Exploratory visualization of RFLP-PCR genomic data using Multidimensional Scaling

S. Espezua-Llerena, C. D. Maciel
School of Engineering of São Carlos, São Paulo University
Av. Trabalhador São Carlense, 400 São Carlos, SP, Brazil
Emails: sespezua@sel.eesc.usp.br, maciel@sel.eesc.usp.br

Abstract

Gel electrophoresis images are common results of biomolecular techniques such as RFLP-PCR (Restriction Fragment Length Polymorphism - Polymerase Chain Reaction). These images are used to discover the genetic relations between organisms. Find patterns in these images are a complex and delayed work if it is performed by humans. Traditionally, the analysis of gel electrophoresis images has been done by biologist using dendrogram representations aiming to capture the relations between organisms in a hierarchical organization. However this representation may be hard to analyze, especially when the information becomes large. This highlights the need to seek new ways of representing this type of data that become more intuitive. One of that methods is MDS (Multidimensional Scaling) which is a method used to transform measurements of similarity (or dissimilarity) between pairs of objects into points in a low-dimensional space, allowing visualize the data in a form that makes it easier to interpret. This paper proposes a new procedure to represent RFLP-PCR images as points in a low-dimensional space, based in a MDS technique. The procedure was applied in a genomic dataset obtained from a Brazilian collection of N_2-fixing bacterial strains belonging to the genus Bradyrhizobium. The results showed the efficacy of the procedure to represent the RFLP-PCR images, facilitating the identification of patterns in a more intuitive way than dendrogram's representations. Also, our procedure allows an appropriate integration with a pattern-recognition algorithm, taking advantages of the visual human skills and the computational power.

1. Introduction

Restriction Fragment Length Polymorphism - Polymerase Chain Reaction (RFLP-PCR) [18] is a useful technique in the biomolecular area. The most important applications of this technique are: genome mapping, localization of genetic disease genes, genetic fingerprinting, paternity testing and taxonomic identification among organisms. This technique is based in the fragmentation of genomic DNA by restriction enzymes, which cut DNA wherever a specific short sequence occurs. The resulting DNA fragments are then separated by length through a process known as gel electrophoresis, resulting in a image that contain a profile of *bands* that can be used in genetic analysis

Usually, this analysis is done by biologist in a visual way, but this can be a complex and delayed work, prone to subjectiveness of the human perceptions. To overcome these limitations, was proposed the use of hierarchical techniques [7] to represent the gel electrophoresis data in a hierarchical organization. This representation is commonly used by biologist to identify genetic relations between organisms. However there are several problems associated with this representation [14, 16], such as : i) isomorphism, when the comparison between branches is symmetric then the direct distance between nodes is not directly related to their true distance; ii) scalability, dendrograms do not scale up especially when the data size becomes large; iii) misclassification, it is difficult to find the level where the dendrogram must be cut; iv) dendrograms are restricted to represent hierarchical organized data, and are not suitable for analyzing data that is not inherently hierarchically organized. These limitations highlight the need of searching new ways to represent this type of data.

Multidimensional Scaling (MDS)[4] is a family of techniques aimed to transform (map) measurements of similarity (or dissimilarity) among pairs of objects into points in a low-dimensional space, with the objective to visualize and present the data in a way that makes it easier to interpret. In MDS, objects with high similarity are represented by close points on the target space[4]. This method enables the users to literally "look" at the data and to visually explore their relationships. The success of the MDS

techniques is due to its application in a wide range of problems. For example, it was applicated in phsychology[10], marketing[5], data mining[9], molecular modeling [17, 3], and others [14, 13, 2, 1]. Like the dendrogram representation, MDS is directly performed from the distance matrix, which can be calculated from different data types (vectors of characteristics, time sequences, images, etc.) using any distance metric.

In this paper, we propose a procedure based in a MDS technique to address the problem of representing RFLP-PCR data. We choose the *Landmark Multidimensional Scaling (LMDS)* technique [6, 12] to perform the mapping. This technique was chosen because its demonstrated computational efficiency and good mapping precision [6, 12]. The distance matrix is computed by extracting the electropherogram sequence of each image and comparing them with the Pearson's correlation index [7]. A visual environment tool was implemented to present the mapped data.

The proposed procedure was applied to the Bradyrhizobium dataset [8], which is a set of RFLP-PCR images obtained from a Brazilian collection of N_2-fixing bacterial strains belonging to the genus Bradyrhizobium. This bacterial are important in agriculture due to its capacity of transforming the nitrogen of the atmosphere (N_2) into plant usable compounds. This dataset was analyzed with the aim to map the images into points in a low-dimensional space, to identify the intrinsic dimensionality of the data, and to visualize and explore the mapped data trying to identify relevant groups.

The results showed the usefulness of the procedure, facilitating the visual identification of patterns. The low-dimensional mapping of the Bradyrhizobium dataset presented in the visual environment showed to be more intuitive than the dendrogram representations presented in [8, 11]. The procedure also allows a proper integration with a pattern-recognition algorithm, taking advantages of the human visual skills and the computational power.

The paper is organized as follows: Section 2.1 addresses a method for the pre-processing of the data. Section 2.2 introduces the mapping with the LMDS algorithm. Section 2.3 explains the visualization and clustering process. Section 3 discusses the experimental results obtained with the Bradyrhizobium dataset. Finally, in Section 4 some comments and conclusions are presented. Also we remark topics for future work.

2. Materials and Methods

2.1. Pre-Processing

Figure 1 shows an example of gel electrophoresis images obtained by the RFLP-PCR technique. This images are photographs taken after the electrophoresis process [18]. In this process DNA (or RNA) molecules are splitted into many fragments by the action of restriction enzymes. These fragments are placed into agarose gel where an electric field is applied. Fragments with similar molecular weight tend to move toward the positive pole at a same rate. After a period of time, the process is stopped and the gel shows a series of *bands* which are darker regions that represent fragments with similar size. The set of bands generated by a specific restriction enzyme in a DNA sample is called *lane*. The identification of the bands in the lanes is important because this information can be used to compare organisms. For example in Figure 1 the lane 1 and 4 have similar band pattern, which can suggests that the respective organism share common traits.

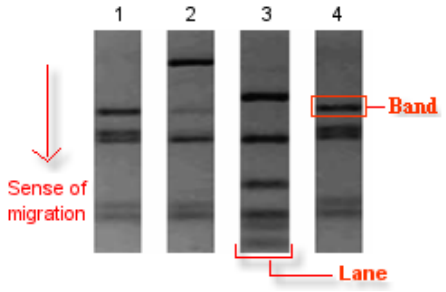

Figure 1. Example of gel electrophoresis images obtained by the RFLP-PCR technique. Each lane correspond to one organism.

Due the RFLP-PCR technique involves manual procedures, many problems are induced in the image lanes, such as: background noise, deformation in the bands, brightness variability, overlapping bands, etc. To tackle these problems we perform a pre-processing phase. First, each lane is separated in a individualized image file, then an electropherogram sequence is computed according to the following procedure:

1. The grayscale pixel matrix of the input lane $R = [r_{i,j}]$ is reduced to a sequence by averaging the columns of the matrix:
$$s[i] = \frac{1}{w} \sum_{j} r_{i,j} \quad (1)$$
where w is the number of pixel columns;
2. The resulting sequence is smoothed by a FIR low pass filter, getting the sequence $s' = Fir(s)$;
3. The sequence s' is shifted downwards in such a way that the relevant information (gel bands) is placed over the zero line, this is, $t[i] = s'[i] - th$, where th is the threshold to consider valid gel band and was empiri-

cally determined as: $th = \mu + 05 \times \sigma$, where μ and σ are the mean and standard deviation of s';

4. The electropherogram sequence e is obtained by zeroing the negative values of t and normalizing the positive values with respect to the maximum value, thus:

$$e[i] = \begin{cases} \frac{t[i]}{Max(t)} & , \quad t[i] > 0 \\ 0 & , \quad t[i] \leq 0 \end{cases} \tag{2}$$

The resulting sequence, called *electropherogram*, contains the useful information of the lane, but with noise reduced and bands easily identifiable. Figure 2 depicts an example of this pre-processing phase.

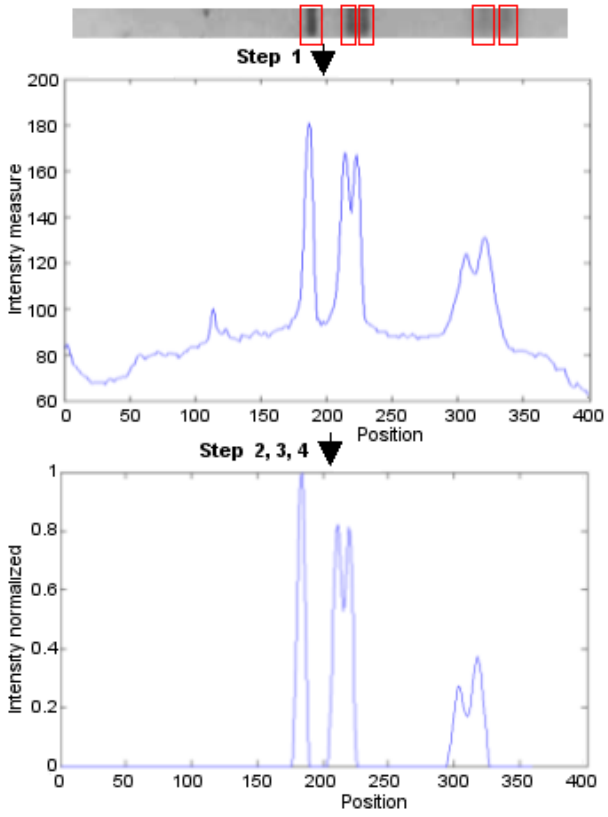

Figure 2. Example of an image pre-processing. The result is an eletropherogram sequence.

As the MDS algorithm needs the distance information to perform the mapping, we need to compute a distance matrix. The similarity between two lanes is calculated using their respective electropherogram sequences. We use the Pearson Coefficient Correlation [7] to measure the similarity between each pair of sequences. This coefficient measures the degree of linear relationship between two sequences, regardless the amplitude scale. It ranges from +1 to -1, where a correlation of +1 means a perfect positive linear relationship between sequences. For two electropherogram e_p and e_q the coefficient correlation is defined as:

$$r_{pq} = \frac{1}{L} \sum_{i=1}^{L} \left(\frac{e_p[i] - \bar{e}_p}{\sigma_p} \right) \left(\frac{e_q[i] - \bar{e}_q}{\sigma_q} \right) \tag{3}$$

where L is the electropherogram length; $\bar{e}_p$ and σ_p are the mean and standard deviation of e_p; $\bar{e}_q$ and σ_q are the mean and standard deviation of e_q.

For a set of n lanes, the distance matrix $\Delta = (\delta_{ij})$ is a $n \times n$ symmetric matrix, each element δ_{ij} representing the dissimilarity between two lanes i and j, calculated as:

$$\delta_{ij} = 1 - r_{ij} \tag{4}$$

Figure 3 shows the process of constructing the distance matrix as described in the previous procedure.

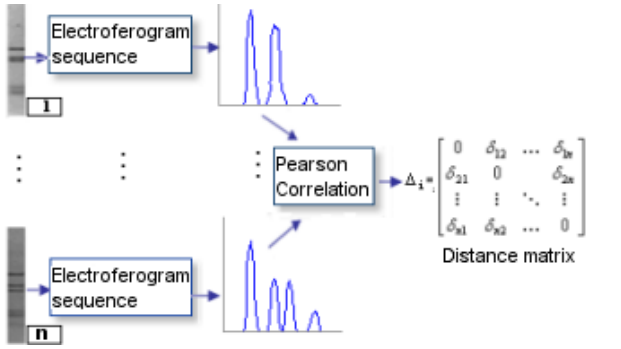

Figure 3. Construction of the distance matrix.

2.2. Mapping with LDMS

In this paper we use the LMDS algorithm [6] to represent the distance information and to determine the relevant dimensionality of the dataset. The choice of this algorithm is justified by its good performance in many datasets [12]. The first step of LMDS is to run the classical Multidimensional Scaling algorithm (CMDS) [4] to map a subset of n choosen points of the dataset, referred as *landmark points* L, whose distance matrix is Δ_n. The second step is calculate the distance-based triangulation procedure, which uses the distances of the already-embedded landmark points to determine where the remaining points should be placed. Finally is carried out an analysis of coordinates by eigenvalues decomposition in the resulting point dataset. LMDS is summarized in Algorithm 1. In this algorithm, $MaxMin()$ is the procedure to select the *landmark points* [6], u is the

mean value of the matrix solution generated by LMDS , U are the eigenvectors, A are the eigenvalues. The possibly $n-1$ dimensional solution $F = U^T \times X$ can be reduced to an approximate 2D or 3D solution by selecting the 2 or 3 first dimensions, respectively.

Algorithm 1 LMDS

Require: $\Delta \leftarrow$ squared distance matrix
$n \leftarrow$ desired number of landmark points
$k \leftarrow$ desired number of output dimensions
Ensure: F
$F \leftarrow$ matrix solution LMDS
1: $N \leftarrow rows(\Delta)$ {data size}
Select landmark points (Use MaxMin)
2: $L \leftarrow MaxMin(N, \Delta, n, 1)$
Calculate the submatrix Landmark Δ_n
3: $\Delta_n \leftarrow D(L, L)$
Call CMDS with Δ_n
4: $[L, A] \leftarrow$ CMDS (Δ_n)
5: $ki \leftarrow min(k, size(A))$ {dimensions for set out}
Calculate the distance-based triangulation
6: $un \leftarrow mean(\Delta_n)$
7: $sqrA \leftarrow \sqrt{A(1:ki)}$
8: $Li \leftarrow \frac{V(:,1:ki)}{ones(n,1) \times sqrA^T}$
9: $F \leftarrow \frac{Li^T(\Delta_n) \times ones(1,N)}{2}$
Calculate the eigendecomposition
10: $u \leftarrow mean(F^T)$
11: $X \leftarrow F - u(:) \times ones(1, N)$
12: $[U, A] \leftarrow eigen(X \times X^T)$
13: $F \leftarrow U^T \times X$

The mapping with LMDS is not perfect and always implies certain distortion between the original distance matrix and the distance matrix obtained from the mapped points. Usually this distortion is measured through a fitness index called *stress*. This index evaluates the quality of the mapping. In this paper we use two stress functions known as *Kruskal's stress* [4], defined as:

$$S1 = \sqrt{\frac{\sum_{i<j}[\delta_{ij} - d_{ij}(X)]^2}{\sum_{i<j} d_{ij}(X)}} \quad (5)$$

$$S2 = \sqrt{\frac{\sum_{i<j}[\delta_{ij} - d_{ij}(X)]^2}{\sum_{i<j} [d_{ij}(X) - \bar{d}]^2}} \quad (6)$$

where $\bar{d}$ is the mean measure of distances; d_{ij} is the distance computed from the mapped points.

2.3. Visualization

To visualize the resulting LMDS mapping, it was implemented a tool in C language using the *Visualization Tool Kit (VTK)* open source library [15]. This library was chosen because its wide variety of functions that make suitable for interactive exploration, taking advantage of the human abilities to explore data in 3D environments. The visual environment assigns colors to the mapped points according to the RGB color space. This environment, join to the addition of xyz-planes, can help the users to identify groups in the data. An example of this environment is showed in the Figure 4.

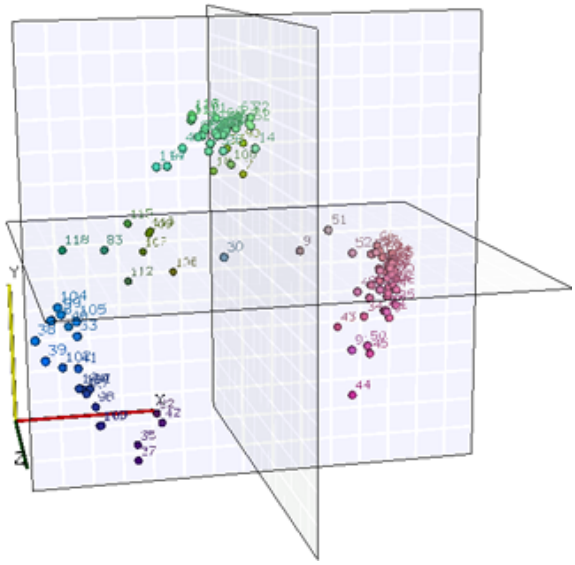

Figure 4. 3D visualization environment in VTK

The visual environment is useful to identify roughly the groups in the data, but this may not be enough when a more precision is required. To overcome this limitation we propose the following procedure: first, let the user to identify visually the number of groups, and second, supply this number of groups to a clustering method and let it to discover these groups. We suggest the use of the K-means [7] algorithm due to its simplicity and computational efficiency.

3. Results

We present here a case of study showing the application of the proposed procedure. The dataset analyzed is formed by RFLP-PCR images corresponding to a Brazilian collection of N_2-fixing bacterial strains belonging to the genus Bradyrhizobium. This symbiotic bacteria is important in agriculture by its capacity to transform the nitrogen of the atmosphere (N_2) into plant usable forms. A detailed information of this dataset can be found in [8].

The procedure of Section 2 was applied in the Bradyrhizobium dataset with the following objectives: map the im-

ages into points in a low-dimensional space, identify the intrinsic dimensionality of the data, visualize and explore the mapped data trying to identify relevant groups.

The dataset is formed by 119 bacterial strains identified by a sequential number. Each strain is described by three lanes, which correspond to the RFLP-PCR analysis of the ribosomal region 16S with three different restriction enzymes: Cfo I, Dde I and Msp I. The lanes were preprocessed according to the procedure 2.1. Then, three distance matrices $D1$, $D2$, $D3$ were computed, one for each restriction enzyme, as is depicted in the Figure 5.

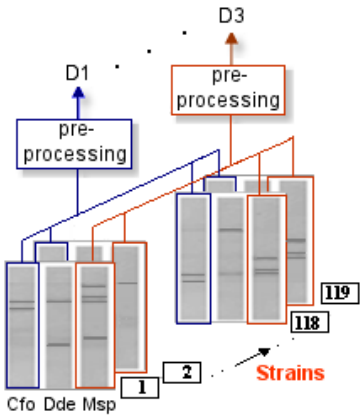

Figure 5. Pre-processing of the Bradyrhizobium dataset. The result are the distance matrices D1, D2 and D3.

The resulting distance matrices were used as inputs for the LMDS algorithm. The number of landmark points n was set to 12 points (10% of the data size). The desired dimension k entered to the LMDS algorithm was set according to the number of positive eigenvalues resulting of mapping the landmark points with CMDS. These dimensions were 7, 6 and 8 for the distance matrices $D1$, $D2$ and $D3$ respectively. Figure 6 shows for each mapping the resulting stress indexes as a function of the dimensionality. To construct these stress curves, it was incrementally varied the coordinate number of the LMDS mapping (from 1 dimension to the mapping dimension) and computed the correspondent Euclidean distance matrices, which were compared with the original dissimilarity matrix according to equations 5 and 6. It can be observed that steady values are reached at 5 dimensions, independent of the restriction enzyme and the stress index. This means that taking more than 5 dimensions does not give major information of the data, and hence, this dimension can be considered the intrinsic dimensionality.

Figure 7 shows a 3D visualization of the LMDS mappings for each distance matrix, considering the three first

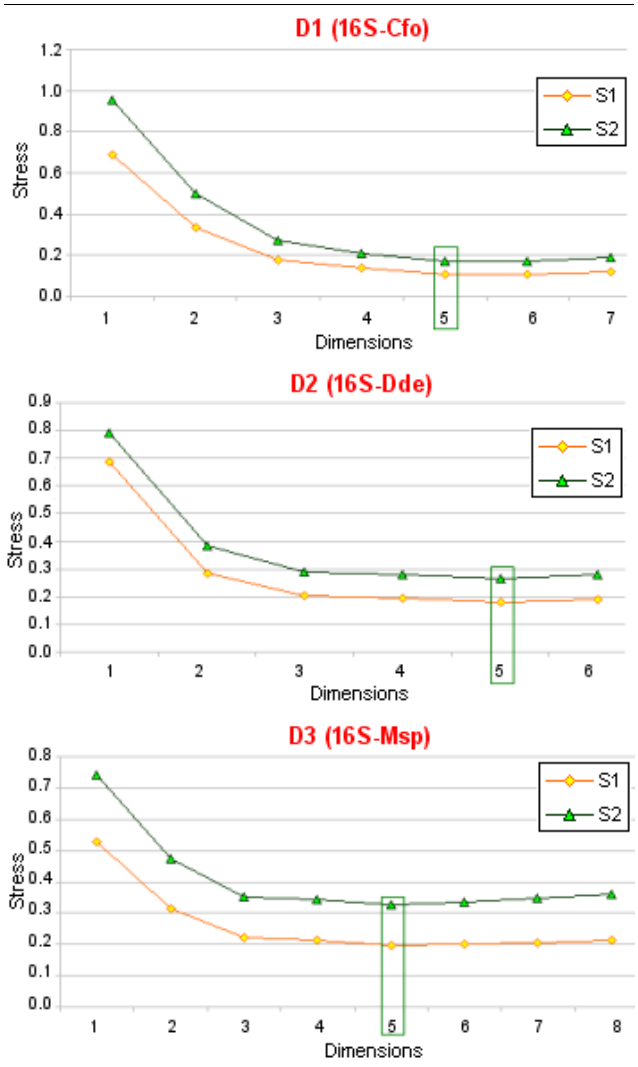

Figure 6. Stress values resulting of mapping the D1,D2 and D3 distance matrices.

dimensions. The quantity of information exhibited in these representations is calculated as the percentages that the 3 higher eigenvalues represent on the total sum of eigenvalues of the covariance matrix, calculated for each mapping. The quantity of information represented in Figure 7 are respectively 73%, 78% and 78% for $D1$, $D2$ and $D3$ mappings. After a visual exploration, we identified 4 groups for the $D1$ and $D3$ mappings and 3 groups for the $D2$ mapping, which are circled in Figure 7. These numbers of groups were introduced to the K-Means algorithm, which grouped the data using all the mapped dimensions. The resulting classification is exhibited in Figure 8 where is possible to observe a great similarity with the visual classification. This fact remarks the utility of the low-dimension representation and the colored visual environment, which allow an easy identi-

fication of patterns, agreeing with the results that K-Means performs in the whole dimensional space.

For reasons of comparison we show in Figure 9 a dendrogram representation for the distance matrix D1. This was created using the Unweighted Pair Group Method with Aritmetic Means (UPGMA) [7]. Note that this representation is more difficult to understand and get knowledge than the MDS mappings. As we can observe, the dendogram does not scale up for large amounts of data. Besides, it is dificult to find the level to cut the tree and get the "natural" partition of the data.

4. Conclusions

In this paper we propose a procedure to map RFLP-PCR images in a low-dimensional space aiming to be visually represented. The application of this procedure in the Bradyrhizobium dataset showed its usefulness to facilitate the identification of patterns in the data. The low-dimensional mapping obtained by the LMDS algorithm and represented in the visual environment showed to be more intuitive than the dendrogram representations found in [8, 11]. Also, the procedure allowed us to integrate the visual exploration with a pattern-recognition algorithm, taking advantage of the human visual skills and the computational power and precision. With this integration we identified between 3 and 4 groups in the analyzed dataset, which are similar with the results found in the literature using dendogram representations. By other hand, the stress indexes can be used to determine the intrinsic dimensionality of the data, which indicates the enough quantity of dimensions of the resulting mapping to be considered when a pattern-recognition algorithm is used. In the next future we will apply the proposed procedure in more datasets to compare and to validate the found patterns from a biological point of view.

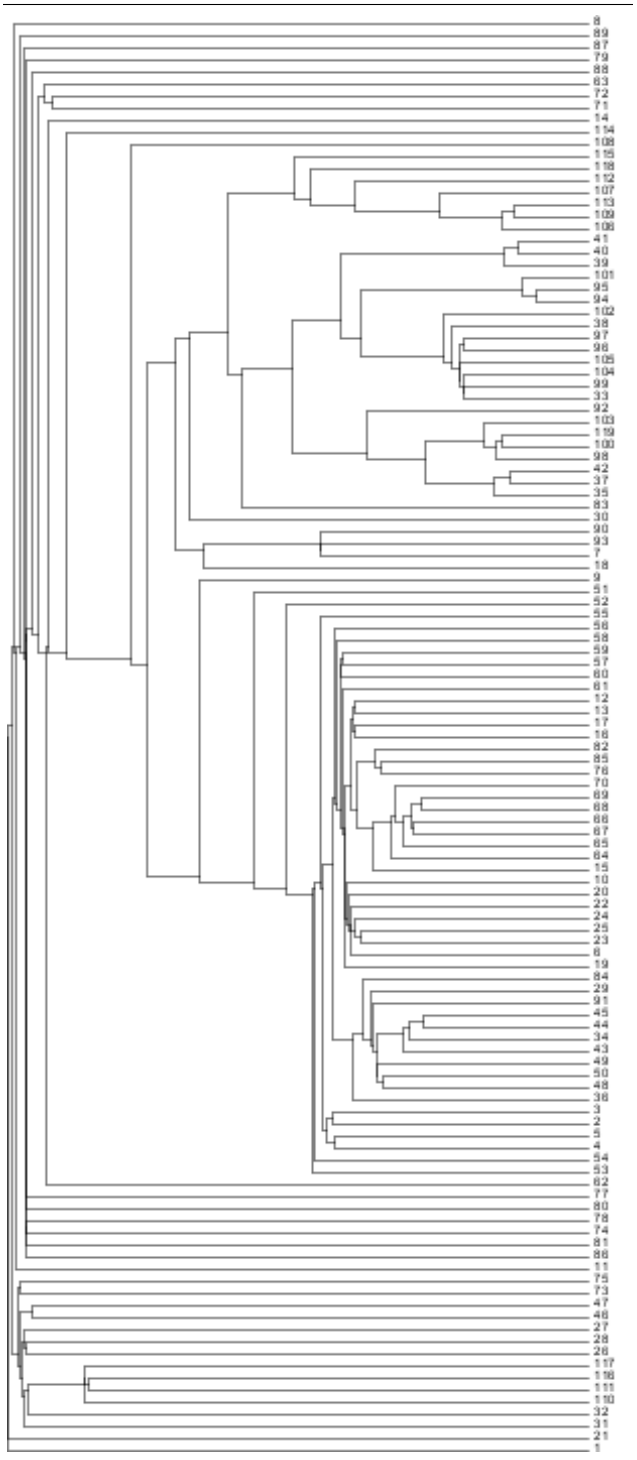

Figure 9. Dendrogram representation for distance matrix D1.

Acknowledgements

The authors acknowledge the School of Engineering of São Carlos (EESC/USP) for the research facilities.

References

[1] H. Abdi, D. Valentin, A. J. O'Toole, S. Chollet, and C. Chrea. Analyzing assesors and products in sorting task: Distatis, theory and applications. *Food Quality and Preference*, 18(4):627–640, 2007.

[2] H. Abdi, D. Valentin, A. J. O'Toole, and B. Edelman. Distatis: The analysis of multiple distance matrices. In *Conference on Computer Vision and Pattern Recognition*, volume 3, pages 42–47. IEEE Computer Society, 2005.

[3] D. Agrafiotis, D. Rassokhin, and V. Lobanov. Multidimensional scaling and visualization of large molecular similarity tables. *Journal Of Computational Chemistry*, 22(5):488–500, April 2001.

[4] I. Borg and P. Groenen. *Modern Multidimensional Scal-*

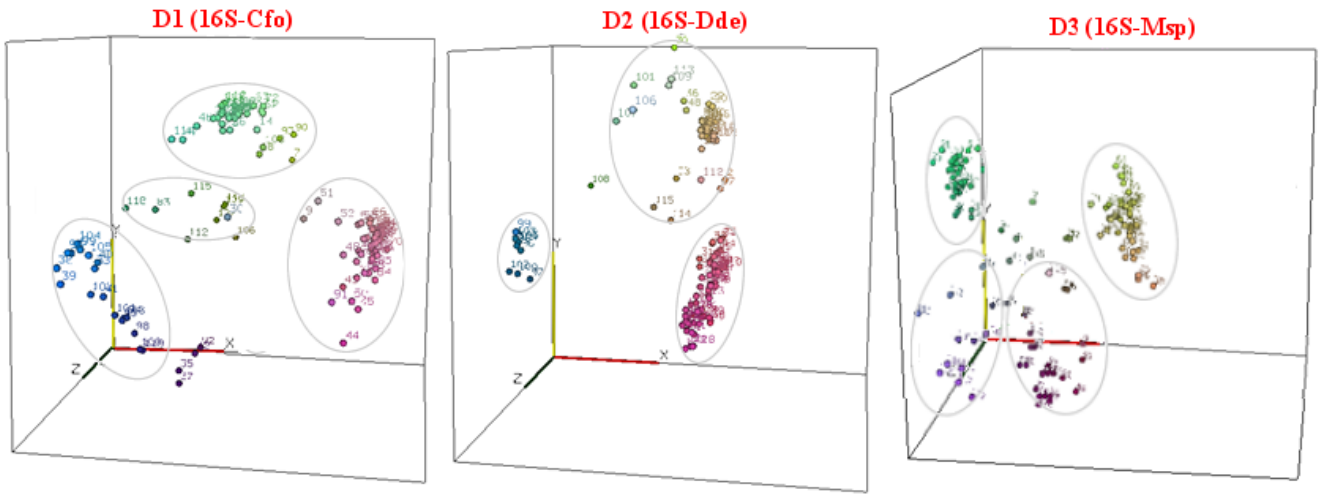

Figure 7. 3D representation of the LMDS mapping performed in the Bradyrhizobium dataset. These representations contain respectively 73%, 78% and 78% of total information for the D1, D2 and D3 mappings. The circles indicate the groups visually identified.

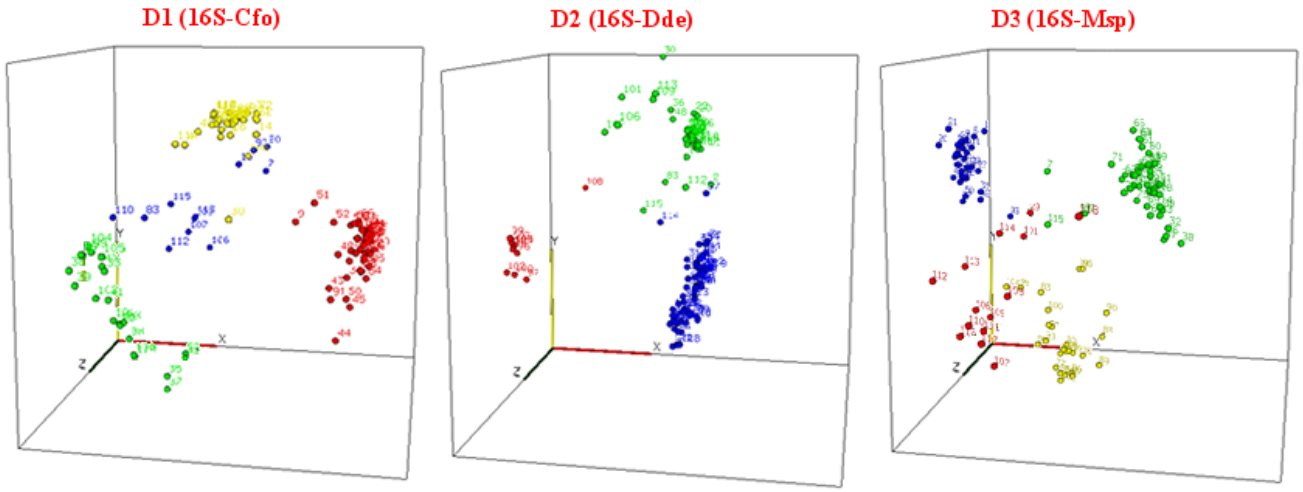

Figure 8. Clustering performed on the LMDS mappings with K-Means algorithm. The number of clusters supplied to K-Means was the visually found in Figure 7. All the dimensions were used to perform the clustering.

ing: Theory and Applications. Springer Press, second edition, 2005.

[5] P. Carrol, D. J.and Green. Psychometric methods in marketing research: Part ii, multidimensional scaling. *Journal of Marketing Research*, 34(2):193–204, 1997.

[6] V. de Silva and J. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction, 2003.

[7] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. A Wiley-Interscience, second edition, 2001.

[8] P. Germano, M. G.and Menna, F. L. Mostasso, and M. Hungria. Rflp analysis of the rna operon of a brazilian collection of bradyrhizobial strains from 33 legume species. *International Journal of Systematic and Evolutionary Microbiology*, 56(1):217?229, 2006.

[9] S. Huang and E. Ward M.and Rundensteiner. Exploration of dimensionality reduction for text visualization. In *Third International Conference on Coordinated and Multiple Views in Exploratory Visualization*, volume 10.1109/CMV, pages

63–74, July 2005.

[10] C. Izmailov, E. N. Skolova, and S. Korshunova. Multidimensional scaling reliability in similarity judgments about environmental sentences. *The Spanish Journal of Psychology*, 8(2):119–133, 2005.

[11] S. T. Milagre. Análise de estabilidade de cluster em uma coleção brasileira de bactérias siazotróficas do bradyrhizobium. Master's thesis, Engenharia Elétrica da Universidade Estadual de Londrina, 2003.

[12] J. C. Platt. Fastmap, metricmap, and landmark mds are all nystrom algorithms. In *10th International Workshop on Artificial Intelligence and Statistics*, pages 261–268, 2005.

[13] J. Priego. A vector space model as a methodological approach to the triple helix dimensionality: A comparative study of biology and biomedicine centres of two european national research councils from a webometric view. *Journal of Scientometrics*, 58(2):429–443, October 2003.

[14] M. Schroeder, D. Gilbert, J. Van Helden, and P. Noy. Approaches to visualization in bioinformatics: from dendrograms to space explorer. *Information Sciences*, 139(1-2):19–57, 2001.

[15] W. J. Schroeder, K. Martim, and B. Lorensen. *The Visualization Toolkit - An Object-Oriented Approach to 3D Graphics*. Prentice-Hall, third edition, 2002.

[16] M. Su and C. Chou. A k-means algorithm with a novel nonmetric distance. In *5th Joint Conference on Information Sciences (JCIS 2000)*, volume 1-2, pages 417–420, 2000.

[17] M. Venkatarajan and W. Braun. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. *Journal Of Molecular Modeling*, 7:445–453, 2001.

[18] A. Zaha. *Biologia Molecular Básica*. Mercado Aberto, third edition, 2003.