Journal of Computational and Applied Mathematics 227 (2009) 308–319

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Multidimensional cluster stability analysis from a Brazilian *Bradyrhizobium* sp. RFLP/PCR data set

S.T. Milagre [a], C.D. Maciel [b,*], A.A. Shinoda [c], M. Hungria [d], J.R.B. Almeida [b]

[a] *Computer Science, Goiás Federal University, Catalão, Brazil*
[b] *Electrical Eng. Department, University of São Paulo, São Carlos, Brazil*
[c] *Electrical Eng. Department, State University of São Paulo, Ilha Solteira, Brazil*
[d] *Soil Biotechnology Laboratory, Embrapa Soja, Londrina, Brazil*

ARTICLE INFO

Article history:
Received 19 June 2006
Received in revised form 30 September 2007

Keywords:
Cluster Analysis
Bradyrhizobium Genus
bioinformatics

ABSTRACT

The taxonomy of the N_2-fixing bacteria belonging to the genus *Bradyrhizobium* is still poorly refined, mainly due to conflicting results obtained by the analysis of the phenotypic and genotypic properties. This paper presents an application of a method aiming at the identification of possible new clusters within a Brazilian collection of 119 *Bradyrhizobium* strains showing phenotypic characteristics of *B. japonicum* and *B. elkanii*. The stability was studied as a function of the number of restriction enzymes used in the RFLP-PCR analysis of three ribosomal regions with three restriction enzymes per region. The method proposed here uses clustering algorithms with distances calculated by average-linkage clustering. Introducing perturbations using sub-sampling techniques makes the stability analysis. The method showed efficacy in the grouping of the species *B. japonicum* and *B. elkanii*. Furthermore, two new clusters were clearly defined, indicating possible new species, and sub-clusters within each detected cluster.

1. Introduction

The ribosomal genes, with emphasis on the 16S rRNA, have been the preferred molecules to trace bacterium phylogenies since they are highly conserved, but with enough variability to enable species cluster analyses, inferring common ancestors and evolutionary progression [19,10]. As a result of the increasing use of ribosomal sequences for taxonomic purposes, identification of genotypic, detection of new species, and environmental monitoring, among others, the deposition of sequencing data in databases that are free to consult is growing exponentially.

Sequencing analysis can be very expensive; however, there are other cheaper methods to analyze ribosomal genes, which can be used as a first approach to evaluate diversity and taxonomic position. It has been shown that the amplification of DNA region coding for ribosomal genes by the PCR (polymerize chain reaction) technique, followed by digestion with restriction enzymes [the RFLP (restriction fragment length polymorphism)–PCR technique] correlates quite well with the sequencing analysis of those genes [28,2,14]. The lower cost of this technique can be useful as a first step to investigate diversity in the tropics, where few studies have been performed, despite wide indications that the region carries the highest levels of diversity known so far. However, the analysis of the electrophoretic profiles produced by RFLP-PCR analysis can be critical for the correct assignment of clusters and species. The results of RFLP-PCR analyses are images with, in most cases, high

* Corresponding address: Electrical Eng. Department, University of São Paulo, Av. Trabalhador São-carlense, 400, 13566-590 São Carlos, SP, Brazil. Tel.: +55 16 3373 9366; fax: +55 16 3373 9372.
E-mail address: maciel@sel.eesc.usp.br (C.D. Maciel).

0377-0427/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2008.03.018

background noise, low contrast and geometrical deformation which may result in different interpretations. Thus, the analysis of the electrophoretic profiles needs to be stable, reproducible, and avoid individual interpretation.

Clustering is widely used in exploratory analysis of biological data. The goal is the partitioning of the elements into subsets, which are called clusters, so that two criteria are satisfied: homogeneity (elements in the same cluster are highly similar to each other) and separation (elements from different clusters have low similarity to each other) [13,25]. The analysis of cluster stability is a means of assessing the validity of data partitioning found by clustering algorithms [24,12,18].

Recently, the research of microorganism population has been increased with the approach of much information from DNA. In [15] the authors described a population structure of the *Bacillus cereus* group (52 strains of *B. anthracis*, *B. cereus*, and *B. thuringiensis*) from sequencing of seven gene fragments. Most of the strains were classifiable into two large sub-groups in six housekeeping gene. As a result there were several consistent clusters with distinct biological interpretations. Also, [6] used viral diseases of tomato caused by monopartite geminiviruses (family *Geminiviridae*) from countries around the Nile and Mediterranean Basins. The molecular biodiversity of these viruses was investigated to better appreciate the role and importance of recombination and to better clarify the phylogenetic relationships and classification of these viruses.

On the other hand, as many DNA regions are incorporated into the analysis, the data are becoming more complex and new approaches need to be developed. In [1] the authors made a comparison of the phylogeny of 38 isolates of chemolithoautotrophic ammonia-oxidizing bacteria based on 16S rRNA and 16S–23S rDNA intergenic spacer region sequences was performed to species affiliations based on DNA homology values. In [20] the phylogenetic relationships of 51 isolates representing 27 species of *Phytophthora* was studied by sequence alignment of the mitochondrially encoded cytochrome oxidase II gene. The authors compared the results from a partition homogeneity from ITS cox II. The study was made from trees constructed by a heuristic search, based on maximum parsimony for a bootstrap 50% majority-rule consensus tree.

The method described here uses clustering algorithms [5] with the matrix of similarity calculated by Pearson correlation [27] from nine restriction enzymes (three for each of the three ribosomal regions). The stability analysis was performed by introducing perturbations using sub-sampling techniques [3,4,18,21]. The consensus trees were generated using the Phylogenic Inference Package (PHYLIP) [7]. This multidimensional approach will consider a set from these combinations. The total number of sets represents 511 experiment combinations. Most of the time, phylogenetic studies are developed from a specific experiment. In our analysis, the experiments were grouped by the resulting number of stable clusters. A consensus tree was performed inside these groups. The main supposition around this procedure is that the consensus tree should be better performed using a similar set obtained from a same number of stable clusters.

This work aimed at the identification of clusters within the genus *Bradyrhizobium*, considering a collection of Brazilian strains and using the multidimensional cluster stability method. The method was performed as a function of the number of enzymes used in the RFLP-PCR analysis of three ribosomal regions. It has been suggested that variability in the 1.5 kb of the 16S rRNA region of *Bradyrhizobium* is very low [26,30]. Thus, two other regions were included in our study, the 23S rRNA, with a longer fragment (about 2.3 kb) and a faster rate of sequence change [19], and the 16S-23S rRNA intergenic spacers (IGS) [26,30].

The paper is organized as follows. In Section 2, we present the concepts of similarity, stable cluster and consensus tree. In Section 3, we present the collection of bacteria and describe the complete method used. Section 4 presents the results and discussion and Section 5 contains the conclusions.

2. Theory

Clustering is one of the most useful tasks in data mining processes for discovery groups and identifying interesting patterns in underlying data. A large data set often consists of many clusters, and some of these clusters may just be the result from noise or from an artifact from the process. Different clustering processed may result in a different partition of the data set. 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. This is the main subject of cluster validation. For a low-dimensional data set, it is clear that visualization of the data set and clusters is a crucial verification of clustering results. In the case of more than three-dimensional data sets, the effective visualization would be a hard task.

Typically, the application of any cluster algorithm needs the choice of specific parameters like number of clusters. The results supplied by the clustering algorithm may depend strongly on this choice. At the lowest resolution, all N points belong to one cluster; on the other hand, one has N clusters of a single point each. As the resolution is changed, data points may be broken into different sub-clusters. In our case, one would like to pursue a specific partitioning of the data that captures a particular important aspect described by a natural clustering in the data set. One of the most important issues in cluster analysis is the evaluation of clustering results to find the partition that best fits the data set. For a comparative analysis of clustering and validation techniques see [12], or for a clustering review see [5].

A clustering C is a partition of data set D into sets $C_1, C_2, \ldots, C_k$ called clusters such that $C_k \cap C_l = 0$ and $\sum_{k=1}^{K} C_k = D$. Let the number of data points in D and in cluster C_k be n_k, $n = \sum_{k=1}^{K} n_k$; it will also be assumed that $n_k > 0$. The parameter K represents the number of non-empty clusters in D. Let a second clustering of the same data set D be $C' = \{C'_1, C'_2, \ldots, C'_k\}$ with individual clusters of size n'_k. An important class of criteria for comparing clustering is based on counting the pairs of points on which two clusterings agree/disagree.

The measures of similarity between two clusters proposed [18,3,21] will be briefly described and discussed in terms of an improvement to adapt to this problem. The matrix representation of a partition is defined by

$$m_{ij} = \begin{cases} 1, & \text{if } d_i \text{ and } d_j \text{ belong to the same cluster} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where d_i and d_j are elements from the data set under study. The partitions C and C' have matrix representations M and M', respectively. The inner product

$$\langle M, M' \rangle = \sum_{i,j} m_{ij} m'_{ij} \tag{2}$$

counts the number of pairs of elements clustered together in both clusterings. This inner product can be normalized [3] into a stability measure by

$$s(M, M') = \frac{\langle M, M' \rangle}{\sqrt{\langle M, M \rangle \langle M', M' \rangle}}. \tag{3}$$

The use of resampling to discover natural clustering is an intensive computational approach. Depending on how large the data set is and the number of sub-samples, the computational resources needed until now have been insufficient and a personal computer may not be the best environment. On the other hand, many works have been done on how to improve this computational performance using a computer cluster for a better performance; see e.g. [22] or [23].

To evaluate the clustering C using resampling [21,18], one considers m new data sets constructed from randomly resampling from M, M', with a sampling ratio f, $0 < f < 1$. To evaluate the clustering C' from M' one considers the metric Eq. (3) between C and C' but using only the data points from M' present in M.

This main idea is to compare a reference cluster obtained from all samples with many clusters from sub-samples of the original dataset. Similarity is calculated between C and C' and the stability is evaluated for the whole collection of similarities.

For a natural partition, [4] and [3] adopted the data set if the similarity is concentrated near one. It can be observed that sub-samples with high similarity have the same general structure as the complete dataset, so this cluster is stable. In accordance with [17] the similarities between C with different clustering C' is an estimation problem where C' is a stochastic process that generates different partitions on different runs. In our approach, we adopted a threshold value and used a hypothesis test with $p < 0.05$ to discern if the sequence of similarities was performed from a stable partition.

The experiments are grouped by the number of clusters that are stable and a consensus tree is obtained for each group. The consensus method used in this study is the Majority Rule (extended) where any set of species that appears in 50% or more of the trees is included. To complete the tree, the other sets of species are considered in the order of the frequency in which they appear, adding to the consensus tree any which is compatible with it until the tree is fully resolved [7].

3. Materials and methods

All strains used are from the Brazilian culture collection of rhizobia, classified as *Bradyrhizobium* in the catalogue of [8]. The data set consists of a 119 strains of *Bradyrhizobium* isolated from 33 legume species, representing nine tribes, and all three sub-families of the family *Leguminosae* were analyzed by RFLP-PCR. The strains have been described elsewhere [11], and the RFLP-PCR process will be briefly described. This study used RFLP-PCR-amplified DNA region coding from 16S, 23S and 16S-23S rRNA intergenic spacer (IGS) from rRNA genes, and three replicates of DNA of each bacterium were used for the amplification. For 16S, universal primers described by [29] were used. The PCR products were then digested with three restriction endonucleases, *Cfo*I, *Msp*I and *Dde*I (Invitrogen - Life Technologies), as recommended by the manufacturers. The fragments obtained were analyzed by electrophoresis in a gel (17 × 11 cm) with 3% agarose, and carried out at 100 V for 4 h. The gels were stained with ethydium bromide and photographed under UV light. RFLP-PCR of the 23S rRNA region was amplified with primers P3 and P4 described by [20]. The PCR products were digested with three restriction endonucleases, *Hha*I (= *Cfo*I), *Hae*III and *Hinf*I, as recommended by the manufacturers. RFLP-PCR of the 16S-23S rRNA intergenic spacer was amplified with primers FGPS1490 and FGPS 132 described by [16]. The PCR products were then digested with the restriction enzymes *Msp*I, *Dde*I and *Hae*III (Invitrogen-Life Technologies), as recommended by the manufacturers. The fragments were visualized as described in the RFLP-PCR of the 16S rRNA region.

Among these strains, six have been shown to belong to the species *B. japonicum* (SEMIA 566, SEMIA 586, SEMIA 5056, SEMIA 5079, SEMIA 5080 and SEMIA 5085) and *B. elkanii* (SEMIA 587 and SEMIA 5019) [9]. Strain SEMIA 5056 is the same as USDA 6, the type of strain for the species *B. japonicum*. Furthermore, two reference strains were included: *B. elkanii* type strain USDA 76 and *B. elkanii* BTAi 1, a strain that nodulates roots and stems of *Aeschynomene* and seems to occupy a distinct phylogenetic position [14]. The DNAs of the strains were analyzed by the RFLP-PCR of 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 (= *Cfo*I), *Hae*III and *Hin*fI) and IGS (*Msp*I, *Dde*I and *Hae*III). Details of the methodology are given elsewhere [11]. The electrophoresis gels (17 × 11 cm) obtained were stained with ethydium bromide and photographed under UV radiation using a digital Kodak DC120 camera (Eastman Kodak).

To simplify the description of the method, a reference name was given for each combination of restriction enzyme and ribosomal region: enzyme **1**(Cfo I -16S), enzyme **2**(Dde I -16S), enzyme **3**(Dde I –IGS), enzyme **4**(Hae III - IGS), enzyme **5**(Hae

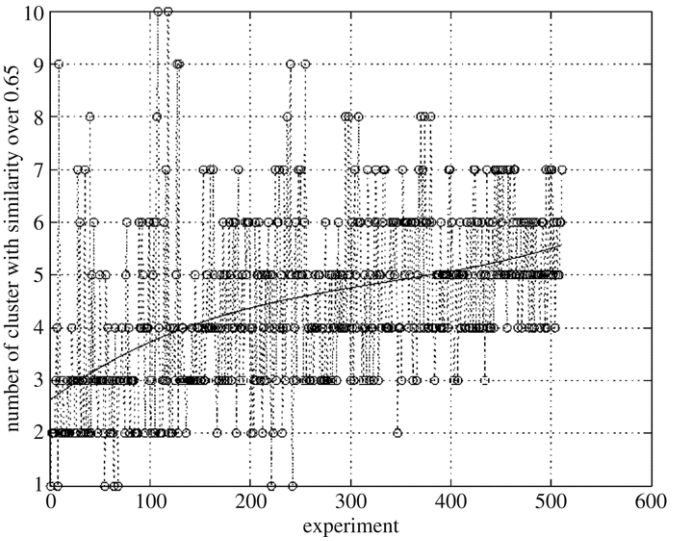

Fig. 1. Number of stable clusters. The x-axis is the experiment number (1–511) and the y-axis is the number of clusters with similarity over 0.65 (represented by circle). The y-axis represents the maximum number of stable clusters ($k = 1, \ldots, 10$) for each experiment obtained with similarities over 0.65 for all sub-samples, as shown in Table 2. The continuous line is the interpolation function of degree three to analyze the tendency of the growth of the number of stable clusters. This shows that the system is not yet stable.

III - 23S), enzyme **6**(Hha I - 23S), enzyme **7**(Hinf I - 23S), enzyme **8**(Msp I - 16S) and enzyme **9**(Msp I – IGS). The first part of the method starts with image processing (noise removal and segmentation of lanes) of the electrophoresis gels. The lanes of the gels were separated one by one into files. These files of images were submitted to a treatment for the removal of background noise, attenuation in the formats of the bands and the removal of tendencies of irregular growth.

After pre-processing, the files of images were normalized making the conversion of the images into numbers and creating a matrix $m \times n$, where m is the bacteria data for one respective enzyme and n is the length of the gel. All combinations of bacteria and enzymes were processed, generating 511 experiments. These combinations were made starting with all bacteria using one enzyme/ribosomal region and followed until nine enzymes were added. All combination of enzyme/ribosomal region are described in Table 1.

The parameters used for evaluation of stability were: numbers of possible clusters present in dataset: $K = 2, \ldots, 10$; fraction of patterns sampled: $f = 0.8$ (95 bacteria); number of sub-sets equal to 25. A cluster has been considered stable when all similarities of 25 sub-sets were over 0.65 and $p > 0.05$. In the second part of the method, a grouping of all experiments by number of stable clusters is performed. A tree is generated in each experiment and grouped by number of stable clusters. Then, these tree collections are processed by the consensus algorithm, using the Majority Rule (extended) [7], generating four consensus trees, one for each partition under study.

4. Results

In the first development, each experiment required one hour of processing, using Octave/Linux and computers Pentium IV with 2.2 GHz and 800 MB of RAM. The processing was divided among seven computers and the processing of the 511 experiments took approximately 360 h. A C program running in a cluster with ten computers (Xeon Dual 2.4 GHz and 1 GB RAM) with Linux - OpenMosix/MPI took eight hours of processing.

In Fig. 1, the x-axis is the experiment number (1–511) and the y-axis is the number of stable clusters for each experiment (represented by a circle). It can be observed that the number of stable clusters increases with the number of experiments, indicating that when new information from the genome is added to the analysis the number of stable clusters increases. The continuous line is a polynomial interpolation function of three degrees to analyze the growth tendency of the numbers of stable clusters. This shows that the system is not stable yet and the inclusion of more regions would be necessary to complete the study.

The numbers of stable clustering are concentrated in three, four, five and six partitions that accumulate 78% of experiments, two partitions accumulated 11% and all others ($k = 7, 8$ and 9) accumulated less than 4% of the total experiments. Only consensus trees belonging to these collections of stable clusters have been considered.

In Fig. 2, the x-axis is the experiment number (1–511) and the y-axis is the stability (represented by a circle). It can be observed that the similarities have high variance for the first experiments, decreasing as the number of experiments increases, tending to concentrate near 0.76 when the number of experiments is around 500 (these experiments use eight and nine enzymes). This can be interpreted as when enzymes are added to the system the initial variance of the system decreases and the similarities tend to reach a stable value.

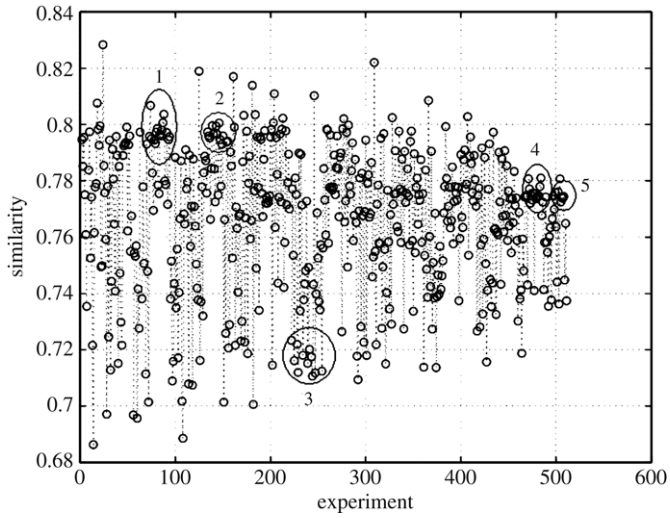

Fig. 2. Mean similarities by number of experiment. The *x*-axis is the experiment number (1–511) and the *y*-axis is the similarity (represented by a circle). The *y*-axis is obtained by calculating the average similarities for each experiment among all stable clusters. The regions 1, 2, 4 and 5 show a set of experiments with high similarities. The region 3 shows a set of experiments with low similarities.

In addition, in Fig. 2, the circles 1, 2, 4 and 5 show a set of experiments with high similarities. In circle 1, the predominant enzyme/ribosomal region is *Cfo*I 16S, and in circle 2, the predominant enzyme/ribosomal regions are *Cfo*I 16S and *Dde*I 16S. In circle 4, the predominant enzyme/ribosomal regions are *Dde*I 16S and *Dde*I IGS and in circle 5, the predominant enzyme/ribosomal regions are *Cfo*I 16S, *Dde*I 16S and *Dde*I IGS. Circle 3 shows a set of experiments with low similarities and the predominant enzyme/ribosomal regions are *Dde*I IGS and *Hae*III IGS. As expected, the 16S region is important for a stable cluster formation, while IGS performs a variability that induces a low stability experiment.

Figs. 3–6 show the dendrograms of the consensus tree from these stable clusters, respectively. Five clusters including the same strains were maintained in these four consensus trees, with differences only in the position inside of each tree. The analysis of the consensus trees were then made in relation to these five clusters, named A, B, C, D and E. Cluster A presented a variation in the placement of the strains inside the sub-clusters and in the lengths of branches among four consensus trees, as well as a high level of variability, with the formation of several sub-clusters. Cluster A grouped all reference strains of the *B. japonicum* species: SEMIA 566, SEMIA 586, SEMIA 5079, SEMIA 5080, SEMIA 5085, and the type strain SEMIA 5056. The small cluster B was similar in both consensus trees and grouped only three strains, SEMIA 6166, SEMIA 6167 and SEMIA 6154. Cluster C grouped two reference strains of *B. elkanii* species, SEMIA 587 and SEMIA 5019 and the strain BTAi 1 (*Bradyrhizobium* sp.). Cluster D grouped the same strains in these four consensus trees, and the same sub-clusters were observed. Type strain USDA 76 of *B. elkanii* fit into an isolated branch. The strains found in cluster E were the same in these four consensus trees; differences were observed in the position within the sub-clusters. The sub-clusters in cluster E differed from all consensus trees.

Clearly, clusters B, D and E were defined besides *B. japonicum* and *B. elkanii*. Furthermore, the strains that fit into those three clusters did not show the physiological properties of the two other described *Bradyrhizobium* species, *B. yuanmingense* and *B. liaoningense* [11]. Cluster D grouped 13 strains, eight from Brazil, four from Paraguay, and the type strain USDA 76. Cluster E grouped ten strains, eight from Brazil, one from Bolivia, and one from Colombia. Tables 2 and 3 contain the list of strains from clusters D and E, respectively, for the four consensus trees.

5. Conclusion

This work presented a method for the identification of possible natural clusters in a Brazilian culture collection of N_2-fixing *Bradyrhizobium* strains. The five clusters identified (A, B, C, D and E) showed high variability inside of the four consensus trees, indicating that these clusters might represent new species or sub-species. Cluster A grouped a major number of strains and grouped all reference strains of the *B. japonicum*; therefore it might also contain sub-species. Cluster B could represent a new species, as the strains were genetically quite dissimilar from reference strains. Cluster C might also represent a new species, since it grouped BTAi 1, a strain that seems to occupy a distinct phylogenetic position [14]. Although cluster C grouped two reference strains of *B. elkanii* (SEMIA 587 and SEMIA 5019), these strains were isolated in Brazil; thus they might be different from USDA 76. Cluster D might possibly contain sub-species, since grouped type strain USDA 76 of *B. elkanii* occupying an isolated branch in the four consensus trees. Finally, cluster E might certainly represent a new species, since the similarity with *B. japonicum* and *B. elkanii* was very low.

The method used in this study presented an efficient way to group the species *B. japonicum* (cluster A) and *B. elkanii* (cluster C). The five clusters (A, B, C, D and E) obtained were stable, since they were conserved in the four consensus trees. The addition of enzyme/DNA regions increased the number of stable clusters, as shown in Fig. 6. The addition of enzymes

Table 1
Description of all experiments using the enzyme nomenclature described in Figs. 1 and 2

Exp.	Enz.	Exp.	Enz.	Exp.	Enz.	Exp.	Enz..	Exp.	Enz.	Exp	Enz.	Exp.	Enz.	Exp.	Enz.
1	1	66	158	131	1235	196	2368	261	12356	326	15789	391	123489	456	245679
2	2	67	159	132	1236	197	2369	262	12357	327	16789	392	123567	457	245689
3	3	68	167	133	1237	198	2378	263	12358	328	23456	393	123568	458	245789
4	4	69	168	134	1238	199	2379	264	12359	329	23457	394	123569	459	246789
5	5	70	169	135	1239	200	2389	265	12367	330	23458	395	123578	460	256789
6	6	71	178	136	1245	201	2456	266	12368	331	23459	396	123579	461	345678
7	7	72	179	137	1246	202	2457	267	12369	332	23467	397	123589	462	345679
8	8	73	189	138	1247	203	2458	268	12378	333	23468	398	123678	463	345689
9	9	74	234	139	1248	204	2459	269	12379	334	23469	399	123679	464	345789
10	12	75	235	140	1249	205	2467	270	12389	335	23478	400	123689	465	346789
11	13	76	236	141	1256	206	2468	271	12456	336	23479	401	123789	466	1234567
12	14	77	237	142	1257	207	2469	272	12457	337	23489	402	124567	467	1234568
13	15	78	238	143	1258	208	2478	273	12458	338	23567	403	124568	468	1234569
14	16	79	239	144	1259	209	2479	274	12459	339	23568	404	124569	469	1234578
15	17	80	245	145	1267	210	2489	275	12467	340	23569	405	124578	470	1234579
16	18	81	246	146	1268	211	2567	276	12468	341	23578	406	124579	471	1234589
17	19	82	247	147	1269	212	2568	277	12469	342	23579	407	124589	472	1234678
18	23	83	248	148	1278	213	2569	278	12478	343	23589	408	124678	473	1234679
19	24	84	249	149	1279	214	2578	279	12479	344	23678	409	124679	474	1234689
20	25	85	256	150	1289	215	2579	280	12489	345	23679	410	124689	475	1234789
21	26	86	257	151	1345	216	2589	281	12567	346	23689	411	124789	476	1235678
22	27	87	258	152	1346	217	2678	282	12568	347	23789	412	125679	477	1235679
23	28	88	259	153	1347	218	2679	283	12569	348	24567	413	125689	478	1235689
24	29	89	267	154	1348	219	2689	284	12578	349	24568	414	125689	479	1235789
25	34	90	268	155	1349	220	2789	285	12579	350	24569	415	125789	480	1236789
26	35	91	269	156	1356	221	3456	286	12589	351	24578	416	126789	481	1245678
27	36	92	278	157	1357	222	3457	287	12678	352	24579	417	134567	482	1245679
28	37	93	279	158	1358	223	3458	288	12679	353	24589	418	134568	483	1245689
29	38	94	289	159	1359	224	3459	289	12689	354	24678	419	134569	484	1245789
30	39	95	345	160	1367	225	3467	290	12789	355	24679	420	134578	485	1246789
31	45	96	346	161	1368	226	3468	291	13456	356	24689	421	134579	486	1256789
32	46	97	347	162	1369	227	3469	292	13457	357	24789	422	134589	487	1345678
33	47	98	348	163	1378	228	3478	293	13458	358	25678	423	134678	488	1345679
34	48	99	349	164	1379	229	3479	294	13459	359	25679	424	134679	489	1345689
35	49	100	356	165	1389	230	3489	295	13467	360	25689	425	134689	490	1345789
36	56	101	357	166	1456	231	3567	296	13468	361	25789	426	134789	491	1346789
37	57	102	358	167	1457	232	3568	297	13469	362	26789	427	135678	492	1356789
38	58	103	359	168	1458	233	3569	298	13478	363	34567	428	135679	493	1456789
39	59	104	367	169	1459	234	3578	299	13479	364	34568	429	135689	494	2345678
40	67	105	368	170	1467	235	3579	300	13489	365	34569	430	135789	495	2345679
41	68	106	369	171	1468	236	3589	301	13567	366	34578	431	136789	496	2345689
42	69	107	378	172	1469	237	3678	302	13568	367	34579	432	145678	497	2345789
43	78	108	379	173	1478	238	3679	303	13569	368	34589	433	145679	498	2346789
44	79	109	389	174	1479	239	3689	304	13578	369	34678	434	145689	499	2356789
45	89	110	489	175	1489	240	3789	305	13579	370	34679	435	145789	500	2456789
46	123	111	479	176	1567	241	4567	306	13589	371	34689	436	146789	501	3456789
47	124	112	478	177	1568	242	4568	307	13678	372	34789	437	156789	502	12345678
48	125	113	469	178	1569	243	4569	308	13679	373	35678	438	234567	503	12345679
49	126	114	468	179	1578	244	4578	309	13689	374	35679	439	234568	504	12345689
50	127	115	467	180	1579	245	4579	310	13789	375	35689	440	234569	505	12345789
51	128	116	459	181	1589	246	4589	311	14567	376	35789	441	234578	506	12346789
52	129	117	458	182	1678	247	4678	312	14568	377	36789	442	234579	507	12356789
53	134	118	457	183	1679	248	4679	313	14569	378	45678	443	234589	508	12456789
54	135	119	456	184	1689	249	4689	314	14578	379	45679	444	234678	509	13456789
55	136	120	589	185	1789	250	4789	315	14579	380	45689	445	234679	510	23456789
56	137	121	579	186	2345	251	5678	316	14589	381	45789	446	234689	511	123456789
57	138	122	578	187	2346	252	5679	317	14678	382	123456	447	234789		
58	139	123	569	188	2347	253	5689	318	14679	383	123457	448	235678		
59	145	124	568	189	2348	254	5789	319	14689	384	123458	449	235679		
60	146	125	567	190	2349	255	6789	320	14789	385	123459	450	235689		
61	147	126	678	191	2356	256	12345	321	15678	386	123467	451	235789		
62	148	127	679	192	2357	257	12346	322	15679	387	123468	452	236789		
63	149	128	689	193	2358	258	12347	323	15689	388	123469	453	245678		
64	156	129	789	194	2359	259	12348	324	12356	389	123478	454	123489		
65	157	130	1234	195	2367	260	12349	325	12357	390	123479	455	123567		

decreased the initial variance of the system, and the mean similarities concentrated near 0.76. For the system analyzed, partitioning into seven clusters ($k = 1, \ldots, 7$) would be sufficient, reducing the time spent in the simulations, which was 360 hours, for the 511 experiments, using seven Pentium IV computers.

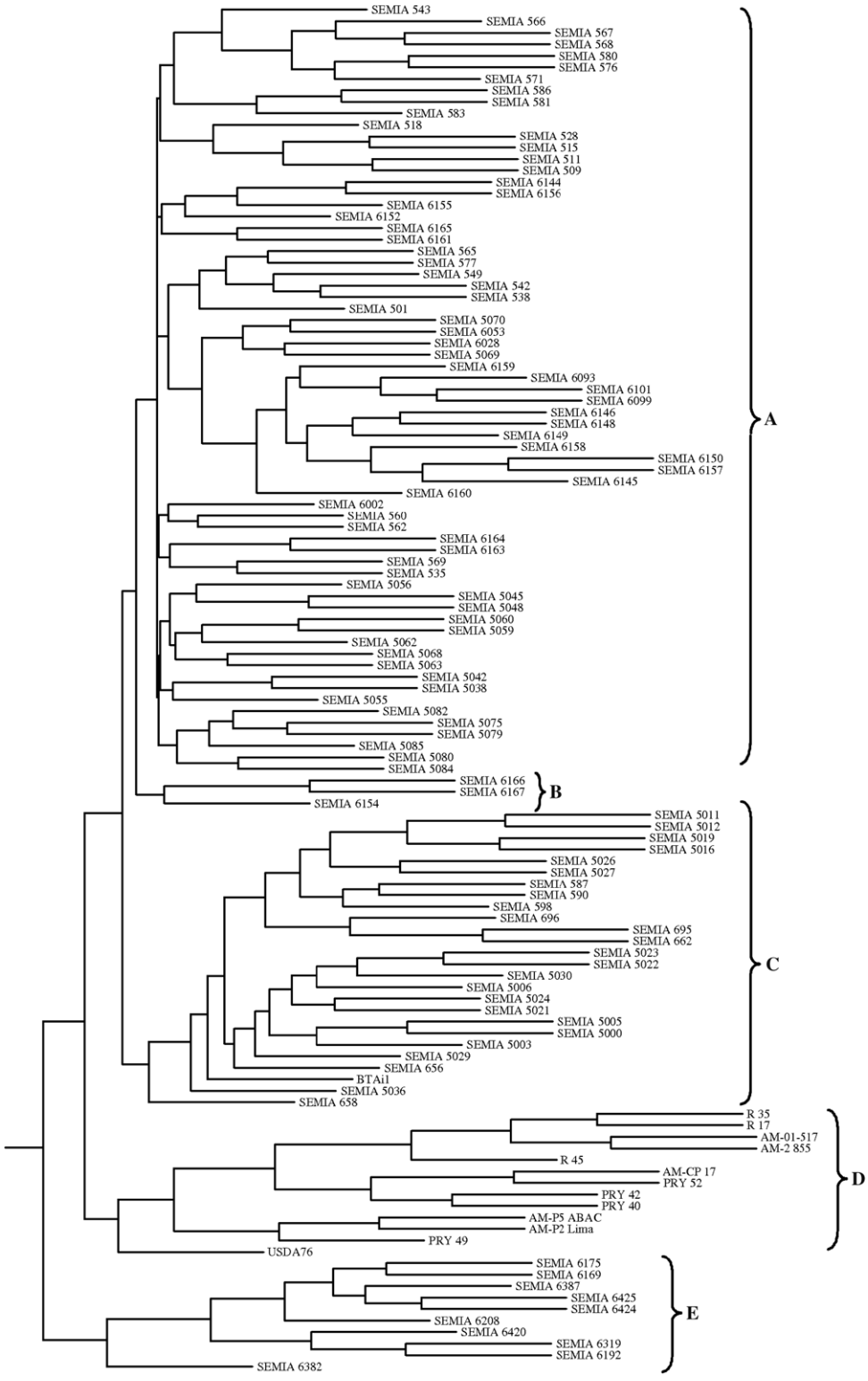

Fig. 3. Consensus tree for three stable clusters ($k = 3$).

Nine enzymes were insufficient to stabilize the system. Enzymes 1 (*Cfo*I 16S) and 2 (*Dde*I 16S) increased the similarities of the experiments and therefore the stability of the clusters. Enzyme 3 (*Dde*I IGS) increased the similarities of experiments when associated with a high number of enzymes, seven and eight, and decreased the similarities of the experiments with

Fig. 4. Consensus tree for four stable clusters ($k = 4$).

four enzymes. Enzyme 4 (*Hae*III IGS) decreased the similarities of the experiments. As expected, the highly conserved ribosomal 16S region was very important for the cluster stability, while the variability of the IGS reduced the experiment's stability.

Fig. 5. Consensus tree for five stable clusters ($k = 5$).

The method in this study is based on the images of electrophoresis gels and no restriction is made in relation to the strains used or number of strains; thus it can be applied to others strains by adjusting some parameters such as number of stable clusters (K) and similarity coefficient (in this work we used > 0.65).

Fig. 6. Consensus tree for six stable clusters ($k = 6$).

Another consideration in relation to the use of the electrophoresis gels is that currently there is no classification for defining gel quality, so even low-quality gels were considered, affecting the final precision of the results. Certainly, the utilization of high-quality image gels will generate results that are more accurate.

Table 2
List of strains from the cluster D for the four consensus trees (Figs. 3–6)

Number	*Strain*	Origin of Nodule/strain
1	R35	Brazil
2	R17	Brazil
3	AM-01-517	Brazil
4	AM-2-855	Brazil
5	R-45	Brazil
6	AM-P5 Abac	Brazil
7	AM-P2 Lima	Brazil
8	AM-CP 17	Brazil
9	PRY-42	Paraguay
10	PRY-49	Paraguay
11	PRY-40	Paraguay
12	PRY 52	Paraguay
13	USDA76	USA

Table 3
List of strains from the cluster E for the four consensus trees (Figs. 3–6)

Number	*Strain*	Origin of Nodule/strain
1	SEMIA 6175	Brazil
2	SEMIA 6169	Brazil
3	SEMIA 6387	Brazil
4	SEMIA 6425	Brazil
5	SEMIA 6424	Brazil
6	SEMIA 6192	Brazil
7	SEMIA 6420	Brazil
8	SEMIA 6382	Brazil
9	SEMIA 6319	Bolivia
10	SEMIA 6208	Colombia

The method presents an important characteristic that is the reproducibility of results because the analysis was made without individual interpretation.

References

[1] Å Aakra, J.B. Utåker, A.P. Röser, H.P. Koops, I.F. Nes, Detailed phylogeny of ammonia-oxidizing bacteria determined by rDNA sequences and DNA homology values, International Journal of Systematic and Evolutionary Microbiology 51 (2001) 2021–2030.
[2] R.C. Abaidoo, H.H. Keyser, P.W. Singleton, D. Borthakur, 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 and Evolutionary Microbiology 50 (2000) 225–234.
[3] A. Ben-Hur, I. Guyon, Detecting stable clusters using principal component analysis, in: M.J. Brownstein, A. Kohodursky (Eds.), Methods in Molecular Biology, Humana press, Clifton, 2003, pp. 159–182.
[4] A. Ben-Hur, A. Elisseeff, I. Guyon, A stability based method for discovering structure in clustered data, in: R. Altman, A. Dunker, L. Hunter, K. Lauderdale, T. Klein (Eds.), Pacific Symposium on Biocomputing, World Scientific, Hawaii, 2002, pp. 6–17.
[5] B. Everitt, Cluster Analysis, Halsted Press, New York, 1993.
[6] C.M. Fauquet, S. Sawyer, A.M. Idris, J.K. Brown, Sequence analysis and classification of apparent recombinant begomoviruses infecting tomato in the Nile and Mediterranean basins, Phytopathology 95 (2005) 549–555.
[7] J. Felsenstein, Software PHYLIP, Phylogeny Inference Package, v. 3.6, Department of Genome Sciences, University of Washington, 2002.
[8] FEPAGRO, Culture Collection Catalogue, 8th ed., Porto Alegre: Fundação Estadual de Pesquisa Agropecuária, 1999.
[9] M.C. Ferreira, M. Hungria, Recovery of soybean an inoculant strains from uncropped soils in Brazil, Field Crops Resources 79 (2002) 139–152.
[10] G.M. Garrity, D.R. Boone, B.W. Castenholz (Eds.), Bergey's Manual of Systematic Bacteriology, 1, 2nd ed., The Williams & Wilkins, New York, 2001.
[11] M.G. Germano, P. Menna, F.L. Mostasso, M. Hungria, RFLP analysis of the rRNA operon of a Brazilian of bradyrhizobial strains from 33 legume species, Journal of Systematic and Evolutionary Microbiology 56 (2006) 217–229.
[12] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, Journal of Intelligent Information Systems 17 (2–3) (2001) 107–145.
[13] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: A review, ACM Computational Survey 31 (3) (1999) 264–323.
[14] A. Jarabo-Lorenzo, E. Velázquez, R. Pérez-Galdona, M.C. Veja-Hernández, E. Martínez-Molina, P.F. Mateos, P. Vinuesa, E. Martínez-Romero, M. León-Barrios, 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 23 (2000) 418–425.
[15] K.S. Ko, J.W. Kim, J.M. Kim, W. Kim, S. Chung, I.J. Kim, Y.H. Kook, Population structure of the Bacillus cereus group as determine by sequence analysis of six housekeeping genes and the plcR gene, Infection and Immunity 72 (2004) 5253–5261.
[16] G. Laguerre, P. Mavingui, M.R. Allard, M.P. Charnay, P. Louvrier, S.I. Mazurier, L. Rigottier-Gois, N.N. Amarger, Typing of rhizobia by PCR and PCR-restriction fragment length polymorphism analysis of chromosomal and symbiotic gene regions: Application to Rhizobium leguminosarum and its different biovars, Applied Environmental Microbiology 62 (1996) 2029–2036.
[17] M.H. Law, A.K. Jain, 2002, Cluster validity by bootstrapping partitions, Technical Report MSU-CSE-03-5. [on line][cited in 23/09/2006]. Available in URL: http://www.cse.msu.edu/cgi-user/web/tech/document?ID=529.
[18] E. Levine, E. Domany, Resampling method for unsupervised estimation of cluster validity, Neural Computation 13 (2001) 2573–2593.
[19] W. Ludwig, K.H. Schleifer, Bacterial phylogeny based on 16S and 23S rRNA sequence analysis, FEMS Microbiological Review 15 (1994) 155–173.
[20] F.N. Martin, P.W. Tooley, Phylogenetic relationships among Phytophthora species inferred from sequences analysis of mitochondrially encoded cytochrome oxidase I and II genes, Mycologia 95 (2003) 269–284.
[21] M. Meilă, Comparing clustering, UW Statistics Technical Report 418, 2003.
[22] OpenMosix Project, [on line][cited in 18/11/2006]. Available in URL: http://openmosix.sourceforge.net/.

[23] MPICH2, [on line][cited in 20/09/2005]. Available in URL: http://www-unix.mcs.anl.gov/mpi/mpich/.
[24] V. Roth, T. Lange, M. Braun, J.A. Buhmann, Resampling approach to cluster validation, in: H. Wolfgang, R. Bernd (Eds.), Computational Statistics, COMPSTAT, Physica-Verlag, Heidelberg, 2002, pp. 123–128.
[25] R. Shamir, R. Sharan, Algorithmic approaches to clustering gene expression Data, in: T. Jiang, T. Smith, Y. Xu, M.Q. Zhang (Eds.), Current Topics in Computational Biology, MIT Press, Massachusetts, 2002, pp. 269–300.
[26] P. van Berkum, J.J. Fuhrmann, Evolutionary relationships among the soybean bradyrhizobia reconstructed from 16S rRNA gene and internally transcribed spacer region sequence divergence, International Journal of Systematic Bacteriology 50 (2000) 2165–2172.
[27] A. van Ooyen, Theoretical Aspects of Pattern Analysis, in: L. Dijkshoom, K.J. Tower, M. Struelens (Eds.), New Approaches for Generation and Analysis of Microbial Fingerprint, Elsevier, Amsterdam, 2001, pp. 31–45.
[28] E.T. Wang, P. van Berkum, X.H. Sui, D. Beyene, W.X. Chen, E. Martínez-Romero, Diversity of rhizobia associated with Amorpha fruticosa from Chinese soils and description of Mesorhizobium amorphae sp, International Journal of Systematic Bacteriology 49 (1999) 51–65.
[29] W.G. Weisburg, S.M. Barns, D.A. Pelletie, D.J. Lane, 16S ribosomal DNA amplification for phylogenetic study, Journal of Bacteriology 173 (1991) 697–703.
[30] A. Willems, R. Coopman, M. Gillis, Comparison of sequence analysis of 16S- 23S rDNA spacer regions, AFLP analysis and DNA–DNA hybridizations in Bradyrhizobium, International Journal of Systematic Bacteriology 51 (2001) 623–632.

S.T. Milagre is a Computer Science Ph.D. candidate.

C.D. Maciel is Professor, Engineering School of São Carlos, University of São Paulo, Brazil. His interest in microbiological statistics studies began with analyses of RFLP data of soil bacteria. Specific methodological interests include natural clustering and applications, the bootstrap method, information theory and signal processing applied to biological studies.

A.A. Shinoda is Professor, Electrical Department, State University of São Paulo at Ilha Solteira, Brazil. His main interest is signal processing applied to biological studies.

M. Hungria is the Chief of lab at Soil Biotechnology Laboratory, Embrapa Soja, Londrina, Brazil. Her work involves nitrogenous fixing bacteria, biodiversity and molecular methods.