J. Ind. Soc. Agril. Statist.
61(1), 2007 : 51-65

Comparative Evaluation of Clustering Techniques for Establishing AFLP Based Genetic Relationship among Sugarcane Cultivars

Ramesh Kolluru, A.R. Rao, V.T. Prabhakaran, A. Selvi[1] and T. Mohapatra[2]
Indian Agricultural Statistics Research Institute, New Delhi
(Received : January, 2006)

SUMMARY

Knowledge about germplasm diversity and genetic relationships among breeding material could be an invaluable aid in crop improvement strategies. Genetic diversity refers to variations within the individual gene loci/among alleles of a gene, or gene combinations, between individual plants or between plant populations. Quite often DNA marker data along with cluster analysis are used to assess genetic diversity of crop germplasm. Choice of genetic distance measures and clustering methods are two major issues in cluster analysis. An attempt is made in this paper to identify a suitable clustering procedure, which could accurately classify sugarcane genotypes, when the AFLP marker data contain missing observations.

Key words : Cluster analysis, Fuzzy clustering, AFLP, Genetic diversity.

1. INTRODUCTION

In recent past, biologists and social scientists began to look for systematic ways to classify their data into homogeneous groups. The advent of computers has revolutionized the modern algorithm-based computing which *inter alia* includes data analysis and classificatory procedures. When we have to deal with large quantities of breeding materials and germplasm accessions used in crop improvement programmes, methods to classify and order genetic variability assume considerable significance. The use of established multivariate statistical algorithms is an important strategy for classifying germplasm, ordering variability for a large number of accessions, or analyzing genetic relationships among breeding materials. Among different algorithms, cluster analysis is most commonly employed and appears particularly useful. Many research workers have discussed different clustering methods based on Partitioning algorithms, Hierarchical algorithms and Projection algorithms. Also, appropriate choice of a genetic distance measure, on the basis of the type of the variable and the scale of measurement, is an important component in the analysis of genetic diversity among a set of genotypes. Although allele frequencies can be calculated for some of the molecular markers, the data is most widely employed to generate a binary matrix for statistical analysis. The commonly used measures of genetic distance (GD) using such binary data are (i) Jaccard's (1908) coefficient (GD_J), (ii) Kulczynski coefficient (GD_K), (iii) Modified Rogers (Rogers, 1972) distance (GD_{MR}) and (iv) Nei and Li's (1979) coefficient (GD_{NL}).

Many a time, molecular marker data contain missing observations due to ambiguity of presence or absence of marker band. Normally, in such situations either the marker information is removed from the analysis or imputed by 0's or 1's. In the former case, one may squander time and resources, whereas in the latter case the estimate of genetic distance becomes biased. In such situations sound statistical techniques need to be employed to impute the missing observations. In this paper, an effort has been made to identify a suitable clustering procedure along with distance measure, which

[1] *Sugarcane Breeding Institute, Coimbatore – 641 007*

[2] *National Research Centre on Plant Biotechnology, Indian Agricultural Research Institute, New Delhi-110 012*

can provide accurate clustering of sugarcane genotypes based on Amplified Fragment Length Polymorphism (AFLP) marker data containing missing observations.

Source of Data

Molecular data based on 1041 AFLP markers generated by Selvi *et al.* (2005) using 28 commercial sugarcane cultivars grown either in the tropical (14) or subtropical (14) regions of India, were provided by one of the co-authors (T. Mohapatra). This data set contained missing observations.

2. GENETIC DIVERSITY AND CLUSTERING ALGORITHMS

Knowledge of genetic variation and genetic relationship among genotypes is an important consideration for efficient rationalization and utilization of germplasm resources. Furthermore, it is useful in the adoption of optimal designs for plant breeding programmes involving the choice of genotypes to cross for developing new populations. Analysis of genetic diversity in germplasm collections can facilitate reliable classification of accessions and identification of subsets of core accessions with possible utility for specific breeding purposes. Among the various procedures used for classifying germplasm, cluster analysis is the most popular one and employs any of the three kinds of clustering algorithms, namely partitioning algorithms, hierarchical methods and projection techniques. For the sake of continuity and clarity of description, the clustering algorithms are discussed briefly in the following.

2.1 Partitioning Algorithms

The partitioning method classifies the data into k groups called clusters, which together satisfy the requirements of partition, i.e. each group must contain at least one object and each object must belong to exactly one group. The algorithm gives partition with as many clusters as specified by the user through k. Since all values of k do not lead to natural clustering, the algorithm is run several times with different values of k and the partition that appears best from the point of view of meaningful interpretation is chosen. In practice it is left to the computer to try all possible combinations and choose the one which is best relative to some numerical criterion. Two robust programs based on partitioning algorithms are Partitioning Around Medoids (PAM) and Fuzzy analysis.

PAM can be applied to objects that are metric measurements or to data which is a dissimilarity matrix. To obtain k clusters, the method selects k objects, called representative objects from the data set and then each remaining object is assigned to the nearest representative object to form clusters such that the average distance of the representative object to all the other objects in its cluster is minimal. The representative object is called the medoid of its cluster and the partitioning method, the k-medoid technique. In other words, medoid is that object of the cluster for which the average dissimilarity with all its companion objects is minimal. The k-medoid method is robust with respect to outliers and can deal with not only interval-scaled measurements but also general dissimilarity coefficients. The result of clustering is displayed as 'silhouettes' (Rousseeuw 1987) on a single diagram, allowing the user to know, which objects lay well within different clusters and which do not, i.e. the quality of the clusters. PAM does not depend on the order in which the objects are presented.

Fuzzy analysis or fuzzy clustering is a generalization of partitioning algorithms and is executed through a program called FANNY. It employs the "fuzziness" principle and avoids "hard" decisions. Instead of saying, "object 'a' belongs to cluster 1," it may state that "object 'a' belongs to cluster 1 with 90% probability; to cluster 2 with 5% probability and to cluster 3 with 5% probability". Its advantage over hard clustering is that it yields a bulk of detailed information on the structure of the data but in the process it accumulates a volume of output which is too much to handle. The algorithm to calculate the membership coefficients is quite different from other clustering methods and does not involve any representative objects. Fuzzy analysis provides an entire $n \times k$ matrix of membership coefficients that may be very hard to interpret because of its mere size. An object is assigned to the cluster with which it has the largest membership coefficient. FANNY yields same kind of graphical display as does PAM, so the two outputs can be compared. The fuzzy clustering technique involves the minimization of the objective function

$$\sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} u_{iv}^2 u_{jv}^2 d(i,j)}{2\sum_{j=1}^{n} u_{jv}^2}$$

subject to the constraints : $u_{iv} \geq 0$ for i = 1, 2, ..., n; v = 1, 2, ..., k and $\sum u_{iv} = 1$ for i = 1, 2, ..., n where d(i, j) is distance between objects i and j, u_{iv} is unknown

membership coefficient of object i for cluster v. In fuzzy clustering, it is possible to visualize a partition in the form of a two-dimensional plot of the entities/cultivars (through multidimensional scaling) on which the clusters are portrayed as ellipses, the plot being known as clusplot.

2.2 Hierarchical Methods

Hierarchical algorithms do not construct a single partition with k clusters but they deal with all values of k in the same run. That is, the partition with k = 1, is part of the output, and also the situation with k = n. In between, all values of k = 2,3,..,n-1 are covered in a kind of gradual transition. There are two kinds of hierarchical techniques: the agglomerative and the divisive. They construct their hierarchy in the opposite directions, possibly yielding quite different results. Agglomerative methods start when all objects are held separate. Then in each step two clusters are merged, until they form a single tree with several branches. On the other hand, divisive methods start when all objects are together and in each following step a cluster is split up, until there are n of them. There exist many agglomerative algorithms, which only differ in their definition of between-cluster dissimilarity. Of them the most commonly used methods are unweighted pair-group average method (UPGMA), single linkage, complete linkage and weighted linkage. The agglomerative and divisive methods appear to be twins because they can be run in the same way and they yield very similar output.

2.3 Projection Techniques

Projection techniques are the methods for displaying (transformed) multivariate data in low-dimensional space. The primary objective here is to fit the original data into a low-dimensional coordinate system such that any distortion caused by a reduction in dimensionality is minimized. Distortion generally refers to the similarities or dissimilarities (distances) among the original data points. Although Euclidean distance may be used to measure the closeness of points in the final low-dimensional configuration, the notion of similarity or dissimilarity depends upon the underlying technique for its definition.

Suppose we have data on multiple characteristics of crop cultivars, to be utilized for classification of these cultivars. Principal Components Analysis (PCA) and Principal Coordinate Analysis (PCoA) are two techniques which can be used to reduce this multivariate data to two-dimensional data in terms of the first two PCA or PCoA axes in respect of different cultivars. The plot of the two axes of PCA or PCoA in XY plane will allow identification of different clusters formed by the cultivars. The third approach useful for the situation is the Multidimensional scaling technique. For details reference can be made to Johnson and Wichern (1993).

3. PROCEDURE FOR IDENTIFICATION OF THE BEST CLUSTERING METHOD

In the present study AFLP molecular marker data of sugarcane crop was considered, and different clustering techniques viz. four agglomerative hierarchical methods (Average, Single, Complete, Weighted) one divisive hierarchical method, two partitioning methods (partitioning around medoids, Fuzzy), three ordination techniques (principal component analysis, principal coordinate analysis and multidimensional scaling) were used for measuring genetic diversity. Here, every technique is used in combination with four distance measures like Jaccard, Kulczynski, Modified Rogers, and Nei and Li. The missing observations present in the data are tackled by different methods viz. removing marker data where observations are missing, impute missing cells with zeros, impute missing cells with ones and impute missing observations by association method. Here, suitable clustering procedures along with distance measures are identified based on three criterion: (i) probability of incorrect classification, (ii) cophenetic correlation, in presence of missing observations, and (iii) the cluster plots.

The sugarcane cultivars, used in the present investigation, are listed in Table 1 along with information on their parentage, year of release and region of adaptation. In all, there are 14 subtropical varieties in one group and 14 subtropical varieties in another group. Further, within the tropical region, the two varieties, Co 8021 and Co 8371, have same male and female parents, thereby forming a subgroup within the main tropical group. It is clear that based on the region of adaptation the cultivars can be set into two groups. These two groups form two main clusters in relation to which the percentage of incorrect classification by a particular clustering method is worked out. The probabilities of incorrect classification are calculated for different combinations of clustering methods viz. agglomerative, complete linkage, single linkage, weighted linkage, divisive, PAM, Fuzzy, and distance measures viz. Jacard,

Table 1. The parentage, year of release and region of adaptation of the sugarcane cultivars used in the study

Sr. No.	Cultivars	Parents		Year of release	Code	Region of adaptation
		Female	Male			
1	Co 1148	P 4383	Co 301	1963	ST1	Subtropical
2	CoJ 64	Co 976	Co 617	1971	ST2	"
3	Co 1158	Co 421		1963	ST3	"
4	Bo 91	Bo 55	Bo 43	1978	ST4	"
5	Co 89003	Co 7314	Co 775	1998	ST5	"
6	CoLK 8102	Co 1158		1986	ST6	"
7	CoS 88230	Co 1148	Co 775	1991	ST7	"
8	CoS 8436	MS 68/47	Co 1148	1987	ST8	"
9	CoPant84211	Co 6806	Co 6912	1996	ST9	"
10	CoPant84211	Co 1148	Co 775	1998	ST10	"
11	Co 7717	Co 419	Co 775	1977	ST11	"
12	Co8347	Co 419	CoC 671	1983	ST12	"
13	Co 87263	Co 312	Co 6806	1994	ST13	"
14	Co87268	BO 91	Co 62399	1994	ST14	"
15	Co 85002	Co 62198		1997	T15	Tropical
16	Co 86010	Co 740	Co 7409	1996	T16	"
17	Co 86032	Co 62198	CoC 671	1994	T17	"
18	Co 87025	Co 7704	Co 62198	1994	T18	"
19	Co 86249	CoJ 64	CoA 7601	1997	T19	"
20	Co 8371	Co 740	Co 6806	1997	T20	"
21	Co 6304	Co 419	Co 453	1973	T21	"
22	CoC 671	Q 63	Co 775	1975	T22	"
23	Co 62175	Co 951	Co 419	1974	T23	"
24	Co 8021	Co 740	Co 6806	1986	T24	"
25	Co 740	P 3247	P 4775	1949	T25	"
26	Co 7219	Co 449	Co 658	1980	T26	"
27	Co 419	POJ 2878	Co 290	1933	T27	"
28	Co 7704	Co 740	Co 6806	1983	T28	"

Table 2. Percentage incorrect classification of sugarcane varieties in two different clusters (C1, C2) using different combinations of clustering methods and distance measures

		Remove missing observations		Replace missing observations by 0's		Replace missing observations by 1's		Replace missing observations by Imputed values	
		Cluster I	Cluster II	Cluster I	Cluster II	Cluster I	Cluster II	Cluster I	Cluster II
Agglomerative	Jaccard	0	85.71	0	92.86	0	92.86	0	85.71
Agglomerative	Kulczynski	0	92.86	0	92.86	0	85.71	0	85.71
Agglomerative	Modified Rogers	0	85.71	0	92.86	0	92.86	0	85.71
Agglomerative	Nei and Li	0	85.71	0	92.86	0	92.86	0	85.71
Complete Linkage	Jaccard	0	85.71	0	92.86	0	92.86	0	85.71
Complete Linkage	Kulczynski	0	92.86	0	92.86	0	85.71	0	85.71
Complete Linkage	Modified Rogers	0	85.71	0	92.86	0	92.86	0	85.71
Complete Linkage	Nei and Li	0	85.71	0	92.86	0	92.86	0	85.71
Single Linkage	Jaccard	0	85.71	0	92.86	0	92.86	0	85.71
Single Linkage	Kulczynski	0	92.86	0	92.86	0	85.71	0	85.71
Single Linkage	Modified Rogers	0	85.71	0	92.86	0	92.86	0	85.71
Single Linkage	Nei and Li	0	85.71	0	92.86	0	92.86	0	85.71
Weighted Linkage	Jaccard	0	85.71	0	92.86	0	92.86	0	85.71
Weighted Linkage	Kulczynski	0	92.86	0	92.86	0	85.71	0	85.71
Weighted Linkage	Modified Rogers	0	85.71	0	92.86	0	92.86	0	85.71
Weighted Linkage	Nei and Li	0	85.71	0	92.86	0	92.86	0	85.71
Divisive	Jaccard	0	85.71	0	92.86	0	7.14	0	85.71
Divisive	Kulczynski	0	42.86	0	92.86	0	85.71	0	85.71
Divisive	Modified Rogers	0	85.71	0	92.86	0	7.14	0	85.71
Divisive	Nei and Li	0	85.71	0	92.86	0	7.14	0	85.71
PAM	Jaccard	0	7.14	28.57	14.29	21.43	7.14	21.43	7.14
PAM	Kulczynski	0	7.14	28.57	14.29	21.43	7.14	21.43	7.14
PAM	Modified Rogers	0	7.14	28.57	14.29	21.43	7.14	21.43	7.14
PAM	Nei and Li	0	7.14	28.57	14.29	21.43	7.14	21.43	7.14
Fuzzy	Jaccard	7.14	7.14	7.14	7.14	7.14	7.14	7.14	7.14
Fuzzy	Kulczynski	7.14	7.14	7.14	7.14	7.14	7.14	7.14	7.14
Fuzzy	Modified Rogers	7.14	7.14	7.14	7.14	7.14	7.14	7.14	7.14
Fuzzy	Nei and Li	7.14	7.14	7.14	7.14	7.14	7.14	7.14	7.14

Kulczynski, Modified Rogers, Nei and Li. These probabilities are worked out separately for different situations of handling missing observations, namely removing missing observations, replacing them by 0s, replacing by 1s, imputing them by association method and are presented in Table 2. The cophenetic correlations, measuring the product moment correlation between the dissimilarity/ similarity indicated by phenogram/ dendrogram and the dissimilarity matrix, under various situations, are presented in the Table 3. The

Table 3. Cophenetic correlation of hierarchical methods for different distance measures in sugarcane cultivars

Clustering Method	Distance Measure	Remove missing observations	Replace missing observations by 0's	Replace missing observations by 1's	Replace missing observations by imputation
UPGMA	Jaccard	0.841	0.856	0.865	0.957
UPGMA	Kulczynski	0.826	0.844	0.848	0.947
UPGMA	Modified Rogers	0.845	0.859	0.867	0.955
UPGMA	Nei and Li	0.838	0.856	0.863	0.962
Complete Linkage	Jaccard	0.841	0.856	0.865	0.957
Complete Linkage	Kulczynski	0.826	0.844	0.848	0.947
Complete Linkage	Modified Rogers	0.845	0.859	0.867	0.955
Complete Linkage	Nei and Li	0.838	0.856	0.863	0.962
Single Linkage	Jaccard	0.841	0.856	0.865	0.957
Single Linkage	Kulczynski	0.826	0.844	0.848	0.947
Single Linkage	Modified Rogers	0.845	0.859	0.867	0.955
Single Linkage	Nei and Li	0.838	0.856	0.863	0.962
Weighted Linkage	Jaccard	0.841	0.856	0.865	0.957
Weighted Linkage	Kulczynski	0.826	0.844	0.848	0.947
Weighted Linkage	Modified Rogers	0.845	0.859	0.867	0.955
Weighted Linkage	Nei and Li	0.838	0.856	0.863	0.962
Divisive	Jaccard	0.711	0.800	0.818	0.901
Divisive	Kulczynski	0.700	0.718	0.802	0.902
Divisive	Modified Rogers	0.721	0.805	0.824	0.898
Divisive	Nei and Li	0.698	0.796	0.816	0.913

non-hierarchical clustering methods do not involve the construction of dendrogram or trees and so the cophenetic correlations are not computable for these cases. Instead, the average silhouette width, an indicator of good quality clusters, is computed for the two clustering methods PAM and Fuzzy analysis, for each mode of handling missing observations, and for each distance measure. The results are given in Table 4. The dendrogram for different agglomerative hierarchical procedures are given in Figs. 1-3, and the cluspots of Fuzzy analysis in Fig. 4. The plots based on PCoA, multidimensional scaling and PCA are given in Figs. 5 to 7.

Table 4. Average silhouette width of partition around medoids (PAM) and fuzzy cluster analysis by using different distance measures for sugarcane

Clustering Method	Distance Measure	Remove missing observations	Replace missing observations by 0's	Replace missing observations by 1's	Replace missing observations by imputation
PAM	Jaccard	0.044	0.063	0.067	0.089
PAM	Kulczynski	0.044	0.063	0.067	0.089
PAM	Modified Rogers	0.044	0.063	0.067	0.089
PAM	Nei and Li	0.044	0.063	0.067	0.089
Fuzzy	Jaccard	0.079	0.084	0.086	0.088
Fuzzy	Kulczynski	0.079	0.084	0.086	0.088
Fuzzy	Modified Rogers	0.079	0.084	0.086	0.088
Fuzzy	Nei and Li	0.079	0.084	0.086	0.088

4. RESULTS AND DISCUSSION

From Table 2, it is evident that the partition methods have much smaller chance of incorrect classification (7 to 43%) than the hierarchical clustering methods (7 to 93%) under various modes of handling missing observations. The partitioning based on fuzzy analysis, however, shows much lesser chance of mis-classification (14%) than PAM method (28 to 35%). Among the distance based clustering methods the hierarchical methods are popular among the plant breeders. Results of this investigation show that, chance of incorrect classification from the use of hierarchical methods is

much lower (86%) when the missing values are imputed by association method, and the probability is the same for all the distance measures used. The use of Nei and Li, modified Rogers and Jaccard distance measures leads to comparatively smaller chances of mis-classification, when the missing values are either removed before analysis or they are replaced by unities. The replacement by zero does not seem to be the worst option to deal with missing observations.

From Table 3, it is seen that the cophenetic correlations are higher (0.90 to 0.96) when missing values are imputed by association method in comparison with other ways of dealing with missing values. Among the distance measures, Nei and Li measure has shown superiority over other measures, followed by Jaccard and Modified Rogers, which are almost at par. This further strengthens the view emerged from the incorrect classification probabilities that, when clustering is based on any hierarchical method, one should use Nei and Li or modified Rogers measure and the imputation should be through association method, for better results.

The average silhouette widths of PAM and Fuzzy clustering methods for different distance measures and different ways of handling missing observations, presented in Table 4 reveal that the values are higher for fuzzy clustering than PAM under different ways of handling missing observations, except association method, where the widths are at par. This once again point to the superiority of fuzzy clustering over PAM method.

The dendrograms, showing classification of varieties into clusters, under average linkage, complete linkage and single linkage are presented in Figs. 1-3. From these figures, it can be inferred that the chance of incorrect classification is reduced with the increase in inter-cluster distance. However, at sub-cluster level, it can be observed that all the sub-tropical varieties are coming under one group.

It is evident from the fuzzy clustering plots (CLUSPLOTs), for the association method (superior to other means of imputation) of handling missing observations (Fig. 4) that all the subtropical varieties (represented by squares) are grouped distinctly into one cluster and tropical varieties (represented by triangles) into the other cluster. Thus, the plots have depicted the correct position regarding the genetic relationships at the population level. The plots obtained from principal coordinate analysis, multi dimensional scaling and principal component analysis are given in Figs. 5-7. These figures also give more or less the same picture as emerged from CLUSPLOTs, but with a little ambiguity. It is observed that the variety Co 86249 (T19), which is a tropical variety, has been wrongly classified into subtropical group, whereas the variety Co 7717 (ST11), which is a subtropical variety, has fallen under tropical group. These two exceptions may be perhaps due to incorrect pedigree information of the cultivars.

According to Mohammadi and Prasanna (2003), hierarchical clustering methods in general and agglomerative hierarchical methods in particular are more commonly employed in the analysis of genetic diversity in crop species. Among the agglomerative hierarchical methods, the unweighted pair-group method using arithmetic averages, popularly known as UPGMA method (Sneath and Sokal 1973) is the most commonly adopted procedure followed by Wards minimum variance method (Ward 1963). Non-hierarchical methods like Fuzzy analysis and PAM are rarely used for analysis of intra-specific genetic variability in crop plants. The reason could be the lack of information about the optimal number of clusters that are needed for accurate classification. In this backdrop the present finding regarding the superiority of Fuzzy analysis over hierarchical methods is important and is expected to motivate the geneticists and plant breeders for using this method in their investigations on genetic diversity of crop plants.

CONCLUSION

Based on the performance of different clustering methods, distance measures and different ways of handling missing observations, the Fuzzy clustering method using either Nei and Li, modified Rogers or Jaccard distance measures can be safely recommended for clustering sugarcane cultivars. For better results, it is necessary that the missing observations are imputed by the association method. It is ideal if the results are crosschecked with those of PCoA.

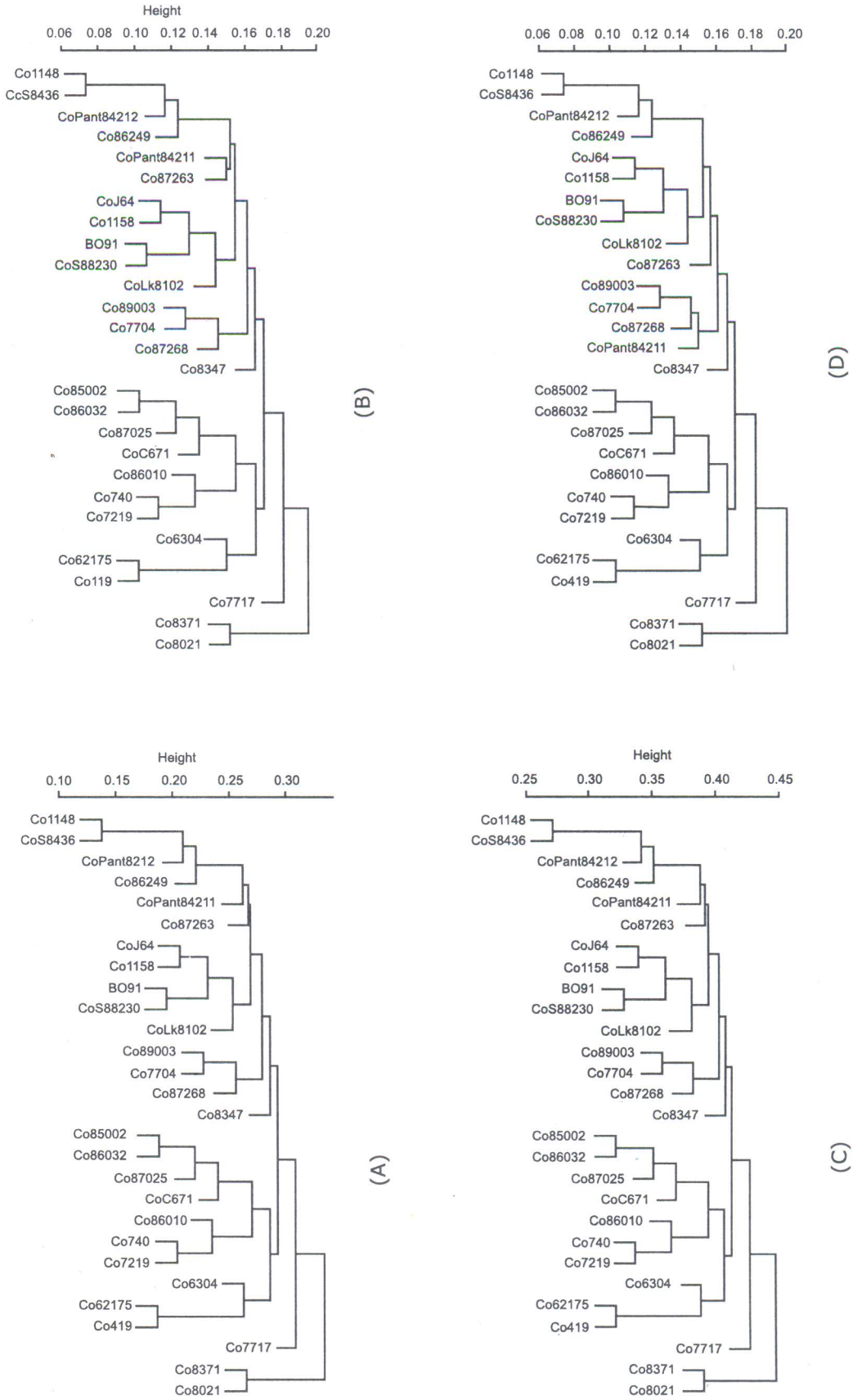

Fig. 1. Dendrograms of average linkage using different distance measures (A – Jaccard, B – Kulczynski, C – Modified Rogers, D – Nei & Li) for sugarcane cultivars where missing observations were imputed by association method.

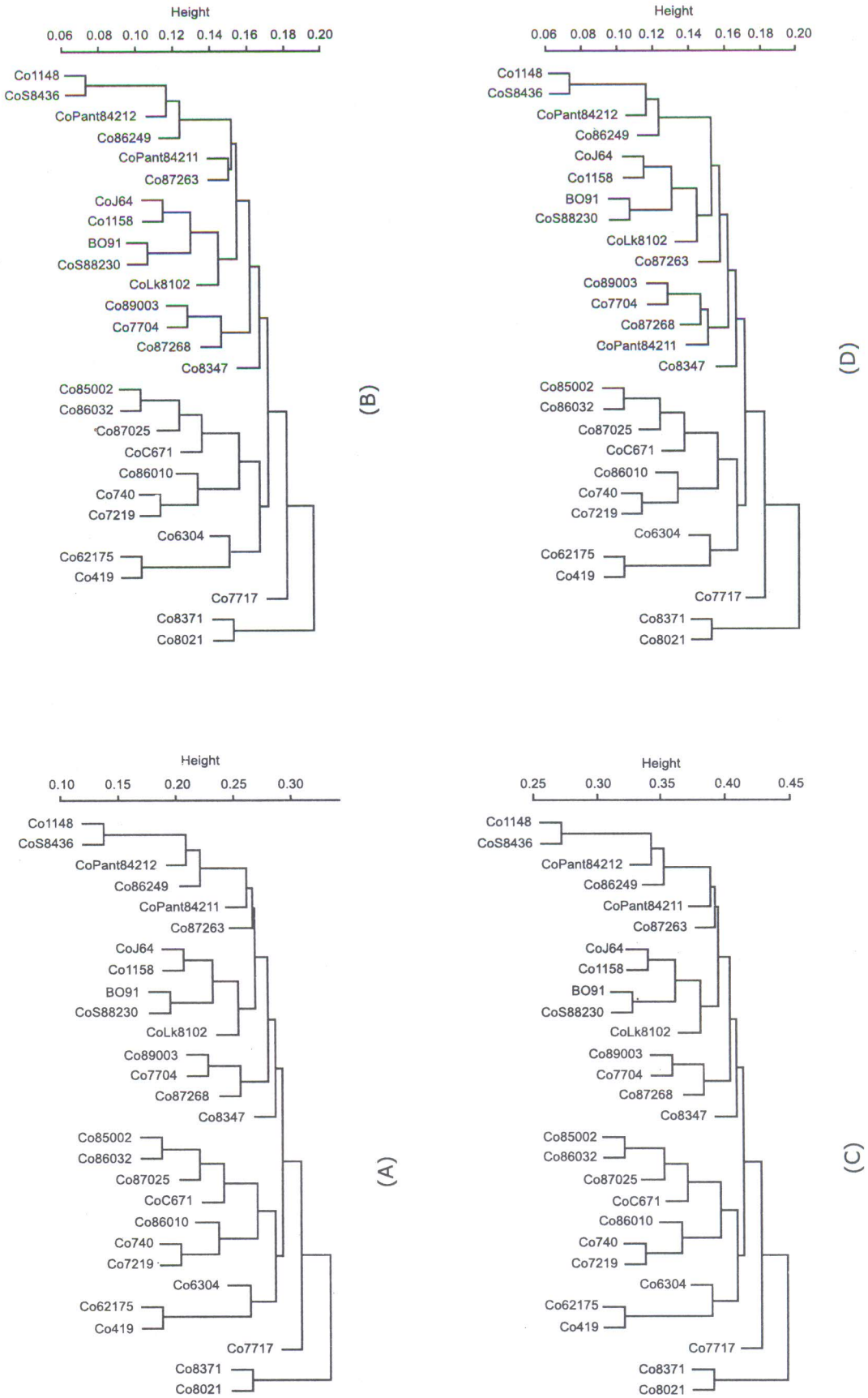

Fig. 2. Dendrograms of complete linkage method using different distance measures (A – Jaccard, B – Kulczynski, C – Modified Rogers, D – Nei & Li) for sugarcane cultivars where missing observations were imputed by association method.

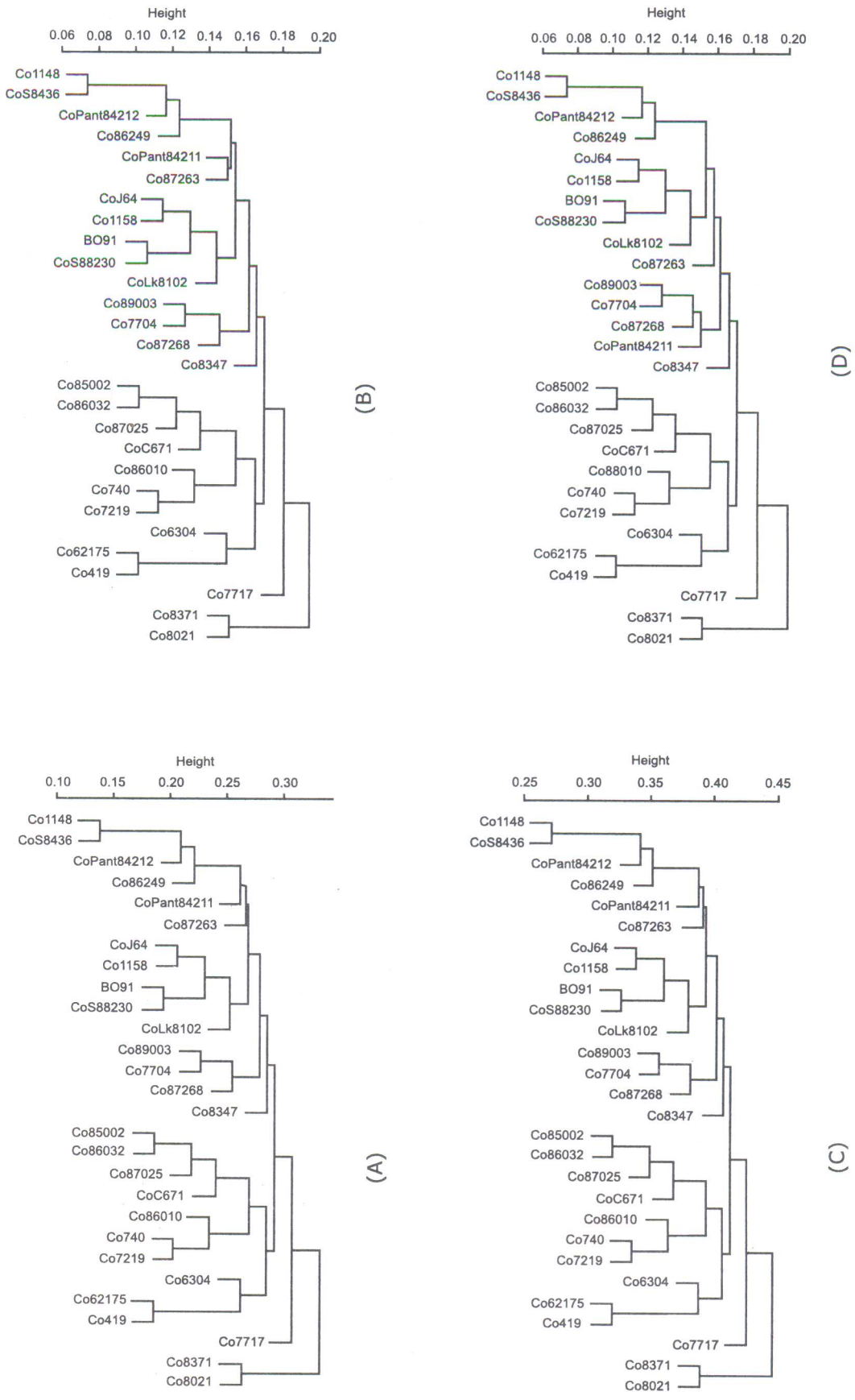

Fig. 3. Dendrograms of single linkage method using different distance measures (A – Jaccard, B – Kulczynski, C – Modified Rogers, D – Nei & Li) for sugarcane cultivars where missing observations were imputed by association method.

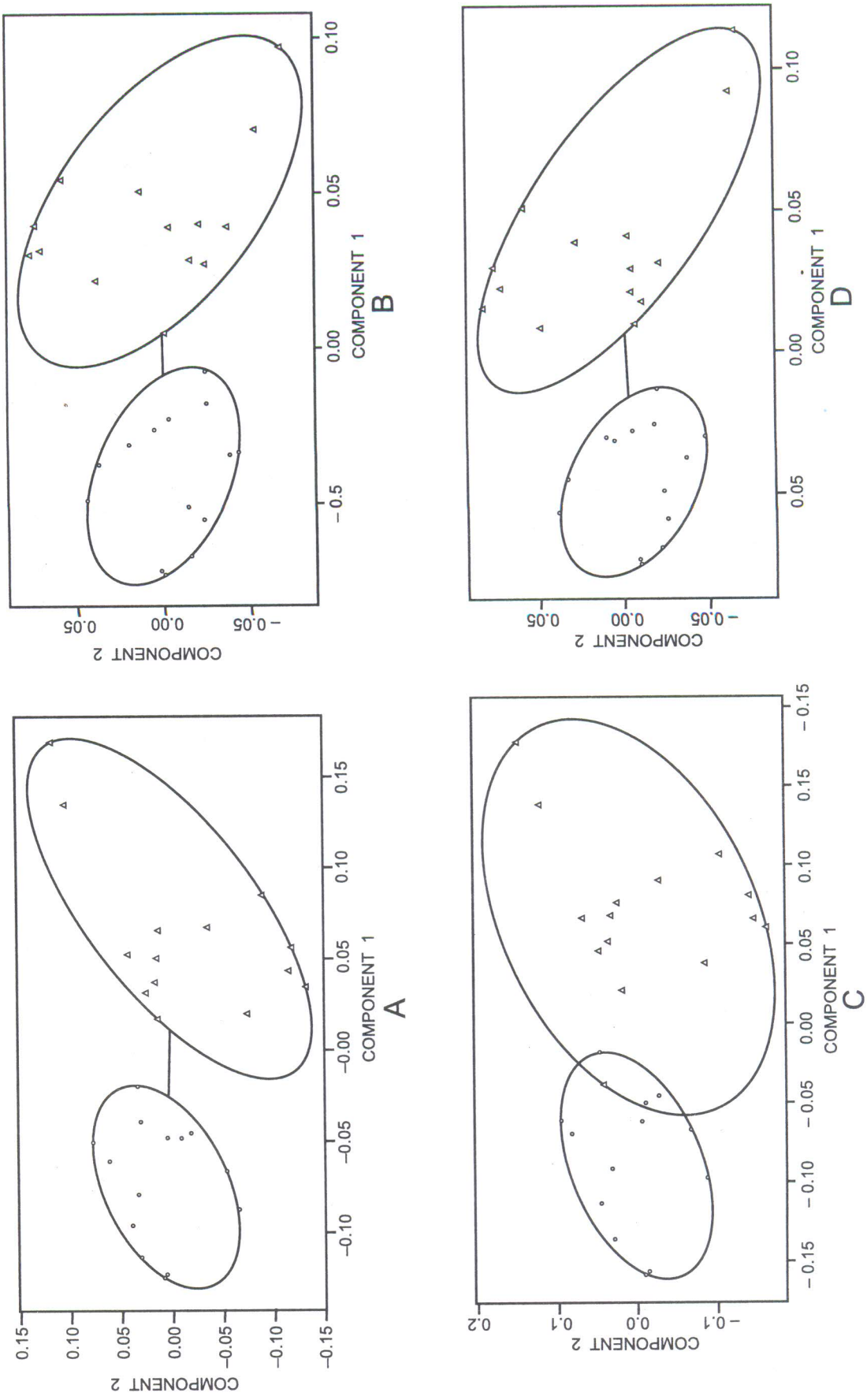

Fig. 4. Clusplots of Fuzzy clustering method using different distance measures (A – Jaccard, B – Kulczynski, C – Modified Rogers, D – Nei & Li) for sugarcane cultivars where missing observations were imputed by association method.

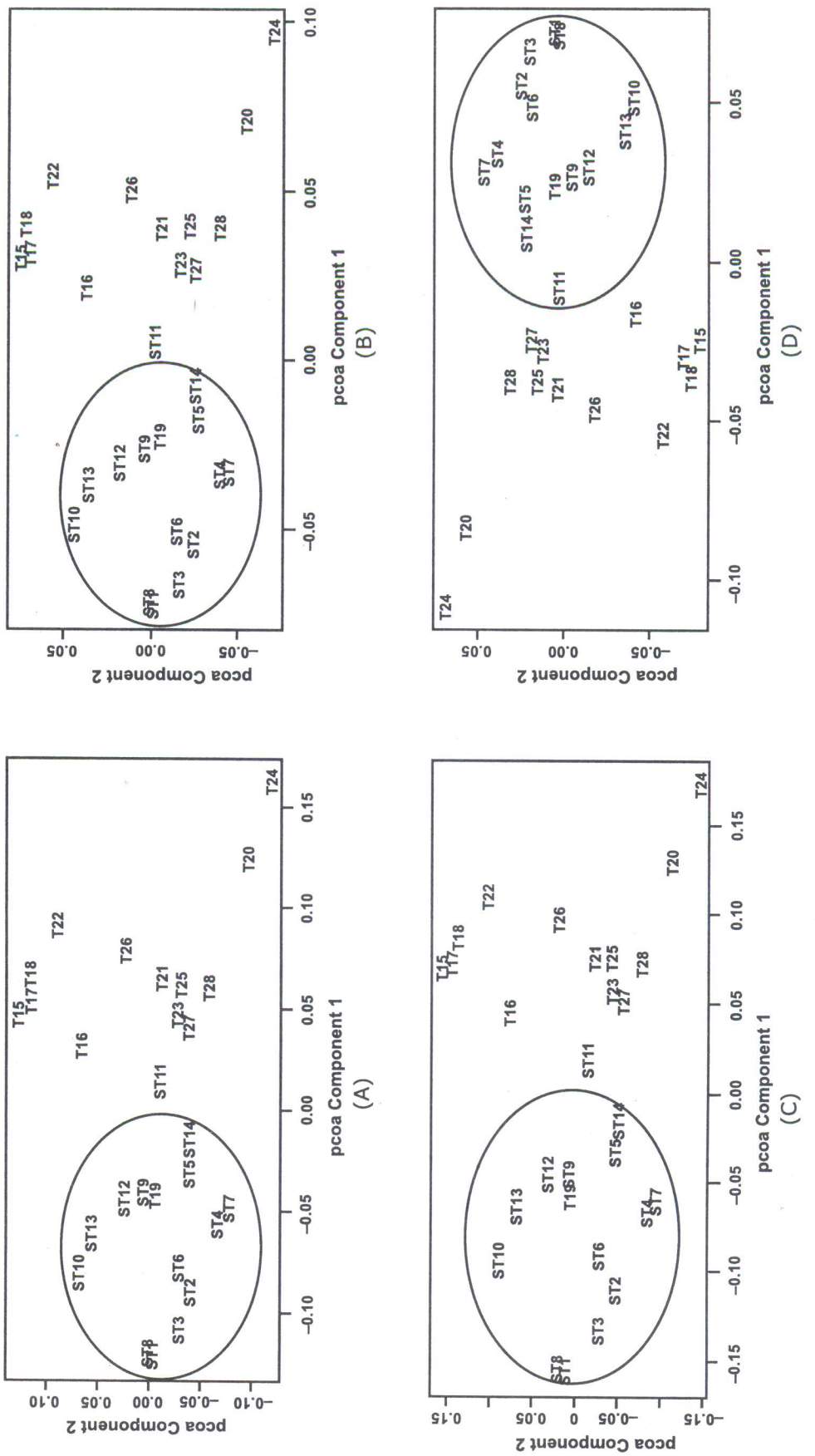

Fig. 5. Principal coordinate analysis method using different distance measures (A – Jaccard, B – Kulczynski, C – Modified Rogers, D- Nei & Li) for sugarcane cultivars where missing observations were imputed by association method.

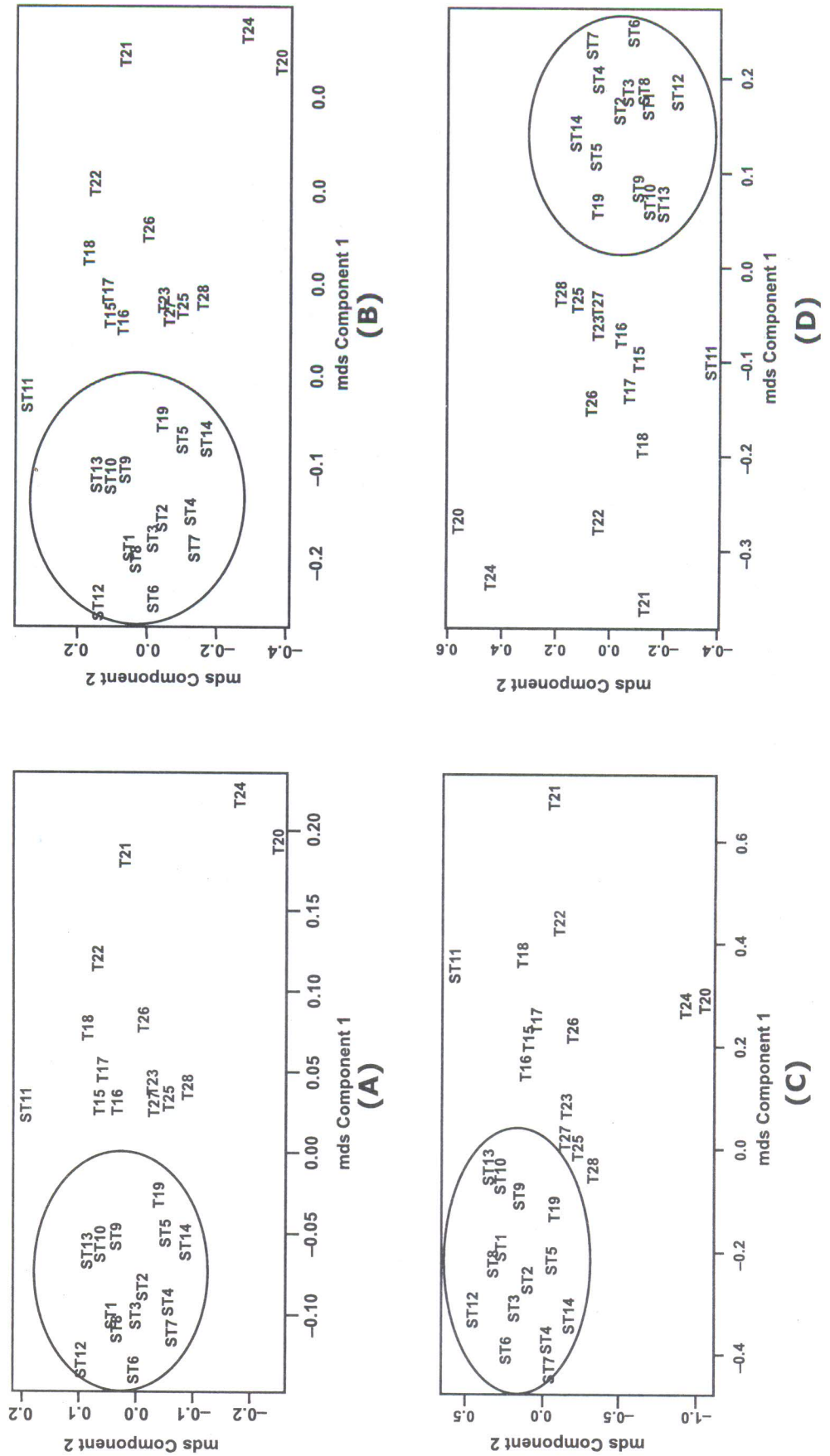

Fig. 6. Multidimensional scaling method using different distance measures (A – Jaccard, B – Kulczynski, C – Modified Rogers, D – Nei & Li) for sugarcane cultivars where missing observations were imputed by association method.

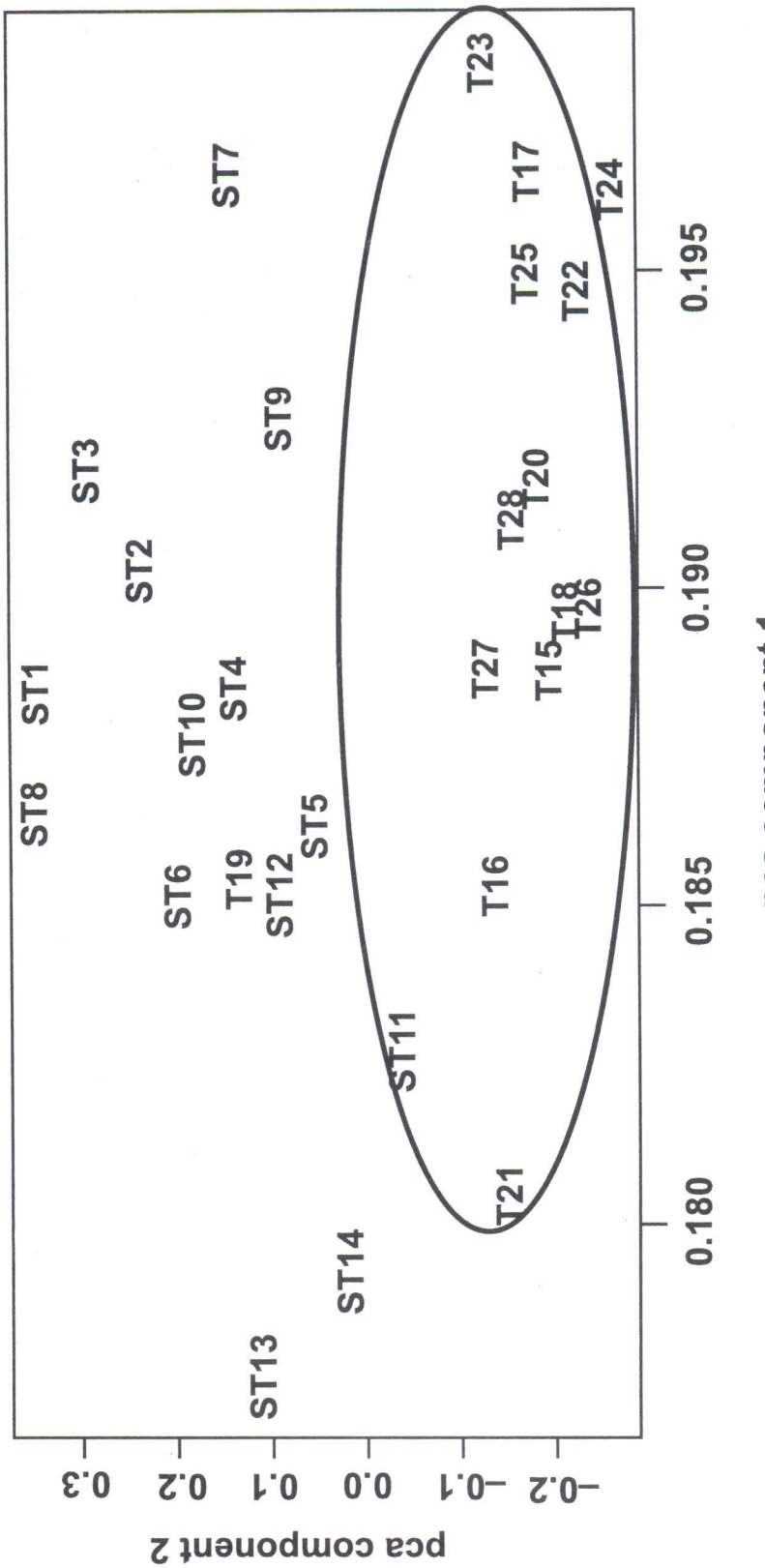

Fig. 7. Principal component analysis for sugarcane cultivars using missing observations imputed by association method.

REFERENCES

Jaccard, P. (1908). Nouvelles researches sur la distribution florale. *Bull. Soc. Vaudoise Sci. Natl.*, **44**, 223-270.

Johnson, A.R. and Wichern, D.W. (1993). *Applied Multivariate Statistical Analysis*. 3rd edition. Prentice-hall, Englewood cliffs, NJ.

Mohammadi, S.A. and Prasanna, B.M. (2003). Analysis of genetic diversity in crop plants-Salient statistical tools and considerations. *Crop Sci.*, **43**, 1235–1248.

Nei, M. and Li, W. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. (USA)*, **76**, 5269–5273.

Rogers, J.S. (1972). Measures of genetic similarity and genetic distance. *Studies in Genetics VII.* Univ. Tex. Publ., **2713**, 145-153.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation to cluster analysis. *J. Comput. Appl. Math.*, **20**, 53-65.

Selvi, A., Nair, N.V., Noyer, J.L., Singh, N.K., Balasundaram, N., Bansal, K.C., Koundal, K.R. and Mohapatra, T. (2005). Genomic constitution and genetic relationship among the tropical and subtropical Indian sugarcane cultivars revealed by AFLP. *Crop Sci.*, **45**, 1750–1757.

Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical Taxonomy*. Freeman, San Francisco.

Ward, J.H., Jr. (1963). Hierarchical grouping to optimize an objective function. *J. Amer. Stat. Assoc.*, **58**, 236-244.