VR BioViewer: A new interactive-visual model to represent medical information

Antonio Gracia, Santiago González, Jorge Veiga and Víctor Robles

Abstract — **Virtual reality (VR) techniques to understand and obtain conclusions of data in an easy way are being used by the scientific community. However, these techniques are not used frequently for analyzing large amounts of data in life sciences, particularly in genomics, due to the high complexity of data (curse of dimensionality). Nevertheless, new approaches that allow to bring out the real important data characteristics, arise the possibility of constructing VR spaces to visually understand the intrinsic nature of data. It is well known the benefits of representing high dimensional data in tridimensional spaces by means of dimensionality reduction and transformation techniques, complemented with a strong component of interaction methods. Thus, a novel framework, designed for helping to visualize and interact with data about diseases, is presented. In this paper, the framework is applied to the Van't Veer breast cancer dataset is used, while oncologists from La Paz Hospital (Madrid) are interacting with the obtained results. That is to say a first attempt to generate a visually tangible model of breast cancer disease in order to support the experience of oncologists is presented.**

Keywords: dimensionality reduction, visualization, bio-informatics, manifold learning, virtual reality spaces, cancer classification, visual data mining, genetic algorithms, similarity structure preservation, genomics.

I. Introduction

Nowadays, around 60 people die of diseases such as cancer every minute. The value is even more concerning if instead of thinking in minutes, we do it in hours or days. It is, therefore, a problem of high social impact that must be solved as quickly as possible. Finding a cure for diseases such as cancer would translate into a much higher life expectancy. In the scientific field, expert biologists are devoted to the study of possible solutions to these kinds of diseases. Among the many approaches, the DNA microarray technology will be the application field of this research.

Medical and biologist experts have a lot of genomic (based on DNA microarray) and clinical information of patients with any disease. Even more, they have available different published gene profiles (*biomarkers*) for diseases as Breast Cancer. However, for them is very difficult to analyze all this information and obtain any conclusion in a short time. Visualization techniques have been of great assistance to experts in different fields of research. These techniques allow to bring out the real important data characteristics, arising the possibility of constructing VR spaces to visually understand the intrinsic nature of data in a very short time. Although some techniques as PCA (Principal Component Analysis) [1] have been used in genomic field to see relationship between genes, these are not frequently used to obtain patterns and conclusions of any gene in any disease. That is because the representation in 3D transforms gene features in 3 new values (x_i,y_i,z_i), but this with the transformation we have lost its scientific significance and there is not any possibility of reverse the transformation.

However, trying to obtain a representation that allows us to see the behavior of all gene features as well as the possibility of real interaction with experts. A novel framework, called VR BioViewer, is presented in this paper, applying it to Van't Veer breast cancer dataset. With this, medical experts from La Paz Hospital (Madrid) can visualize patients as points in a VR space, gene features as different axis represented in 3D, and can interact with the model: moving any gene axis (f. e. two genes are related, thus their axis have to be closed), seeing gene information of other past researches (using literature), or clinical information of patients by clicking in any point.

The structure of the paper is as follows: Next section presents DNA microarray technology. Section 3 analyzes briefly a state of art about Visualization and the use of optimization in visualization. Section 4 describes the VR BioViewer framework. Finally, conclusions, future lines and acknowledgments are presented in the last sections.

II. DNA Microarray

DNA microarrays [2,3,4,5] are a relatively new and complex technology used in molecular biology and medicine. Microarrays present unique opportunities in analyzing gene expression and regulation in an overall cellular context. This technology has been applied in diverse areas ranging from genetic and drug discovery to disciplines such as virology, microbiology, immunology, endocrinology

Antonio, Santiago, Jorge and Victor are from the Department of Computer Architecture, Universidad Politécnica de Madrid in Spain. (emails: {agracia,sgonzalez,jveiga,vrobles}@laurel.datsi.fi.upm.es).

and neurobiology. Microarray technology is the most widely used technology for the large-scale analysis of gene expression because it provides a simultaneous study of thousands of genes by single experiment.

A DNA microarray consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides (shorts molecules consisting of several linked nucleotides, between 10 and 60, chained together and attached by covalent bonds), called Expressed Sequence Tags (ESTs), each containing several molecules of a specific DNA sequence. This can be a short section of a gene or other DNA element.

There are several biological steps [6, 7] in the design and implementation of a DNA microarray experiment: Probe, Chip Manufacture, Sample preparation, Assay (*Hibridization* [6]), Readout and Informatics.

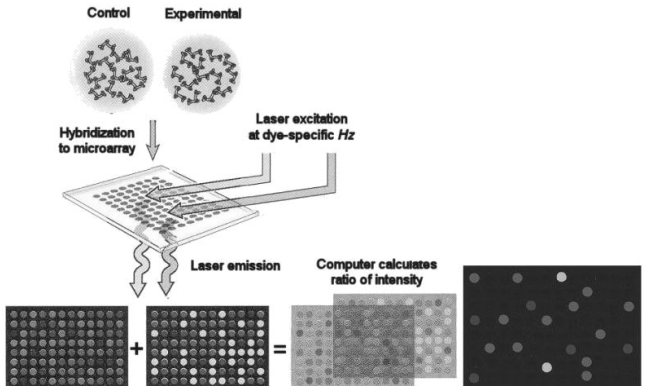

Fig. 1. Biological process of DNA Microarray analysis. Image from Gibson & Muse 2002

A. Breast Cancer dataset

The Van't Veer dataset [8] on Breast Cancer has been considered to use in this research. Van't Veer results have been approved by the FDA (Food and Drug Administration) and were applied in a genomic profiling test, called MammaPrint, that predicts whether patients will suffer breast cancer relapse or not. The data is divided into two groups, learning and validation instances. The training data consist of 78 patients, 34 of whom are patients that developed distance metastases within 5 years (poor prognosis). The rest of the dataset (44 patients) are the ones who remained healthy from the disease after their initial diagnosis for an interval of 5 years (good prognosis). The second group of patients (validation dataset) consists of 19 patients, 12 patients with poor prognosis and 7 with good prognosis. DNA microarray analysis was used to determine the mRNA expression levels of approximately 24,500 genes for each patient. All the tumors were hybridized against a reference pool made by pooling equal amounts of RNA from each patient.

- **Preprocessing:** Obviously real data have a lot of redundancy, as well as incorrect or missing values, depending on some factors. So, as first step, we carry out some pre-processing in order to clean up and prepare the data. We also discard variables with low internal variance or low Pearson correlation with outcome.

  Several pre-processing algorithms have been carried out through the training data. Firstly, we have discarded genes that are replicated. Next, we have discarded patients that had more than 80% of missing gene values. All data have been background corrected, normalized and log-transformed using Lowess Normalization [9]. Missing values were estimated using a 15-weighted nearest neighbor algorithm [10] (kNN Impute).

- **Biomarker selection:** The objective of this paper is not to obtain a feature selection of 24,500 genes expressions, but to create a model that represents this data efficiently. Thus, the microarray data is filtered to the 70 Van't Veer [8] selected genes (accepted by the FDA as breast cancer biomarkers), and these are the features we will use to represent the data in BioViewer.

III. Optimization and Visualization

The role of visualization techniques in the knowledge discovery process (KDD) [28] is well known. The increasing complexity of the data analysis procedures makes it more difficult for the user to extract useful information out of the results generated by the various techniques. So, a graphical representation is appealing from the point of view of the user. Besides, if this visualization is complemented with elements coming from a VR space, such as interaction or immersion in real time, we obtain a suitable framework for visualizing high dimensional data. The creation of VR spaces for visualizing high-dimensional dataset means a bigger insight of the underlying patterns or trends in data. To achieve that, we try to visually stimulate the human skills of understanding the intrinsic nature of data. The key is interaction. Interacting with 3D representations allow us to observe, for example, how clustered data is or the variation over time. These features make VR a much more intuitive environment than traditional ones for representing high dimensional data.

In many cases, the input data of these representations come directly from optimization techniques [27]. There are several examples that have been previously reported in [11,12,13]. Here, a multi-objective (MOO) optimization is used to the visualization of high dimensional datasets (i.e. leukemia or lung cancer). MOO optimization [26] studies

optimization problems involving more than one objective function and the goal is to find one or more optimal solutions. It is very common to use evolutionary algorithms for solving MOO problems (MOEA in general and MOGA if based on genetic algorithms). An evolutionary algorithm has four different steps: initialization, mutation, recombination and selection.

- **Initialization:** This initial phase focuses on select a population randomly.

- **Mutation:** the idea is to create a new individual v, for each one of the population x_i. To make it possible we need to select three different individuals of the population (r1, r2, r3). Where r_i has to be different and no one can be the parent. Consider that, F is a parameter in the interval [0, 2].

- **Recombination:** this operator generates the crossover between the element of the population x_i and the new individual generated, v. Implementation is made based on the exponential criteria.

- **Selection:** each element has a metric to evaluate how good or bad is the individual, this metric is called fitness. Comparing the new element generated v and the selected x_i, the algorithm select the one which has the best fitness.

Several multi-objective optimization algorithms inspired by this principles have been proposed. Among them, VEGA [14], HLGA [15], NSGA, NSGA-II [16,17,18], SPEA [19] and many others. In any case, a MOO optimization is out of scope in this study. Instead, an optimization technique involving two different available objective functions is carried out, particularly a differential evolution algorithm.

Handling large amounts of data arises a problem known as 'curse of dimensionality'. The high dimensional nature of genomic datasets makes difficult a straightforward analysis. So, it requires using several techniques for overcoming these problems. Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality. Ideally, the reduced representation has a dimensionality that corresponds to the intrinsic dimensionality of the data. The intrinsic dimensionality of data is the minimum number of parameters needed to account for the observed properties of the data. Dimensionality reduction is important in many fields, since it facilitates visualization, classification and compression of high-dimensional data, by mitigating the curse of dimensionality and other undesired properties of high-dimensional spaces [20]. Among this techniques, there are also another that accomplish a transformation of the involved features (transformation-based dimensionality reduction) i.e. 'Star Coordinates' algorithm [21]. It constructs a low dimensional space composed of a linear combination of the attributes.

IV. VR BioViewer

Here, VR BioViewer framework is presented. We divide it into two basic modules. The first one deals with the optimization part and the second one is for visualizing the results produced after the optimization. So, we are going to explain both methods in order to get a clear understanding of the pipeline of the process pipeline.

A. Optimization

Seeing the optimization module as a black box model (fig. 2), we can consider that the input is the n-dimensionality data and the output is the best distribution of the axes. That is, the distribution that better preserves the intrinsic geometry of n-dimensional data, after a representation using star coordinates algorithm.

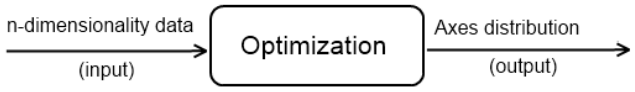

Fig.2. Optimization process of the axis

To set up the optimization module we implemented the differential evolution. The population of our model is the set of the different spatial distribution of the axes, and the initialization is done in a randomly way. So, the optimization tries to find the best configuration of the axis in order to preserve, as possible , the geometry of the n-dimensional data. Note that each axis corresponds to each one of the attributes in the input data, and has three components normalized in spherical coordinates (*t, p, r*) where *r* is the radius, *p* is the azimuthal angle in the interval [0, π] and *t* corresponds to the polar angle, interval [0, 2π]. In fact, we have to transform this problem to the cartesian coordinates to evaluate de distances. In the following formula, the transformation between spherical and cartesian coordinates is shown.

$$\begin{Bmatrix} x = r\sin(t)\cos(p) \\ y = r\sin(t)\sin(p) \\ z = r\cos(t) \end{Bmatrix}$$

At the beginning, the model has to calculate pair to pair distances of all the instances in the initial data. So this is stored in a matrix that 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.

In the study, the Manhattan distance metric ($L1$ norm) is considered because is consistently more preferable than the

Euclidean distance metric (*L*2 norm) for high dimensional data mining applications [22].

So, the aim is to preserve the distance of the data with high dimensionality and resemble the low dimensional data. As consequence of this, the model extracts the geometry of the initial data and this geometry is mapped into the low dimensionality data.

Before calculating the fitness, it is needed a representation of the data with the optimized axes, by means of the star coordinates algorithm. Then, the *generated distance matrix* is obtained and it represents the distances between all the instances of the data using these new optimizated axes. Needless to say, that this algorithm makes a linear combination of all the axes, or attributes, of the data.

Finally, the *generated matrix* and the *target matrix* are compared. To establish the comparison between two rows of those matrixes we implemented the *Pearson correlation*. To get a value of the fitness there are to methods: (1) based on the mean of the correlations and (2) based on a threshold. In the first option all the correlations between same rows in both matrixes are computed, and have a arithmetic mean of them. In the second one, however, all the correlations are obtained and for each one a threshold is evaluated to each of them, so then one vote is counted. Lastly, all the votes are added up and the result is divided between the number of instances.

B. Visualization

Once the optimization module has produced the results, the aim is just to obtain a 3D representation of the data in order to study in detail possible patterns, trends or outliers in the dataset as well as separation of classes, if it is possible. The input data of the visualization module consist of the set of optimized axes that will make possible a successful 3D embedding of the intrinsic geometric structure of the n-dimensional manifold. This results in a structure preservation with a minimum information loss, depending on the quality of the optimization process.

A visualization tool, 'Unity3D' [23], is used. Unity3D is a game engine designed for the creation of multiple 3D powerful interactive contents. The implemented visualization 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. The background of this dimensionality transformation is the star coordinates algorithm. The original algorithm 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. In this framework a 3D mapping is used, so it can be described as the next formula:

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

Where d_{ji} is the *j-th* data with the *i-th* value, min_i 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. Figure 3 illustrates an example of the final position of a data point in a 8-dimensional dataset (the example is 2D but easily extensible to 3D by dimensional analogy).

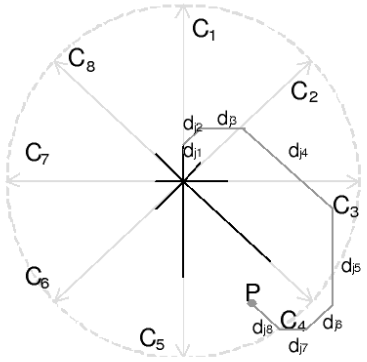

Fig. 3. Process of obtaining the final position of a data point for a 8-dimensional dataset (two-dimensional space).

Once the representation is done, the fundamental objective is that the expert oncologist classify, in a visually way, the breast dataset by a representation with colored spheres according to the class (black: relapse (R); gray: non-relapse (N.R) in breast cancer disease). It would be very useful if the oncologist could establish any kind of relations or connections between patients or genes, so a new mechanism of interaction based on the variation of the positions of the axes is provided. This variation will generate in real time a new spatial distribution of the spheres. Interacting with the coordinate axis could provide us valuable information. There might be many observations the expert might be interested in.

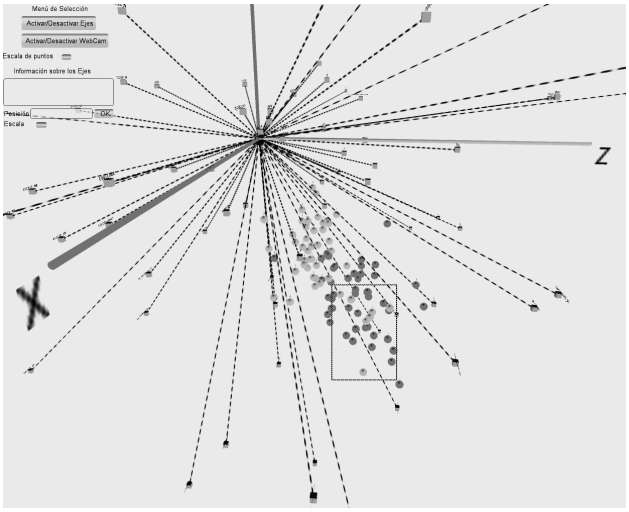

Fig. 4. Final visualization after the optimization process. Black spheres represent patients who suffered a relapse in breast cancer disease. Gray spheres represent healthy patients who do not relapse. Dotted lines represent coordinate axes (genes). A separation of classes is clearly visible. Last, the dotted square shows healthy patients in the relapse side.

The first one is related with clustering. Just looking at the distribution of data points. Playing with the coordinate axes, scaling or rotating. Finally one must see how spheres move in and out of clusters. The second one is correlation. The idea is the same, the oncologist could interact with the coordinate axes, multiple of them at the same time, scale a number of them at the same time, and observe how spheres move. He also has the possibility of turn on tracking so that consecutive movements of a data point are represented as lines so he can better observe where spheres go. Then he can examine, the direction points go, that tells you how the genes are correlated. It is important to mention that every operation above described must be supported by the expert's opinion in order to achieve valid and useful conclusions. Thus, a new tool that could make easier the acquisition of knowledge from medical data is provided.

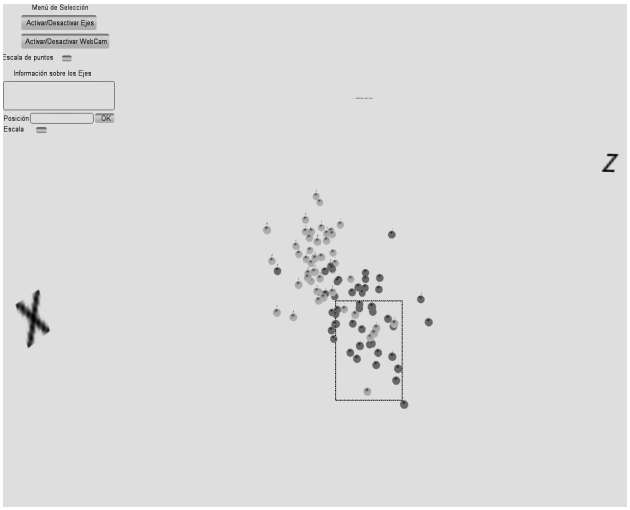

Fig. 5. Final visualization after the optimization process. No coordinate axes.

Fig. 4 and 5 shows a clear separation of the two classes. A possible interpretation is that in a n-dimensional space, the intrinsic nature of data makes evident a real division between patients who have relapsed and non-relapsed in breast cancer disease. After all, a three-dimensional embedding of the data's intrinsic geometric structure is mapped from the original n-dimensional space, what it means that the original geometry is being approximated.

If the results are analyzed with more detail, it is noticeable that several gray spheres lie on the black side. In fact, if the optimization process is continually repeated, in all cases the same representation is obtained. Moreover, always the same N.R patients lie on the black side. This fact could be exclusively seen from a medical point of view and it is possible that it has a major importance. For example, a priori patients with similar values for attributes must be grouped together but, why always the same N.R patients lie on the black side if the are supposed to be on the gray one? What does it means? Maybe, a valid interpretation could be that these patients consumed a certain kind of medicine during the chemotherapy process, so they didn't suffer a relapse. In this case, the next step could be to identify this medicine and the biomarkers, with the help of an expert.

Finally, the knowledge acquisition process is complemented with a strong interaction component. The expert also has the possibility of interacting with the application by means of rotating, scaling, or deleting axes in real time (fig.6). An optimal tradeoff between graphic quality and performance has been achieved. The expert can navigate and become absorbed into the environment thanks to the head tracking system. It works in the following way: just by using a standard webcam and FaceAPI software [24]. The visualization algorithm recognizes the data stream, previously filtered by FaceAPI software, and makes a mapping of the head movements to the main camera.

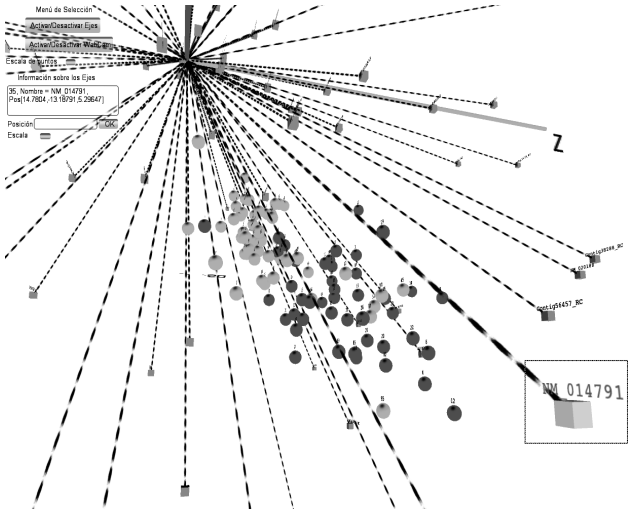

Fig. 6. Interacting with the coordinate axes. The dotted square shows a selected gen , 'NM_014791'.

C. Results

Here, final results have been separated into two categories depending on the objective function used in the optimization process. The first one represents the method based on the mean of the correlations, and the second one is based on the threshold method.

Table 1 shows the experimental settings in the optimization process.

Parameters	**Experimental values**
Number of generations	*15000*
Population size	*50*
Crossover probability	*0.7*
Mutation factor	*0.5*
Distance Metric	*Manhattan (L1 norm)*

Table 1. Experimental settings for computing the set of optimized axes.

Different parameters are shown. For example, the number of generations or the size of the initial population in differential evolution algorithm. We can also observe the crossover probability (C.P) and the mutation factor (M.F). Regarding to the distance metric, a Manhattan distance is used.

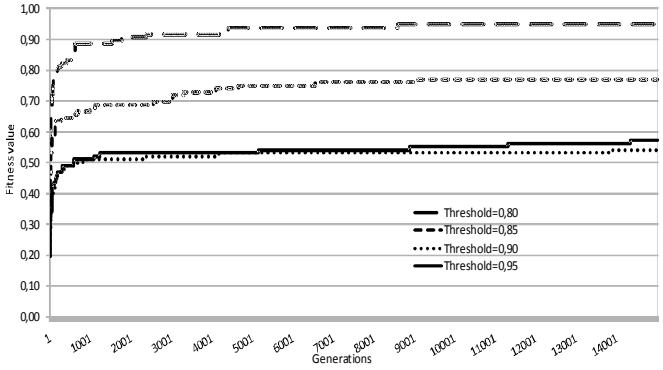

Fig. 7. Evolution of the fitness value using different threshold values. Threshold method.

Different fitness values are obtained according to the threshold value we use (fig. 7). For example, using a near 1 threshold will cause obtaining a small fitness value. For a 0.95 value, the explanation is that we are voting only if the correlation between the same row in *target matrix* and *generated matrix* is greater or equal than 0.95. The final fitness value for a 0.95 threshold is 0.572917, it means that the 57.2917% of individuals in the population (55 of 96 patients) are correlated each other with a threshold greater or equal than 0.95. Nevertheless, if a 0.80 threshold value is used, the 0.947917 fitness value is obtained, so 91 of 96 patients are correlated with a threshold greater or equal than 0.80. These high correlations between pair of distances in both matrixes (*target* and *generated* by optimization) show that it is necessary to find a tradeoff between the final fitness value and the geometry preservation in data visualization.

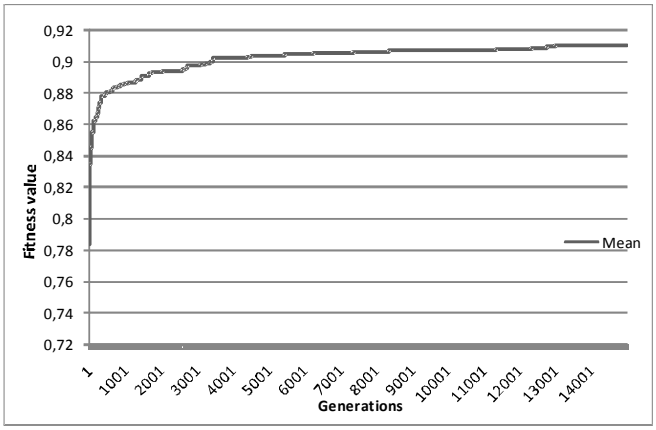

Fig. 8. Evolution of the fitness value for different threshold values. Method based on the mean of the correlations.

Figure 8 illustrates the method based on the mean of the correlations. It is simpler than the previous one. Here, the main benefit of this method is that a more uniform convergence is obtained. In consequence, the fitness value for every individual of the population is being fitted gradually. Instead, the drawback is that it penalizes high correlations and gives more importance to low correlations, because of the arithmetic mean.

V. Conclusions and Future Lines

This paper presented a new VR framework for visualizing and interacting with high dimensional DNA microarray datasets. The development is supported by an optimization process. It tries to find an optimal solution to preserving the intrinsic geometric structure of the n-dimensional manifold where the original dataset is embedded.

Taking into account the biological nature of data and the limitations for a computer engineer in terms of understanding and interpreting the data visualization, it is necessary to complement the final results with the experience of a specialist. For that reason, it is needed a close collaboration with the expert in order to validate appropriately the results.

Due to the medical nature of data and considering they are labeled, a direct approach based on supervised classification is carried out. The axes optimization process is done independently of the data class, so the same concept could be applied to different fields, for example in semi-supervised classification or unsupervised classification. In other words, the underlying idea is based on the possibility of extrapolating this development to any dataset that requires a visual and interactive representation.

It could be useful to use VR BioViewer to get better insight into prediction issues, particularly survival analysis in cancer. Using different data sources, for example the patient clinical information and the medicines used during the treatment, a complete visual study about who are alive for a given period could be done. Another possibility is to identify the biomarkers and add additional information about them, or giving them more intensity in the final representation.

Regarding to the distance metric used to get a suitable geometry preservation, it is considered that the use of geodesic distances over the neighborhood graph could improve the manifold learning [25]. It will be considered for next revisions.

These are preliminary results for which further experimentation would be required and they show a potential for advising and supporting the criterion of the oncologist expert.

Acknowledgements

The authors are grateful to the Blue Brain Project Team (http://cajalbbp.cesvima.upm.es/) and the La Paz Hospital, especially Cristobal Belda, for their assistance of medical knowledge. They also thankfully acknowledge the computer resources, technical expertise and assistance provided by the Centro de Supercomputación y Visualización de Madrid (CeSViMa) and the Spanish Supercomputing Network. It is also necessary to thank Ramón Toral from Operative Systems Laboratory, Universidad Politécnica de Madrid. This work is partially supported by the Madrid Regional Authority (Comunidad de Madrid) and the Universidad Rey Juan Carlos under the URJC-CM-2010-CET-5185 contract.

References

[1] Jolliffe, I. T. Principal Component Analysis. Springer-Verlag. pp. 487, 1986.

[2] S. Knudsen. A biologist's guide to Analysis of DNA microarray data. JohnWilley and Sons, 2002.

[3] J. Quackenbush. Computational analysis of microarray data. Nat Rev Genet, 6(2):418–427, June 2001.

[4] M. Schena, R. A. Heller, T.P. Theriault, K. Konrad, E. Lachenmeier, and R.W. Davis. Microarrays: biotechnology's discovery platform for functional genomics. Trends Biotechnol, 7(16):301–306, July 1998.

[5] Wolfgang Huber, Anja Von Heydebreck, and Martin Vingron. Analysis of microarray gene expression data. In in Handbook of Statistical Genetics, 2nd edn. Wiley, 2003.

[6] J. Quackenbush. Computational approaches to analysis of dna microarray data. Methods Inf Med, 45 Suppl 1:91–103, 2006.

[7] D. J. Lockhart and E. A. Winzeler. Genomics, gene expression and dna arrays. Nature, 405(6788):827–836, June 2000.

[8] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. art, M. Mao, H. L. Peterse,K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415(6871):530–536, January 2002.

[9] J. Quackenbush. Microarray data normalization and transformation - nature genetics.

[10] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. Bioinformatics, 17(6):520–525, June 2001.

[11] J. J. Valdés and A. J. Barton, "Virtual reality spaces for visual data mining with multiobjective evolutionary optimization: Implicit and explicit function representations mixing unsupervised and supervised properties," in IEEE Congress of Evolutionary Computation (CEC 2006). Vancouver: IEEE, July 16-21 2006, pp. 5592–5598.

[12] J. J. Valdés, A.J. Barton: Visualizing High Dimensional Objective Spaces for Multi-objective Optimization: A Virtual Reality Approach. Submitted to CEC 2007. Congress on Evolutionary Computation. 2007.

[13] J. J. Valdés, Alan J. Barton: Multiobjective evolutionary optimization for visual data mining with virtual reality spaces: application to Alzheimer gene expressions. GECCO 2006: 723-730.

[14] R. Schaffer, "Multiple objective optimization with vector evaluated genetic algorithms," in Proc. First International Conference on Genetic Algorithms, 1985, pp. 93–100.

[15] P. Hajela and C. Lin, "Genetic search strategies in multicriterion optimal design," Structural Optimization, vol. 4, pp. 99–107, 1992.

[16] N. Srinivas and K. Deb, "Multiobjective optimization using nondominated sorting in genetic algorithms," Evol. Comput., vol. 2, no. 3, pp. 221–248, 1994.

[17] K. Deb, S. Agarwal, A. Pratap, and T. Meyarivan, "A fast elitist nondominated sorting genetic algorithm for multi-objective optimization: Nsga-ii," in Proceedings of the Parallel Problem Solving from Nature VI Conference, Paris, France, 16-20 September 2000, pp. 849–858.

[18] K. Deb, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii." in IEEE Transaction on Evolutionary Computation, vol. 6 (2), 2002, pp. 181–197.

[19] E. Zitzler and L. Thiele, "Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach," IEEE Transactions on Evolutionary Computation, vol. 3, no. 4, pp. 257– 271, 1999.

[20] L.O. Jimenez and D.A. Landgrebe. Supervised classification in high-dimensional space: geometrical,statistical, and asymptotical properties of multivariate data. IEEE Transactions on Systems, Man and Cybernetics, 28(1):39 – 54, 1997.

[21] E. Kandogan, Visualizing Multi-dimensional Clusters, Trends, and Outliers Using Star Coordinates, KDD 2001, pp. 107-116, 2001.

[22] Aggarwal C., Hinneburg A., Keim D.A.: On the Surprising Behavior of Distance Metrics in High Dimensional Space,in Proc. of 8th International Conference on Database Theory, ICDT 2001, London, pp. 420-434.

[23] Reference Manual, 2009. Unity3D. [online] Available at: <http://www.unity3d.com> [Accessed 05 January 2011].

[24] FaceAPI software, 2010. FaceAPI. [online] Available at: <http://www.seeingmachines.com/product/faceapi/> [Accessed 15 January 2011].

[25] J. B. Tenenbaum, V. de Silva, J. C. Langford, A Global Geometric Framework for Nonlinear Dimensionality Reduction, Science 290, (2000), 2319–2323.

[26] C.A. Coello, D.A. Van Veldhuizen, G.B. Lamont, Evolutionary Algorithms for Solving Multi-Objective Problems. Kluwer Academic Publishers, 2002.

[27] J.A. Foster. Computational genetics: Evolutionary computation. Nature Reviews Genetics, 2:428–436, June 2001.

[28] Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann, 1st edition, September 2000.