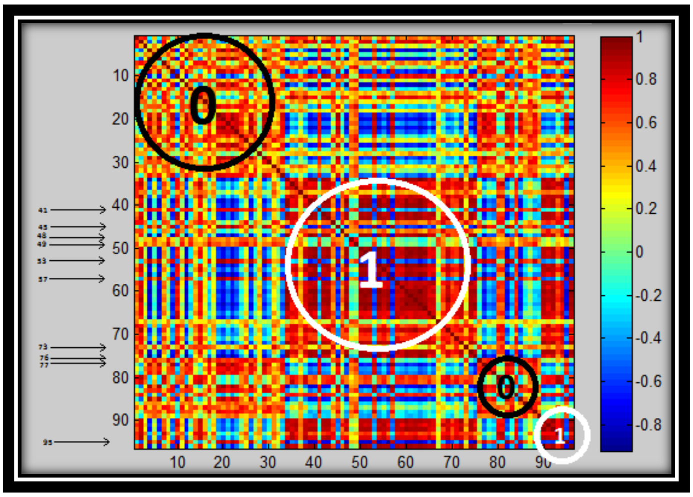


Figure 17. Pearson Correlation matrix obtained in Matlab (P_{veer}). According to the labels in the original dataset, the two different classes (0: *non-relapse,* and 1: *relapse* patients) have been marked in the figure. Black circles indicate *non-relapse* patients. White circles indicate *relapse* patients. The 'red Intruders' have been marked.

If we take a look at P_{veer}, one could be able to corroborate this hypothesis by MedVir. In figure 17 the *P* matrix is shown and the 'red Intruders' have been indicated by black arrows. We see that, absolutely every one of the 'red Intruders' identified through MedVir fit in the matrices as patients who do not have the same typology as other patients who have been labeled with, as they have a different color in the matrix, therefore indicating dissimilarity (see patients marked with black arrows in the clusters formed in the main diagonal). This is an example of obtaining knowledge through MedVir, and the effectiveness of MedVir framework has been demonstrated by P_{veer}.

Besides that, when visualizing with MedVir we can also guess a set of patients that are labeled as relapse patients (in blue) and yet are overlapping with non-relapse patients (in red). We call these patients the 'blue intruders' (1, 2, 7, 10, 14, 15, 16, 18, 25, 28, 31, 32, 80, 81 and 84. Figure 18). This is the opposite case above, these are patients with gene expression profiles similar to patients who did not finally relapse in the disease, and nevertheless the 'blue intruders' relapsed in the disease. Here, the chemotherapy treatment may not take effect, or just those patients were exposed to environments that favored the relapse in the disease.

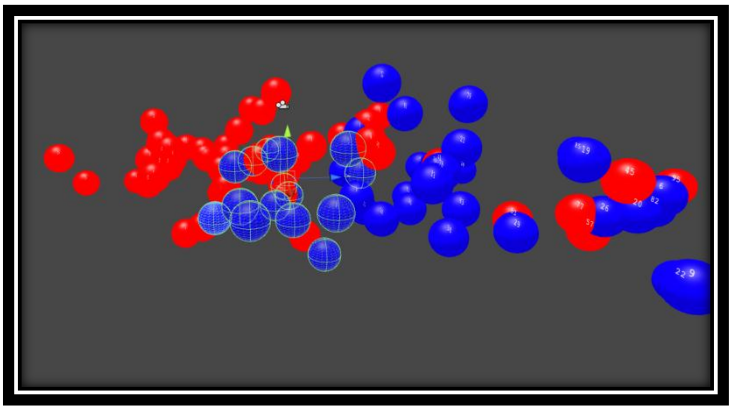


Figure 18. The 'blue Intruders'. They are patients labeled as *relapse* and they have similar gene expression profile to non-*relapse* patients.

The correlation matrix P_{veer} is shown again (see patients marked with black arrows in the clusters formed in the main diagonal).

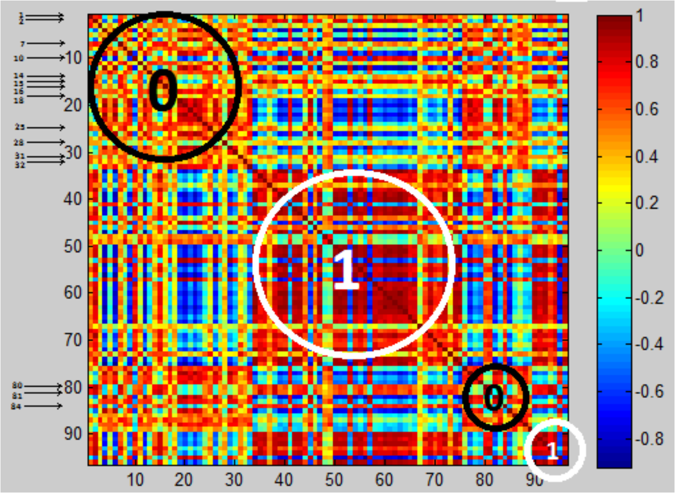


Figure 19. Pearson Correlation matrix obtained in Matlab. The 'blue Intruders' have been marked in black arrows.

5.3 - Golub's dataset

Figure 20 shows the tridimensional visualization of the Golub's dataset using MedVir. As said above, the colored spheres refer to the dataset samples (in this case, they are patients). The coordinate axes represent each of the gene expressions included in the study (50 selected features by *Golub et al*). The blue color represents *ALL* (Acute Lymphoblastic Leukemia) tissue samples while the red one represents cancer samples identified as type *AML* (Acute Myeloid Leukemia). Almost a complete discrimination between the two classes is appreciated, which suggests that the gene profile selection conducted by the author is correct [52].

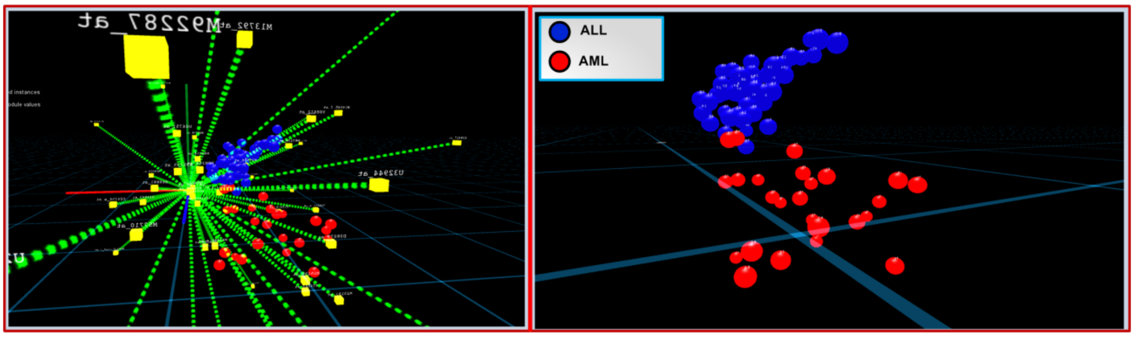


Figure 20. Visualization of the Golub's dataset. The blue spheres represent ALL tissue samples. Nevertheless, red spheres represent AML samples.

Drawing conclusions… The first thing we can see when visualizing Golub's dataset in MedVir is two groups clearly separated and distinct. At first sight, something that stands out is the difference in ***density*** in both clusters. The group of patients labeled as *ALL* patients (in blue) is much more compact and spatially delimited than *AML* patients (in red). The latter group is much more spatially dispersed, which a priori could be considered a greater variation in the differences between the gene expression profiles of these patients.

This observation in MedVir could lead to the discovery that this group is likely to have sub-groups of patients. Although all patients in this cluster have been labeled by the expert as AML type, they may not have been taken into account possible sub-types of AML patients. Let's see the matrices M_{golub} and P_{golub} of this dataset.

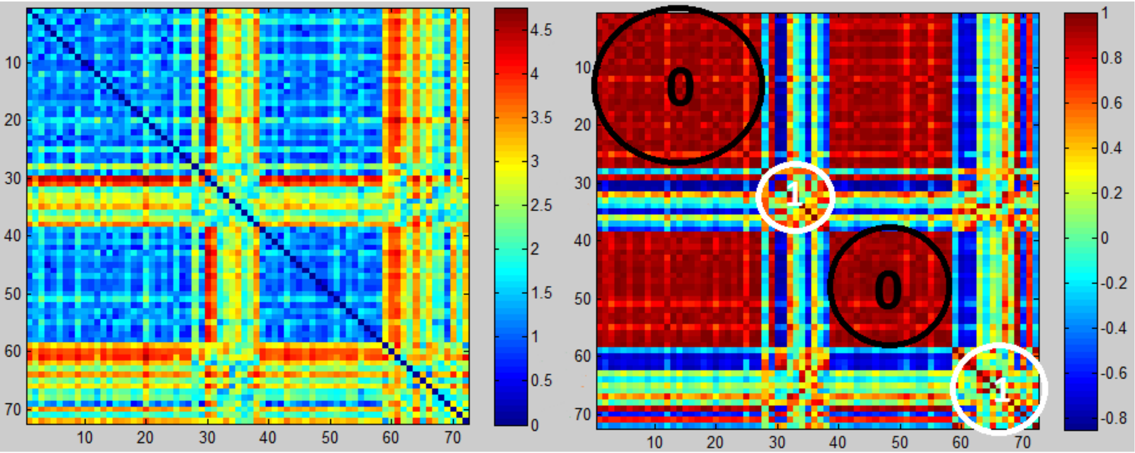


Figure 21. Distance and Pearson Correlation matrix obtained in Matlab. Left: *M* matrix of Golub's dataset (*M*$_{\_golub}$) . Right: *P* matrix of Golub's dataset (*P*$_{\_golub}$). According to the labels in the original dataset, the two different classes (0: *ALL*, and 1: *AML* patients) have been marked in the figure. Black circles indicate *ALL* patients (1 to 27, and 39 to 58). White circles indicate *AML* patients (28-38, 59-72).

As guessed, inside the group labeled as 1 (AML) there are significant differences between patients. Just take a look at the matrix *P*$_{\_golub}$ to observe that patients number 28-38 and 59-72 have a more heterogeneous nature than the other patients labeled as 0 (ALL). This would explain the dispersion of this group that will result in greater spatial separation when representing the data in MedVir.

The ***second*** assessment we conducted after running MedVir is that, at first glance, there are two AML patients ('red Intruders'. Figure 22) that are overlapped with ALL patients (blue ones). Intuition tells us that it could be that these two patients, having gene expression profiles similar to those of ALL type, could in fact belong to ALL leukemia instead of AML. *Maybe the expert did not take into account this observation. Probably he considered these patients as AML patients while in fact the type of leukemia they were suffering was of ALL type.*

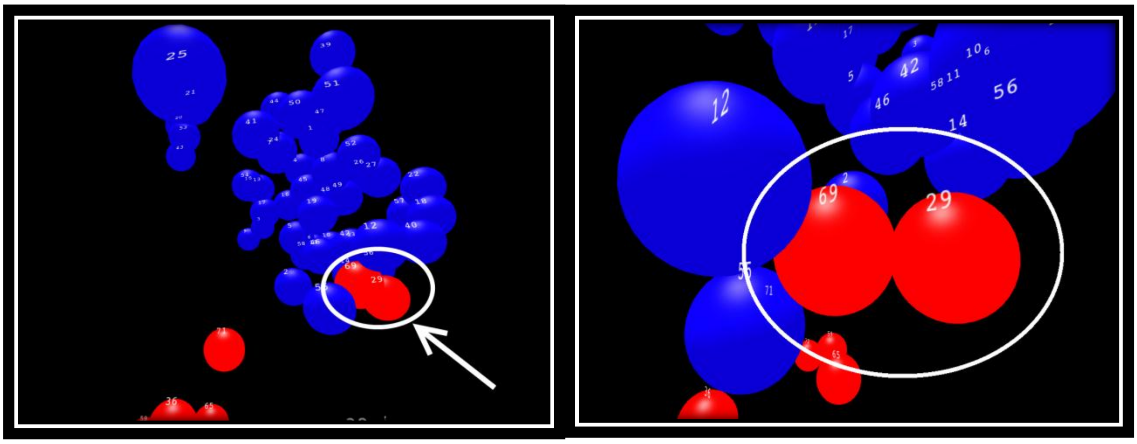


Figure 22. Red intruders found by MedVir. ID: patients number 29 and 69.

As said above, in many times we need to perform a further analysis in order to study this observation with detail. The DR process of a dataset of these characteristics (50 attributes) into a three-dimensional environment, always involves a minimal loss of information. This fact could be minimally causing distortions in the representation of the data and placing these two patients in these wrong positions.

So, we make use again of the M_{golub} and P_{golub} in order to verify if MedVir is properly representing the data.

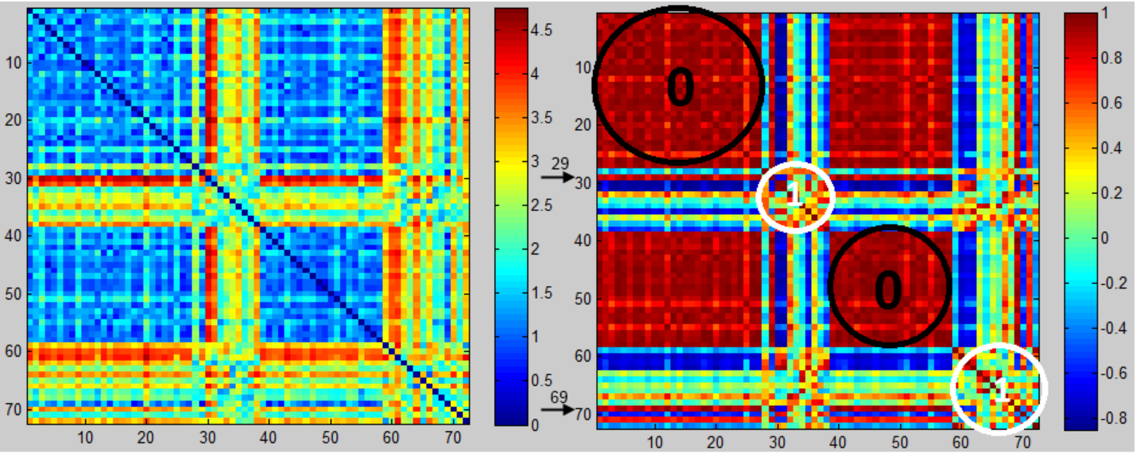


Figure 23. Red intruders in M_{golub} and P_{golub} matrices (marked by black arrows, Matlab). ID: patients number 29 and 69.

M_{golub} and P_{golub} matrices successfully confirm the hypothesis thrown by MedVir (Figure 23). In these matrices, the patients under analysis have been identified by black arrows (29 and 69). Observing the tonality distribution along the row it is obvious that both patients have a very similar nature to the rest of patients labeled as 0 (ALL). As it can be seen, the distance between the 'red intruders' and ALL patients is minimal. This observation could lead us to conclude with a very high degree of probability that the 'red intruders', actually, are patients suffering from ALL leukemia type that have been misdiagnosed. In this case, it would be appropriate to *report* this fact to the expert in order for him to check again these two patients.

In spite of everything, we need be aware of working with different set of attributes could lead to very different results. So, if we are working with the set of features selected by Golub et al, the conclusions we are drawing will be valid.

This is another clear example where MedVir can be *effectively* and *efficiently* used to detect possible misclassified instances and obtaining underlying knowledge in the data in a quickly and visual way. M_{golub} and P_{golub} matrices are used to corroborate the geometry preservation capabilities of MedVir when representing multidimensional data.

6 - Conclusions and futures lines

This research consists in the development of a tool for visualizing multidimensional biological data in a three-dimensional environment. The tool should allow, through interaction, the acquisition of underlying knowledge in the data.

We have explored the local and global geometry preservation of data when a DR process is carried out. Our research also indicated that MedVir, a new approach of DR based on heuristic optimization and visualization techniques, can obtain a very effective geometry preservation compared with the best known DR algorithms.

MedVir can visualize multidimensional data in 3D, so this allows conclusions to be obtained in a more simple, intuitive and quickly way. MedVir has been designed to provide a framework that makes easier the interaction of the expert with the data representation, for example, by asking for additional information about a instance, modifying the importance of an attribute from one specific study, removing the contribution of several attributes in the study, or even identifying a set of features that are supposed to have a great impact in the set of features.

In the case of DNA Microarray Data, the application of MedVir framework on a particular genetics study provides valuable information to the expert biologist. Therefore, MedVir is presented as a quick diagnostic tool in order to help experts find an interesting gene profiles. For example, once the patients have been represented in 3D according to their gene expression values, the classification of a new patient could be achieved by introducing its gene expression profile in MedVir and observing how its corresponding sphere is placed in the three dimensional space. The closer the new sphere is of a given existing cluster, the more likely that new patient will belong to that cluster.

Besides, the expert will be able to draw some conclusions based on their visual skills and previous experience on the domain. MedVir allows the visualization of the quality of the gene profile of a study, and infer knowledge related to the classification of new patients and misclassified patients. These misclassified patients could suggest the urgent need for a re-diagnosis process by the oncologist. MedVir also provides the possibility of identify the set of attributes that have the greatest influence on the gene expression profile, as well as certain gene expressions that do not provide information for the separation of the classes.

A specific sample belonging to a patient previously labeled as normal (absence of cancer), could suggests a carcinogenic intrinsic nature if that sample is very close spatially located to the rest of samples labeled as carcinogenic. Furthermore, MedVir could detect patients who have finished their treatment ahead of schedule or even detect non-interesting genes of a particular gene profile. It is also possible to visualize and analyze relationships between genes by means of the interaction of the expert with the axes in the representation. In addition, because the data used in this research are about the classification of patients based on their gene profiles, MedVir allows these gene profiles to be analyzed using the 3D representation of data colored according to their class.

However, the manipulation of data once they are represented in 3D can create some ambiguity in the interpretation and perception of the experts. It seems logical to think that an axis (gene expression) located far away from the origin of the coordinate system means that this has more influence on the representation of the samples. Translating this situation into the biological world, this gene expression has more importance in the study and should be selected in the gene profile. However, the movement of this axis in the 3D world, even, to place this axis close to other axes can create some confusion to experts. Thus, one of the future lines of this work is to create a set of rules that allows all of the possible interactions with the data to be better interpreted.

Other future lines of research will be oriented towards improving the MedVir framework in terms of interaction. On one hand, the interaction of the expert with the 3D visualizer using Kinect hardware in order to access to the data with a simple movement of a finger. In addition, several improvements in the usability and interaction between the expert and systems are necessary in order to be more useful and simple for the expert.

Regarding data mining techniques, supervised and unsupervised algorithms could improve the final analysis. For example, clustering validation techniques can advise experts to take into account or not certain groups of samples. Feature Subset Selection (FSS) could complete the entire MedVir framework in order to be able to work directly with thousands of features. Sometimes, the gene profile selection conducted by the author is not the best. Furthermore, a more complex analysis of multiobjective (local-global) optimization could be interesting to study.

Another important aspect is related to the interpretation of the lengths of the axes in MedVir. Different experiments are being carried out in order to obtain a final ranking of the attributes that have a major influence in the representation of the data geometry, as well as those who have no influence. This could give an explanation of the importance of each attribute in a given study. Thus, in a very near future, MedVir will perform an underlying Feature Subset Selection process in order to obtain a ranking of the most important attributes.

In addition, this work is under the framework of *Cajal Blue Brain* project, where data from morphological features of neurons are being used. A possibility is to visualize multidimensional neuronal datasets and obtain knowledge of these representations. For example, a FSS process could be performed in order to select a subset of attributes that better segment typologies of neurons, and then visualize the data in MedVir to classify or re-label different types of neurons. Moreover, the aim is to apply MedVir to other fields such as magnetoencephalography data or, in a global scale, any set of biological and medical data.

Another possibility of improving results could be the inclusion of a mechanism for integrating different sources, for example genomic and clinical data [62]. The final visualization could be enriched to a great extent.

7 - References

1. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armañanzas R, Santafe G, Perez A, Robles V: Machine learning in bioinformatics. Brief Bioinform. 2006, 86-112.

2. Lopez M, Mallorquin P, Vega M: Microarrays and DNA biochips. Genoma Spain 2002.

3. Lee JA, Verleysen M: Nonlinear dimensionality reduction. New York; London: Springer 2007.

4. Julio J. Valdés, Alan J. Barton, Robert Orchard: Genetic Programming for Exploring Medical Data using Visual Spaces. Genetic and Evolutionary Computation: Medical Applications 2011

5. Addison D: Intelligent Computing Techniques: A Review. Telos Pr 2004.

6. Jirapech-Umpai T, Aitken S: Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. BMC Bioinformatics 2005, 6:148.

7. Su Y, Murali TM, Pavlovic V, Schaffer M, Kasif S: RankGene: identification of diagnostic genes based on expression data. Bioinformatics 2003, (12):1578-1579.

8. Li T, Zhang C, Ogihara M: A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics (Oxford, England) 2004, (15):2429-2437.

9. Lee JM: Introduction to Topological Manifolds (Graduate Texts in Mathematics). Springer 2000.

10. Van der Maaten LJP, Postma EO, van den Herik HJ: Dimensionality Reduction: A Comparative Review 2007.

11. Boyd S, Vandenberghe L: Convex Optimization. Cambridge University Press 2004.

12. De Silva V, Tenenbaum JB: Global Versus Local Methods in Nonlinear Dimensionality Reduction. In Advances in Neural Information Processing Systems 15 2003:705-712.

13. Cox TF, Cox MAA: Multidimensional Scaling. London: Chapman & Hall, har/dis edition 1994.

14. Sammon JW: A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers 1969, (5).

15. Demartines P, Herault J: Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. IEEE Trans. Neural Netw. 1997, 148-154.

16. Tenenbaum J, Silva V, Langford J: A global geometric framework for nonlinear dimensionality reduction. Science 2000, 290(5500):2319-2323.

17. Estévez PA, Chong AM: Geodesic Nonlinear Mapping Using the Neural Gas Network. In IJCNN 2006:3287-3294.

18. Lee JA, Verleysen M: Nonlinear dimensionality reduction of data manifolds with essential loops. Neurocomputing 2005, 29-53.

19. Lee JA, Lendasse A, Verleysen M: Curvilinear Distance Analysis versus Isomap. In Proceedings of ESANN'2002, 10th European Symposium on Artificial Neural Networks 2000:185-192.

20. Lee JA, Lendasse A, Donckers N, Verleysen M: A robust non-linear projection method. In ESANN'00 2000:13-20.

21. Schölkopf B, Smola A, Müller KR: Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. 1998, (5):1299-1319.

22. Scölkopf B, Smola AJ, Müller KR: Kernel principal component analysis, Cambridge, MA, USA: MIT Press 1999 :327-352.

23. Ball WWR, Coxeter HSM: Mathematical Recreations and Essays, 13th ed. New York: Dover 1987.

24. Kohonen T, Schroeder MR, Huang TS (Eds): Self-Organizing Maps. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 3rd edition 2001.

25. Bishop CM, Williams CKI: GTM: The generative topographic mapping. Neural Computation 1998, 215-234.

26. Lee JA, Archambeau C, Verleysen M: Locally Linear Embedding versus Isotop. In ESANN 2003:527-534.

27. Jolliffe IT: Principal Component Analysis. Springer, 2nd edition 2002.

28. Hotelling H: Analysis of a complex of statistical variables into principal components. J. Educ. Psych. 1933, 24.

29. Roweis ST, Saul LK: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 2000, (5500):2323-2326.

30. Saul LK, Roweis ST: Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds. Journal of Machine Learning Research 2003,119-155.

31. Belkin M, Niyogi P: Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In Advances in Neural Information Processing Systems 2001:585-591.

32. Belkin M, Niyogi P: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation 2003, (6):1373-1396.

33. Nadler B, Lafon S, Coifman RR, Kevrekidis IG: Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. Applied and Computational Harmonic Analysis 2006, 113-127.

34. Lafon S, Lee AB: Diffusion Maps and Coarse-Graining: A Unified Framework for Dimensionality Reduction, Graph Partitioning, and Data Set Parameterization. IEEE Transactions on Pattern Analysis and Machine Intelligence 2006, (9):1393-1403.

35. Dijkstra EW: A Note on Two Problems in Connexion with Graphs. Numerische Mathematik 1959, 1:269-271.

36. Balasubramanian M, Schwartz EL: The Isomap Algorithm and Topological Stability. Science 2002, (5552):7.

37. L Chen AB: Local Multidimensional Scaling for Nonlinear Dimension Reduction, Graph Drawing, and Proximity Analysis. Journal of the American Statistical Association 2009.

38. Venna J, Kaski S: Local multidimensional scaling. Neural Networks 2006, (6-7):889-899.

39. Kaski S, Nikkila J, Oja M, Venna J, Toronen P, Castren E: Trustworthiness and metrics in visualizing similarity of gene expression. BMC Bioinformatics 2003, 4:48.

40. Lee JA, Verleysen M: Quality assessment of dimensionality reduction: Rank-based criteria. Neurocomputing 2009, 72(7-9):1431-1443.

41. Lee JA, Verleysen M: Rank-based quality assessment of nonlinear dimensionality reduction. In ESANN 2008:49-54.

42. Valdés JJ, Barton AJ: Hybrid unsupervised/supervised virtual reality spaces for visualizing cancer databases: an evolutionary computation approach. In Proceedings of the 9th international work conference on Artificial neural networks, IWANN'07 2007:1028-1035.

43. Kandogan E: Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In KDD '01: Proceedings of the seventh ACM SIGKDD international

conference on Knowledge discovery and data mining, New York, NY, USA: ACM 2001:107-116.

44. Gracia A, González S, Veiga J, Robles V: VR BioViewer - A new interactive-visual model to represent medical information. In MSV '11: Proceedings of the 2011 International Conference on Modeling, Simulation and Visualization Methods., Las Vegas, NV, USA 2011:40-46.

45. Storn R, Price K: Differential Evolution {A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. Journal of Global Optimization 1997, (4):341-359.

46. Muelas Pascual, Santiago (2011) Advances in Hybrid Evolutionary Computation *for Continuous Optimization.* Tesis (Doctoral), Facultad de Informática (UPM).

47. Unity Technologies: Unity3D 2009, [http://www.unity3d.com/].

48. Hardin J, Mitani A, Hicks L, VanKoten B: A robust measure of correlation between two genes on a microarray. BMC Bioinformatics 2007, 220.

49. Van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002, (6871):530-536.

50. Van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R: A gene-expression signature as a predictor of survival in breast cancer. The New England journal of medicine (25):1999-2009.

51. Stirewalt DL, Meshinchi S, Kopecky KJ, Fan W, Pogosova-Agadjanyan EL, Engel JH, Cronk MR, Dorcy KS, McQuary AR, Hockenbery D, Wood B, Shelly Heimfeld B, Radich JP: Identification of genes with abnormal expression changes in acute myeloid leukemia. Genes, Chromosomes and Cancer 2008, 47:8-20.

52. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999, (5439):531-537.

53. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR: Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 2002, (6870):436-442.

54. Knudsen S: A Biologist's Guide to Analysis of DNA Microarray Data. John Wiley & Sons 2002.

55. Quackenbush J: Computational analysis of microarray data. Nature Reviews Genetics 2001, (6):418-427.

56. Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, Davis RW: Microarrays: biotechnology's discovery platform for functional genomics. Trends in biotechnology 1998, (7):301-306.

57. Huber W, Heydebreck AV, Vingron M: Analysis of Microarray Gene Expression Data. In in 'Handbook of Statistical Genetics', 2nd edn, Wiley 2003.

58. Fernandez-Teijeiro A, Betensky R, Sturla L, Kim J, Tamayo P, Pomeroy S: Combining gene expression profiles and clinical parameters for risk stratification in medulloblastomas. J Clin Oncol. 2004, 22(6):994-8.

59. Huang CJ, Liao WC: Application of probabilistic neural networks to the class prediction of leukemia and embryonal tumor of central nervous system. Neural Processing Letters 2004, 19(3):211-226.

60. Daemen A, Gevaert O, De Moor B: Integration of clinical and microarray data with kernel methods. Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference 2007, :5411-5415.

61. Hollander M, Wolfe DA: Nonparametric Statistical Methods, 2nd Edition. Wiley-Interscience, 2 edition 1999.

62. González S, Guerra L, Robles V, Peña JM, Famili F: CliDaPa: A new approach to combining clinical data with DNA microarrays. Intell. Data Anal. 2010, 14:20-223.

63. Day, WHE (1986). "Computational complexity of inferring phylogenies from dissimilarity matrices". *Bulletin of Mathematical Biology* 49: 461–7.