PERFORMANCE ANALYSIS OF RESTRICTION ENZYMES BASED IN RESULTS OF CLUSTER STABILITY

ANÁLISE DE DESEMPENHO DE ENZIMAS DE RESTRIÇÃO BASEADO NOS RESULTADOS DE ESTABILIDADE DE *CLUSTER*

Carlos Dias Maciel[1] e Selma Terezinha Milagre[2]

[1]Universidade de São Paulo, Departamento de Engenharia Elétrica
Av. Trabalhador São-Carlense, 400
CEP: 13566-590 São Carlos, SP Brasil
E-mail: maciel@sel.eesc.usp.br
[2]Universidade Federal de Goiás, Departamento de Ciências da Computação
Av. Dr. Lamartine Pinto de Avelar, 1120
CEP: 75700-000 Catalão, GO Brasil
E-mail: selma@catalao.ufg.br

ABSTRACT

This paper presents an application of a method aiming to compare a performance of the restriction enzymes with the results obtained by the analysis from cluster stability within a Brazilian collection of 119 *Bradyrhizobium* strains. The stability has been studied as a combination of six restriction enzymes used in the RFLP-PCR analysis and three ribosomal regions using three restriction enzymes per region, each combination forms a pair, thus there are nine pairs: pair 1 (*Cfo* I 16S), pair 2 (*Dde* I 16S), pair 3 (*Dde* I IGS), pair 4 (*Hae* III IGS), pair 5 (*Hae* III 23S), pair 6 (*Hha* I 23S), pair 7 (*Hinf* I 23S), pair 8 (*Msp* I 16S), pair 9 (*Msp* I IGS). The analysis of cluster stability is a way to validate the partitioning of data encountered through any conventional clustering algorithms. The aim is to compare a reference cluster obtained from all of samples with several clusters from subsamples of original dataset. For this study, the sampling ratio was 0.8 and 25 datasets were made from subsamples. The similarity was calculated between pairs of samples of the data and the stability was computed using the whole collection of similarities. For the system analyzed, were generated 511 experiments (all combinations from 1 up to 9 pairs = 9!-1 pairs) and the number of possible clusters varied from 2 to 10. The results indicated that groupings up to 7 clusters were sufficient to achieve the stable result with a reduction of time expense in simulations. The pairs 1 and 2 increased the similarities of the clustering process. The pair 3 increased the similarities of experiments when associated with pairs 7 and 8, and decreased the similarities of the experiments with pair 4. Pair 4 decreased the similarities of the experiments.
Keywords: restriction enzymes, cluster stability, *Bradyrhizobium.*

RESUMO

Este artigo apresenta a aplicação de um método objetivando comparar o desempenho de enzimas de restrição com os resultados obtidos pela análise de estabilidade de *cluster* em uma coleção brasileira de 119 estirpes de *Bradyrhizobium.* A estabilidade foi estudada como uma combinação de seis enzimas de restrição usadas na análise RFLP-PCR e três regiões ribossomais utilizando três enzimas de restrição por região, cada combinação forma um par, então existem nove pares: par 1 (*Cfo* I 16S), par 2 (*Dde* I 16S), par 3 (*Dde* I IGS), par 4 (*Hae* III IGS), par 5 (*Hae* III 23S), par 6 (*Hha* I 23S), par 7 (*Hinf* I 23S), par 8 (*Msp* I 16S), par 9 (*Msp* I IGS). A análise de estabilidade de *cluster* é uma forma de validar o particionamento dos dados encontrados por meio de qualquer algoritmo de agrupamento convencional. O objetivo é comparar um *cluster* de referência obtido a partir de todas as amostras com diferentes *clusters* de subamostras do conjunto original de dados. Para este estudo, a taxa de amostragem foi 0,8 e 25 subamostras. A similaridade foi calculada entre pares de amostras dos dados e a estabilidade foi calculada usando a coleção completa de similaridades. Par o sistema analisado, foram gerados 511 experimentos (todas as combinações de 1 até 9 pares=9!-1 par) e o número de possíveis *clusters* variou de 2 a 10. Os resultados indicaram que agrupamentos até 7 *clusters* foram suficientes para alcançar o resultado estável com redução de tempo gasto nas simulações. Os pares 1 e 2 aumentaram as similaridades dos processos de agrupamento. O par 3 aumentou as similaridades dos experimentos quando associados com os pares 7 e 8, e diminuiu as similaridades dos experimentos com o par 4. O par 4 diminuiu as similaridades dos experimentos.
Palavras-chave: enzimas de restrição, estabilidade de *cluster*, *Bradyrhizobium.*

1 – INTRODUCTION

The rapid accumulation of data regarding DNA chains since the 1970s had a great impact upon phylogenies. The advent of various molecular techniques, in particular the polymeric chain reaction (PCR), resulted in both,

accumulation of data in DNA sequences and in an unprecedented level of activities in molecular phylogenies. DNA and protein sequences are strictly hereditary entities and, for various reasons, molecular data are frequently more adaptable than morphological data to quantitative treatments [10].

An analysis by DNA sequences can be expensive and complex, but there are other methods cheaper to analyze ribosomal genes that can be used as an initial method for assessment of diversity and taxonomic position. It has been shown that amplification of DNA regions coding for ribosomal genes by PCR (polymerase chain reaction) technique, followed by digestion with restriction enzymes [RFLP (restriction fragment length polymorphism)-PCR technique] correlates quite well with an analysis of the sequences of these genes [3, 12, 14, 21, 23]. Restriction enzymes are proteins, isolated from bacteria that cut the DNA into smaller fragments. The number of fragments produced is established through the number of restrictions recognized by the enzymes used.

The fragments obtained can be separated into respective size groups using a technique of gel electrophoresis. The biggest ones are heavier and take longer to be displaced through the gel, a step that a chain with smaller pieces are lighter and are moved more rapidly through the gel. The analysis of electrophoretic patterns, produced by the RFLP-PCR method is critical for the correct identification of clusters and species. Therefore, an analysis of this technique needs to be reliable, reproducible and not susceptible to individual interpretation.

Exploratory procedures are, often, useful in understanding of the complex nature of relationships in multivariate dataset. Groupings can supply an informal way to access the dimensionality, indicating trends and suggesting hypotheses relative to relationships. This process is based upon the similarities or dissimilarities among data and the goal is the partitioning of the elements into subsets, which are called clusters. In clustering no information about classes is assumed except the number of clusters [22]. A different clustering process may result in different partitions of data sets, depending on a specific criterion used in clustering. One of the most important issues in cluster analysis is the evaluation of clustering results to find the partitioning that best fits the underlying data [4, 24].

An analysis of cluster stability is a way to validate the partitioning of data encountered through clustering algorithms [11, 26, 29]. There is no still no consensus about the definition of what be a "natural grouping". The method of cluster stability defined by [1, 2] is a process used to detect the presence of clusters in data and is the basis for such definition as well. The results of clustering are easily corrupted through addition of perturbations in the system, such as a small subsample. The method of cluster stability is based on these subsamples. The stability of the group is then assured with a defined pattern of partitions (clusters) that are more important [1, 2].

The method described here uses a hierarchical clustering algorithm with distances (differences) calculated by average linkage clustering [4, 10, 16, 24]. A matrix of similarity is calculated using the coefficient of correlation of Pearson [5]. The Pearson correlation coefficient is a measure of association between two bacteria, and is defined by:

$$S_{ij} = \frac{\sum_{k=1}^{n}\left(C_{k,i}-\overline{C_i}\right)\left(C_{k,j}-\overline{C_j}\right)}{\sqrt{\sum_{k=1}^{n}\left(C_{k,i}-\overline{C_i}\right)^2}\sqrt{\sum_{k=1}^{n}\left(C_{k,j}-\overline{C_j}\right)^2}} \quad (1)$$

Where $S_{i,j}$ is the similarity between bacterium i and j, $C_{k,i}$ is the value that object i takes on for character k (in this work represented by the lanes from image gels), and $\overline{C_i}$ is the mean of all the character values of bacterium i. A value of the correlation coefficient equal to 1 indicates that there is a perfect association, while the value 0 (zero) indicates that there is no association.

The analysis of cluster stability was done introducing perturbations into the system using a technique of sub-sampling [1, 2, 11, 18].

The dataset was clusterized using the similarity matrix obtained by Equation (1). Thus each bacterium was signed to a cluster and compared itself and with the other 118 obtaining the representation matrix of reference D^1, giving by the Equation (2).

$$D_{ij} = \begin{cases} 1, \text{ if } d_i, d_j \text{ belong to the same cluster, and i}\neq\text{j} \\ 0, \text{otherwise} \end{cases} \quad (2)$$

The matrix D^1 is square (119 x 119) where the principal diagonal is zero. The same steps are done for all subsamples generating for each one a representation matrix of reference D^2. The inner product, given by Equations (3), (4) and (5), counts the number of pairs of elements clustered together in both clustering. This inner product can be normalized [1] into a stability measure by Equation (6).

$$\left\langle D^1, D^2 \right\rangle = \sum_{i,j} D^1_{ij} D^2_{ij} \quad (3)$$

$$\left\langle D^1, D^1 \right\rangle = \sum_{i,j} D^1_{ij} D^1_{ij} \quad (4)$$

$$\left\langle D^2, D^2 \right\rangle = \sum_{i,j} D^2_{ij} D^2_{ij} \quad (5)$$

$$S\left(D^1, D^2\right) = \frac{\left\langle D^1, D^2 \right\rangle}{\sqrt{\left\langle D^1, D^1 \right\rangle \left\langle D^2, D^2 \right\rangle}} \quad (6)$$

The objective of this work is to analyze the performance of the restriction enzymes on the three ribosomal regions in relation to the results obtained in cluster identification of the genus *Bradyrhizobium*, considering one collection of Brazilian strains [17] and using the cluster stability method. The paper is organized as follows. In section 2, we present the collection of

bacteria and describe the complete method used. Section 3 presents the results and discussions while section 4 contains the conclusions.

2 – MATERIALS AND METHODS

A Brazilian culture collection of 119 strains for *Bradyrhizobium*, isolated from thirty-three legume species, representing nine tribes and all three subfamilies of the family *Leguminosae* were analyzed by RFLP-PCR. The DNAs of the strains were analyzed by the three ribosomal regions followed by the digestion with three restriction enzymes per region, as follows: 16S rRNA (*Cfo* I , *Msp* I and *Dde* I) , 23S rRNA (*Hha* I, *Hae* III and *Hinf* I) and IGS (*Msp* I, *Dde* I and *Hae* III). Details about the methodology are given in [14]. The electrophoresis gels (17 x 11 cm) obtained were stained with ethydium bromide and photographed under UV radiation using a digital Kodak DC120 (Eastman Kodak). With the objective of simplifying the method, the combinations between restriction enzymes and ribosomal regions were enumerated (Table 1).

Table 1 – Relation of restriction enzymes used in this work and respective ribosomal regions.

Number	restriction enzyme	ribosomal region
1	*Cfo* I	16S
2	*Dde* I	16S
3	*Dde* I	IGS
4	*Hae* III	IGS
5	*Hae* III	23S
6	*Hha* I	23S
7	*Hinf* I	23S
8	*Msp* I	16S
9	*Msp* I	IGS

The method has begun with a processing of photographs of the DNA gels resulting from a method of electrophoresis. The images were processed to remove the background noise and lane segmentation. All the combinations within the ribosomal regions and restriction enzymes were processed, generating 511 experiments. Each experiment was applied to bacteria dataset. For example, 001 to 009 represents the class of experiments with only one restriction enzyme and ribosomal region, experiments 010 to 045 are all combinations of restriction enzymes and ribosomal regions (C_9^2).

Following, for each experiment, a cluster stability analysis was done as proposed by [1, 2] and adapted to this situation. We define a reference cluster containing the 119 bacteria analyzed and a fraction of sampled patterns (f=0.8) equal to 95 bacteria being the subsamples to be analyzed. The number of possible stable clusters present in conjunction with data utilized was: k = 2,…, 10. Thus, a comparison of the reference cluster with the sample cluster was made for each partition. Within each section a count was made of the pairs of grouped patterns together in both clusters and can also be interpreted how a number of structures in common to the diagrams of the reference cluster and of the subsample, generating a measure of similarity within pairs (Equation 6). This process was repeated for 25 subsamples, exemplified in Table 2.

A cluster has been considered stable when all similarities of 25 subsamples were over 0.65. Therefore, the similarity (coefficients of correlation) is calculated between pairs (cluster of reference and resample cluster) and the stability is a result of the analysis of the calculated similarities of the whole combination.

Table 2 – Relation of similarities (experiment 293). For each section (k = 2, 3, ..., 10 clusters) the similarities were calculated among 25 subsamples relative to the reference.

Subsample	K= 2	K = 3	.....	K = 9	K= 10
1	0.793	0.687	.....	0.380	0.461
2	0.802	0.720	.....	0.428	0.443
3	0.777	0.740	.....	0.399	0.403
.	.	.	.....	.	.
.	.	.	.....	.	.
.	.	.	.....	.	.
.	.	.	.....	.	.
24	0.785	0.670	.....	0.473	0.538
25	0.793	0.757	.....	0.425	0.436

The average similarity was calculated by average among all similarities of the stable clusters.

3 – RESULTS

In Figure 1, the *x*-axis represents the number of the experiment and the *y*-axis represents the number of stable clusters with similarities above 0.65. It can be seen that the number of stable clusters increase with the number of the experiment, indicating that the addition of new information from the genome brings an increase of the number of stable clusters (solid line). A large number of experiments was concentrated in 4, 5 and 6 clusters. The dotted line is a polynomial function of interpolation of the third degree (Equation 7) used to analyze the trend of growth of number of stable clusters and shows that the system still has not reached stability. This equation was obtained using Polyfit function of MATLAB® [25]. The Polyfit function finds the coefficients of a polynomial P(X) using the number of experiments (X axis - value 1 to 511), the number of stable cluster in each experiment (Y) and degree of the function (N). The X axis is determinate by the experiment number; the Y axis is determinate by the Polyval function of MATLAB® [25].

$$P = P(1) * X^3 + P(2) * X^2 + P(3) * X + P(4) \qquad (7)$$

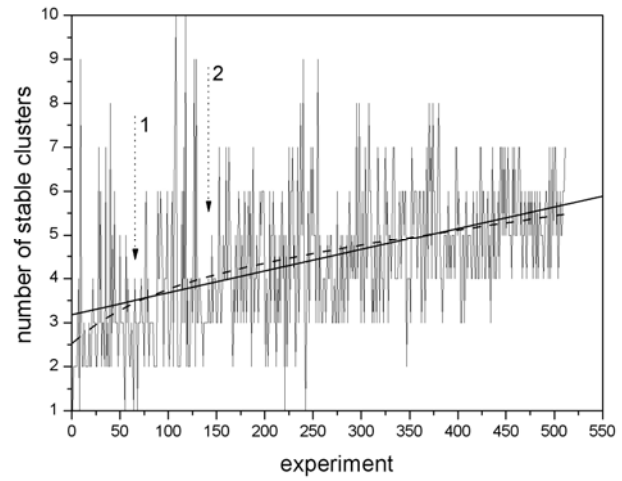

Figure 1: The number of stable clusters with coefficients of correlation over 0.65. The X-axis is the number of experiment (1-511) and the Y-axis is the number of clusters with similarity over 0.65. The Y-axis represents the maximum number of stable clusters (k = 1, 2, …, 10) for each experiment obtained with similarities over 0.65 for all subsamples, as show in Table 2. The dash line is the interpolation function of degree three and the continuous line is the linear function to analyze the trend of the growth of the number of stable clusters. Arrows 1 e 2 indicate regions with a small number of stable clusters.

In Figure 2, the *x*-axis represents the number of the experiment and the *y*-axis represents the average similarity for each experiment. The values for the *y*-axis were obtained by calculating for each experiment the average similarity within all the stable groups. The dotted lines show that the first experiments have the most variance and the variance decreases with the addition of more enzymes. It can be interpreted that when information are added to the system the variance of the system decreases and the similarities trend to reach stable value.

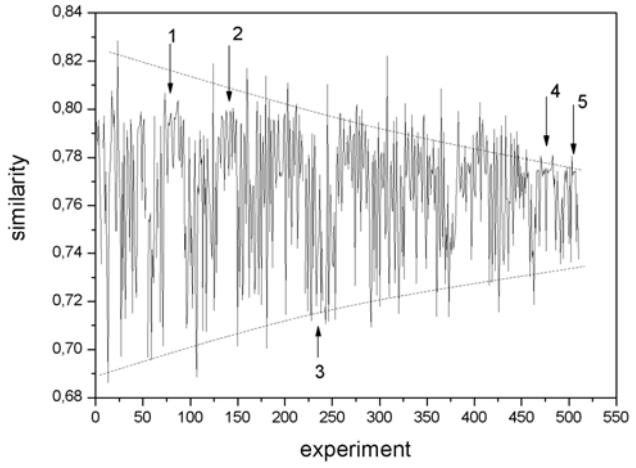

Figure 2: Mean similarities by number of experiment. The X-axis is the number of experiment (1-511) and the Y-axis is the similarity. The regions 1, 2, 4 and 5 show a set of experiments with high similarities. The region 3 shows a set of experiments with low similarities. The dash lines show the trend of the similarities to a reach stable value.

Table 3 shows the enzyme/ribosomal region predominant in the regions indicated by the arrows of Figure 2. Arrows 1, 2, 4 e 5 show groupings that maintain high similarities. In arrow 1 the predominance is of the enzyme/ribosomal region *Cfo* I 16S. In arrow 2 the predominance is of the enzyme/ribosomal region *Cfo* I 16S and *Dde* I 16S. In arrow 4 the predominance is the enzyme/ribosomal region *Dde* I 16S and *Dde* I IGS. In arrow 5 the predominance is of the enzyme/ribosomal region *Cfo* I 16S, *Dde* I 16S and *Dde* I IGS. The arrow 3 shows groupings of experiments that maintain low similarities, what the predominance was the pair of enzyme/ribosomal region *Dde* I IGS and *Hae* III IGS. Both cases the IGS region have the highest variance in the same population.

Table 3 – Relation of enzyme/ribosomal region predominant

Arrows	Enzyme/ribosomal region predominant
1	*Cfo* I 16S
2	*Cfo* I 16S and *Dde* I 16S
3	*Dde* I IGS and *Hae* III IGS
4	*Dde* I 16S and *Dde* I IGS
5	*Cfo* I 16S *Dde* I 16S and *Dde* I IGS

4 – CONCLUSION

The addition of enzymes/ribosomal region increased the number of stable clusters, as shown in Figure 1. Initial variance of the system decreased when enzymes were added and the similarities trended to be concentrated in a stable value (near 0.76). Groupings until seven clusters (k = 2, ... , 7), for the system analyzed, were sufficient to achieve a reduction of time expense in simulations, that were 360 hours, using 7 Pentium IV computers, with Linux.

The utilization of 9 combinations of enzyme/ribosomal regions increases the cost of biological analysis and the time of processing, at the same time that makes the analyses of results much more complex.

The Figures 1 and 2 show the trend to diminish the variations and to concentrate in 5 stable clusters (Figure 1) with 0.76 of similarity (coefficient of correlation) (Figure 2). Regions of high similarities (i.e, Regions 1 and 2) of Figure 2 are the regions with a few stable clusters (Regions 1 and 2 in Figure 1), indicating that a bigger similarity comes with the few stable clusters.

Enzymes number 1 (*Cfo* I 16S) and number 2 (*Dde* I 16S) increased the similarities of the experiment. The enzyme number 3 (*Dde* I IGS) increased the similarities of experiments when associated with several enzymes, seven and eight, and decreased the similarities of the experiments with four enzymes. Enzyme number 4 (*Hae* III IGS) decreased the similarities of the experiments. As a main result it should be considered only the regions 16S and 23S for this kind of study keeping the same restriction enzymes as shown in Table 1. In this case the time expense should be reduced to 240 hours.

REFERENCES

[1] Ben-Hur, A. and Guyon, I. In: Brownstein, M. J. and Kohodursky, A. eds. "Methods in Molecular Biology". Humana press, Clifton, pp. 159-182, 2003.

[2] Ben-Hur, A., Elisseeff, A. and Guyon, I. "A stability based method for discovering structure in clustered data". In Pac. Symp. Biocomputing. R. Altman, A. Dunker, L. Hunter, K. Lauderdale, and T. Klein, eds. World Scientific, Hawaii, pp. 6-17, 2002.

[3] Jarabo-Lorenzo, A., Velázquez, E., Pérez-Galdona, R., Veja-Hernández, M. C., Martínez-Molina, E., Mateos, P. F., Vinuesa, P., Martínez-Romero, E. and León-Barrios, M. "Restriction fragment length polymorphism analysis of 16S rDNA and low molecular weight RNA profiling of rhizobial isolates from shrubby legumes endemic to the Canary Islands". Systematic and Applied Microbiology, Stuttgart, v. 23, pp. 418-425, 2000.

[4] Jain, A. K., Murty, M. N. and Flynn, P. J. "Data Clustering: A Review". ACM Computing Surveys. New York, v. 31, n. 3, pp. 264-323, 1999.

[5] van Ooyen, A. "Theoretical Aspects of Pattern Analysis". In: L. Dijkshoom, K. J. Tower and M. Struelens eds. New Approaches for Generation and Analysis of Microbial Fingerprint, Elsevier, Amsterdam, pp. 31-45, 2001.

[6] Willems, A., Coopman, R. and Gillis, M. "Comparison of sequence analysis of 16S- 23S rDNA spacer regions, AFLP analysis and DNA-DNA hybridizations in Bradyrhizobium". International Journal of Systematic Bacteriology, v. 51, pp. 623-632, 2001.

[7] Everitt, B. S., Landau, S. and Leese, M. "Cluster Analysis". London: Arnold, 2001.

[8] Woese, C. R. Bacterial evolution. "Microbiology Reviews", Washington, v. 51, pp. 221-271, 1987.

[9] Jordan, D. C. "Rhizobiaceae Conn 1938". In: Krieg, N. R. and Holt, J. G., eds., Bergeys's Manual of Systematic Bacteriology, Baltimore, pp. 235-244, 1984.

[10] Graur, D. and Li, W., "Fundamental of Molecular Evolution". 2nd ed., Sinauer Associates, Suderland, Massachusetts, 2000.

[11] Levine, E. and Domany, E. "Resampling Method for Unsupervised Estimation of Cluster Validity". Neural Computation, v. 13, pp. 2573-2593, 2001.

[12] Wang, E. T., van Berkun, P., Sui, X. H., Beyene, D., Chen, W. X. and Martínez-Romero, E. "Diversity of rhizobia associated with Amorpha fruticosa from Chinese soils and description of Mesorhizobium amorphae sp. nov.". International Journal of Systematic Bacteriology, Washington, v. 49, pp. 51-65, 1999.

[13] Felsenstein, J. "Software PHYLIP", Phylogeny Inference Package, v. 3.6. Department of Genome Sciences, University of Washington., 2002.

[14] Laguerre, G., Allard, M. R., Revoy, F. and Amarger, N. "Rapid identification of rhizobia by restriction fragment length polymorphism analysis of PCR-amplified 16S rRNA genes". Applied and Environmental Microbiology, Washington, v. 60, pp. 56-63, 1994.

[15] Laguerre, G., Mavingui, P., Allard, M. R., Charnay, M. P., Louvrier, P., Mazurier, S. I., Rigottier-Gois, L. and Amarger, N. "Typing of rhizobia by PCR and PCR-restriction fragment length polymorphism analysis of chromossomal and symbiotic gene regions: application to Rhizobium leguminosarum and its differents biovars". Applied and Environmental Microbiology, Washington, v. 62, pp. 2029-2036, 1996.

[16] Quackenbush, J. "Computational analysis of microarray data". Nature Reviews Genetics, Maryland, v. 2, pp. 418-427, 2001.

[17] Germano-Silva, M. "Avaliação da diversidade genética de Bradyrhizobium pela análise de genes ribossomais". Universidade Estadual de Londrina, Londrina, Brazil (M. Sc. Dissertation), 2003.

[18] Meilă, M. "Comparing Clustering". UW Statistics Technical Report, pp. 418, 2003.

[19] Garrity, G. M., Holt, J. G. "The road map to the Manual". In: Garrity, G. M. Boone, D. R. and Castenholz, B. W., eds. Bergey's Manual of Systematic Bacteriology, v. 1, 2nd ed., New York, 2001.

[20] van Berkum, P. and Fuhrmann, J. J. "Evolutionary relationships among soybean bradyrhizobial reconstructed from 16S rRNA gene and internally transcribed spacer region sequence divergence". Int. Journal of Systematic Bacteriology, Washington, v. 50, pp. 2165, 2000.

[21] Vinuesa, P., Rademaker, J. L. W., de Bruijn, F. J. and Werner, D. "Genotypic characterization of Bradyrhizobium strains nodulating endemic woody legumes of the Canary Islands by PCR-Restriction Fragment Length Polymorphism analysis of genes encoding 16S rRNA (16S rDNA) and 16S-23S rRNA intergenic spacers, repetitive extragenic palindromic PCR genomic fingerprinting, and partial 16S rRNA sequencing". Applied and Environmental Microbiology, Washington, v. 64, pp. 2096-2104, 1998.

[22] Johnson, R. A. and Wichern, D. W. In: "Aspects of Multivariate Analysis". Applied Multivariate Statistical Analysis. Prentice Hall, New Jersey, 1992.

[23] Abaidoo, R. C., Keyser, H. H., Singleton, P. W. and Borthakur, D. "Bradyrhizobium spp. (TGx) isolates nodulating the new soybean cultivars in Africa are diverse and distinct from bradyrhizobia that nodulate North American soybeans". International Journal of Systematic Evolutionary Microbiology, Washington, v. 50, pp. 225-234, 2000.

[24] Shamir, R. and Sharan, R. "Algorithmic Approaches to Clustering Gene Expression Data". In: Jiang, T.; Smith, T.; Xu, Y.; Zhang, M. Q., eds., Current Topics in Computational Biology. Massachussets, MIT Press, pp. 269-300, 2002.

[25] The Mathworks, Inc. "Software MATLAB, The Language of Technical Computing", v. 6.1.0.450, 2001.

[26] Roth, V., Lange, T., Braun, M. and Buhmann, J. A. "A Resampling Approach to Cluster Validation". In Härdle, W. and Röns, B. eds. Computational Statistics, 123, Heidelberg, Physica-Verlag, 2000.

[27] Weisburg, W. G., Barns, S. M., Pelletie, D. A. and Lane, D. J. "16S ribosomal DNA amplification for phylogenetic study". Journal of Bacteriology, Washington, v. 173, pp. 697-703, 1991.

[28] Ludwig, W. and Schleifer, K. H. "Bacterial phylogeny based on 16S and 23S rRNA sequence analysis". FEMS Microbioogy, Amsterdan, v. 15, pp. 155, 1994.

[29] Halkidi, M., Batistakis, Y. and Vazirgiannis, M. "On Clustering Validation Techniques". Journal of Intelligent Information Systems, v. 17, Issue 2-3, pp.107-145, 2001.