Scientific Research and Essays Vol. 5(16), pp. 2318-2326, 18 August, 2010
Available online at http://www.academicjournals.org/sre
ISSN 1992-2248 ©2010 Academic Journals

Full Length Research Paper

Compare various combinations of similarity coefficients and clustering methods for *Olea europaea sativa*

M. Sesli[1*] and E. D. Yegenoglu[2]

[1]College of Tobacco Expertise, Celal Bayar University, Republic of Turkey, 45210 Akhisar, Manisa, Turkey.
[2]Akhisar Vocational College, Celal Bayar University, Republic of Turkey, 45210 Akhisar, Manisa, Turkey.

Accepted 3 June, 2010

The aim of the study was to compared the genetic similarity coefficients (Jaccard, Dice, Simple Matching) and different clustering methods (UPGMA, WPGMA, Single Linkage and Complete Linkage) combinations for cultivated olives. A total of 12 samples, Gemlik, Manzanilla, Edremit, Domat, Uslu and Memecik cultivars were screened with RAPD-PCR analysis by Operon random primers OP-Q kit. The closest samples based on their genetic similarity values have been found as 'Edremit and Gemlik 5' and the most distant ones have been found as 'Manzanilla 1 and Gemlik 3' in both of Dice and Jaccard coefficients; whereas, in simple-matching coefficient, the closest samples based on their genetic similarities have been determined as Edremit and Gemlik 5 and the most distant ones have been determined as Manzanilla 1 and Gemlik 3 and also Manzanilla 1 and Gemlik 4. The results from Mantel test of original matrices show that the correlation between Jaccard and Dice similarity matrices was high and significant (0.9971). UPGMA clustering for Dice coefficient was given a highest cophenetic correlation as 0.9571 and Complete linkage clustering for Simple Matching was yielded a lowest correlation value as 0.8992. The results obtained from consensus indices shown that Consensus fork index was found (CI_C =0, 9000) in Jaccard and Dice coefficients. Simple Matching coefficient had very low values with the Dice and Jaccard coefficients (CI_C =0.1000). PCO analysis provided results matching up one-to-one with the data obtained from Dice and Jaccard coefficient UPGMAs.

Key words: Clustering methods, genetic similarity coefficients, PCO, randomly amplified polymorphic DNA, olive.

INTRODUCTION

Cultivated olives (*Olea europaea sativa L)*, being produced in Turkey for thousands of years, is a product with significant economic value. It is preferred very much by the producers in order to benefit both from its oil and grains (Sesli et al., 2006).

Randomly amplified polymorphic DNA (RAPD) is used as genetic marker for determining the genetic diversity of cultivated olives (Belaj et al., 2001; Besnard et al., 2001; Wu et al., 2004).

RAPD technique is used successfully in the determination of genetic polymorphism and similarity of cultivated olives, and studies show that it yields positive results in determining redundancy of genetic diversity in olive trees and consequently the polymorphism of genetic markers (Martins-Lopes et al., 2007; Mekuria et al., 2002; Gemas et al., 2000).

*Corresponding author. E-mail: meltem.sesli@bayar.edu.tr. Tel: (+90) 236 412 68 96. Fax: (+90) 236 413 70 58.

Abbreviations: UPGMA, Unweighted pair group method with arithmetic; **WPGMA,** weighted pair-group method using arithmetic averages; **PCO,** principal co-ordinate; **RAPD,** random amplified polymorphic DNA; **PCR,** polymerase chain reaction.

Table 1. Provinces where cultivated olives were supplied.

Olives	Sample amounts	Place of supply	Province
Gemlik	5	Olive production research institute	Izmir, Bornova, Turkiye
Manzanilla	3	Olive production research institute	Izmir, Bornova, Turkiye
Domat	1	Olive production research institute	Izmir, Bornova, Turkiye
Memecik	1	Olive production research institute	Izmir, Bornova, Turkiye
Edremit	1	Sapling planters	Akhisar, Manisa, Turkiye
Uslu	1	Sapling planters	Akhisar, Manisa, Turkiye

Table 2. Correlation coefficients from Mantel test of original matrices.

	Simple matching	Jaccard	Dice
Simple matching	**********		
Jaccard	0.3587	**********	
Dice	0.3538	0.9971[1]	**********

[1]: Significant ($p < 0.05$).

The data obtained in conclusion to RAPD are compared with different similarity coefficients and clustering methods (Da Silva Meyer et al., 2004). Using similarity coefficients, for example Sorensen-Dice, Jaccard and Simple Matching, affects the results of Unweighted Pair Group Method with Arithmetic (UPGMA), Weighted Pair-Group Method using Arithmetic averages (WPGMA), Complete Linkage and Single Linkage clusters (Jackson et al., 1989; Duarte et al., 1999). Principal co-ordinate (PCO) analysis is used in the estimation of genetic similarity between plants (Baldoni et al., 2006).

In the study, the comparison of different similarity coefficients and clustering methods in cultivated olives was discussed.

MATERIALS AND METHODS

Plant material

The cultivated olives were transferred to the glasshouse; young leaves were collected and stored in liquid nitrogen until DNA extraction. A total of 12 samples were extracted by Doyle and Doyle method (1987) (Table 1).

DNA extraction and RAPD-PCR analysis

Twenty different decamer primers were used for RAPD analyses of *Olea europaea sativa.* A total of twenty primers from Kit OP-Q (Operon Technologies, Alameda, CA, USA) were used for RAPD-PCR analysis. PCR was performed on an Eppendorf MasterCycler Thermal Cycler in a total volume of 25 µl. PCR mix including 25 ng template DNA, 2.42 µl. 10 X PCR reaction buffer (with $MgCl_2$, Sigma), 0.44 µl. dNTP (Sigma), 1 µM primer, and 0.13 µl Taq DNA polymerase (Sigma). The amplification reactions were carried out for 60 s at 94°C as an initial denaturation. The PCR program comprised 35 cycles with 20 s at 94°C; 20 s at 35°C; 30 s at 72°C and a final extension performed at 72°C for 5 min.

Amplification products were loaded onto 1.5% agarose gels (Sigma) in 0.5 X TBE buffer with 0.5 µg/ml ethidium bromide at a constant 100 V. For evaluating the base pair length of bands, a DNA ladder (Sigma, Fermentas) was loaded on the first lane of each gel. After the separation of PCR products by agarose gel electrophoresis, gels were visualized with the Photo Print (Vilber Lourmat, France) imaging system and analyzed by BioOne D++ software (Vilber Lourmat, France).

Data analysis

Gels were visualized with Photo Print (Vilber Lourmat, France) imaging system and analysis of RAPD bands were performed by BioOne D++ software (Vilber Lourmat, France). The RAPD bands (markers) were scored as 1 if present and 0 if absent.

Original dendrograms were developed through UPGMA, WPGMA, Single Linkage and Complete Linkage clustering methods over the original matrices calculated from Simple Matching, Dice and Jaccard similarity coefficients. In the calculation of dendrograms, NTSYS Pc 2.1 program was used. Clustering methods and similarity coefficients were tested by applying SIMQUAL, SAHN, TREE procedures in NTSYS Pc 2.1. As the clustering method in SAHN module, WARN was selected in the option of UPGMA, WPGMA, Single Linkage and Complete Linkage and Tie Method (Rohlf, 2000). Cophenetic correlation coefficients were calculated by using COPH and MXCOMP procedures for each combination.

Original matrices were compared by applying Mantel test (Mantel, 1967) in the option of MXCOMP in NTSYS Pc 2.1 program for the comparison of original matrices by implementing Simple Matching, Dice and Jaccard similarity coefficients. The correlation values obtained are shown in Table 2 (Rohlf, 2000).

Mantel test was applied in NYSYS Pc 2.01 program through MXCOMP procedure in the comparison of original matrices and the matrices obtained from cophenetic values. Table 2 shows the correlation coefficients of data from original similarity matrices and cophenetic values as calculated with the Mantel test. The correlation coefficients calculated with the Mantel test enables the finding of correlation between the similarity matrix and the phenetic trees obtained as a result of cluster analysis. The correlation matrices calculated show the goodness of fit of cluster analysis in accordance with the similarity matrix.

Relationships among cultivated olives were studied using a principal coordinates analysis (PCO) according to Jaccard similarity coefficient with FAMD 1.23 (Schlüter and Harris, 2006).

The dendrograms bases of different coefficients were compared by consensus fork index (CI_C). The CI_C index provides a relative estimate of the dendrogram similarities and was calculated using NTSYS pc 2.1 (Rohlf, 2000).

Table 3. Genetic similarity of cultivated olives and the cophenetic correlation coefficients of matrices obtained from cophenetic values.

Clustering/similarity	Simple matching	Jaccard	Dice
UPGMA	0.9193	0.9566	0.9571
WPGMA	0.9175	0.9556	0.9557
Single Linkage	0.9088	0.9217	0.9130
Complete Linkage	0.8992	0.9326	0.9383

Table 4. Consensus fork index among the dendrograms (UPGMA) produced by similarity coefficients among cultivated olives by RAPD marker.

	Jaccard	Dice	Simple matching
Jaccard	**********	0.9000[1]	0.1000
Dice		**********	0.1000
Simple matching			**********

1: Significant ($p<0.05$).

RESULTS AND DISCUSSION

Ten of the twenty primers were given scorable bands (OP-Q 7, OP-Q 8, OP-Q 9, OP-Q 10, OP-Q 14, OP-Q 15, OP-Q 16, OP-Q 17, OP-Q 19, OP-Q 20). In this research, total of 84 bands were derived from primer kit OP-Q. Maximum bands were observed from OP-Q 16 primer with 34 bands and minimum numbers of bands were observed from OP-Q 7 and OP-Q 8 with 3 bands.

The correlation values obtained from the comparison of original matrices by applying the Mantel test are shown in Table 2 (Rohlf, 2000).

The results from Mantel test of original matrices show that the correlation between Jaccard and Dice similarity matrices was high and significant (0.9971), however, correlations between Jaccard, Dice and Simple matching coefficients were very low to compared the Jaccard and Dice correlation. Table 3 shows the original similarity matrices and the correlation coefficients of data from cophenetic values as calculated with Mantel test.

Table 3 shows the cophenetic correlations from the genetic similarity matrices and cophenetic matrices.

UPGMA clustering for Dice coefficient was given a highest cophenetic correlation as 0.9571 and Complete linkage clustering for Simple Matching was yielded a lowest correlation value as 0, 8992.

Comparing the results of clustering method/similarity coefficients, Jaccard and Dice coefficients produced high cophenetic correlations; however, Simple Matching coefficients had lower cophenetic correlations. Because of their highest correlation rate, Dice similarity with UPGMA clustering method is evaluated as a convenient combination for detecting the genetic relationships between cultivated olives.

Since simple matching coefficient and complete linkage clustering method had a lowest correlation value, bear in mind that the combination is an inconvenient composition for detecting the genetic relationships between cultivated olives.

Koopman et al. (2001) found that Jaccard similarity with UPGMA yielded highest correlation rate for AFLP data set from *Lactuca*, s.l. species. In this study, Dice similarity with UPGMA yielded highest correlation value but the differences between Dice and Jaccard similarity coefficients with UPGMA clustering was not so far (0.9566 and 0.9571, respectively). Koopman et al. (2001) also reported that Dice coefficient with UPGMA clustering method was suitable for AFLP data set. Since, AFLP and RAPD markers have dominant characteristics; UPGMA with Dice coefficient was useful for genetic analysis of the cultivated olives.

For the evaluation of trees from UPGMA clustering with genetic similarity coefficients, consensus indices were also calculated for the each combination of coefficient and UPGMA clustering. The results from Consensus Indices were shown in the Table 4.

The results obtained from consensus indices shown that Consensus fork index was found (CI_C =0.9000) in Jaccard and Dice coefficients. Simple Matching coefficient had very low values with the Dice and Jaccard coefficients (CI_C =0.1000). This is in line with the findings from dendrograms and shows why obtained different results from the dendrogram formed with Simple Matching coefficient as compared to the others.

The closest samples based on their genetic similarity values have been found as ‘Edremit and Gemlik 5’ and the most distant ones have been found as ‘Manzanilla 1 and Gemlik 3’ in both of Dice and Jaccard coefficients. In the coefficient including negative co-occurrences such as simple-matching, the closest samples based on their genetic similarity values were determined as ‘Edremit and Gemlik 5’ and the most distant samples as ‘Manzanilla 1 and Gemlik 3’ and ‘Manzanilla 1 and Gemlik 4’ different from other coefficients.

The dendrograms derived from the combination of genetic similarity coefficients and different clustering methods were given in Figure 1, Figure 2, Figure 3 and Figure 4 and also, Figure 5 shows that the PCO analysis of cultivated olives.

In accordance with the result obtained with PCO (Figure 5), the proportions obtained from the X-axis, Y-axis and Z-axis were 16.57, 12.32 and 10.08%, respectively.

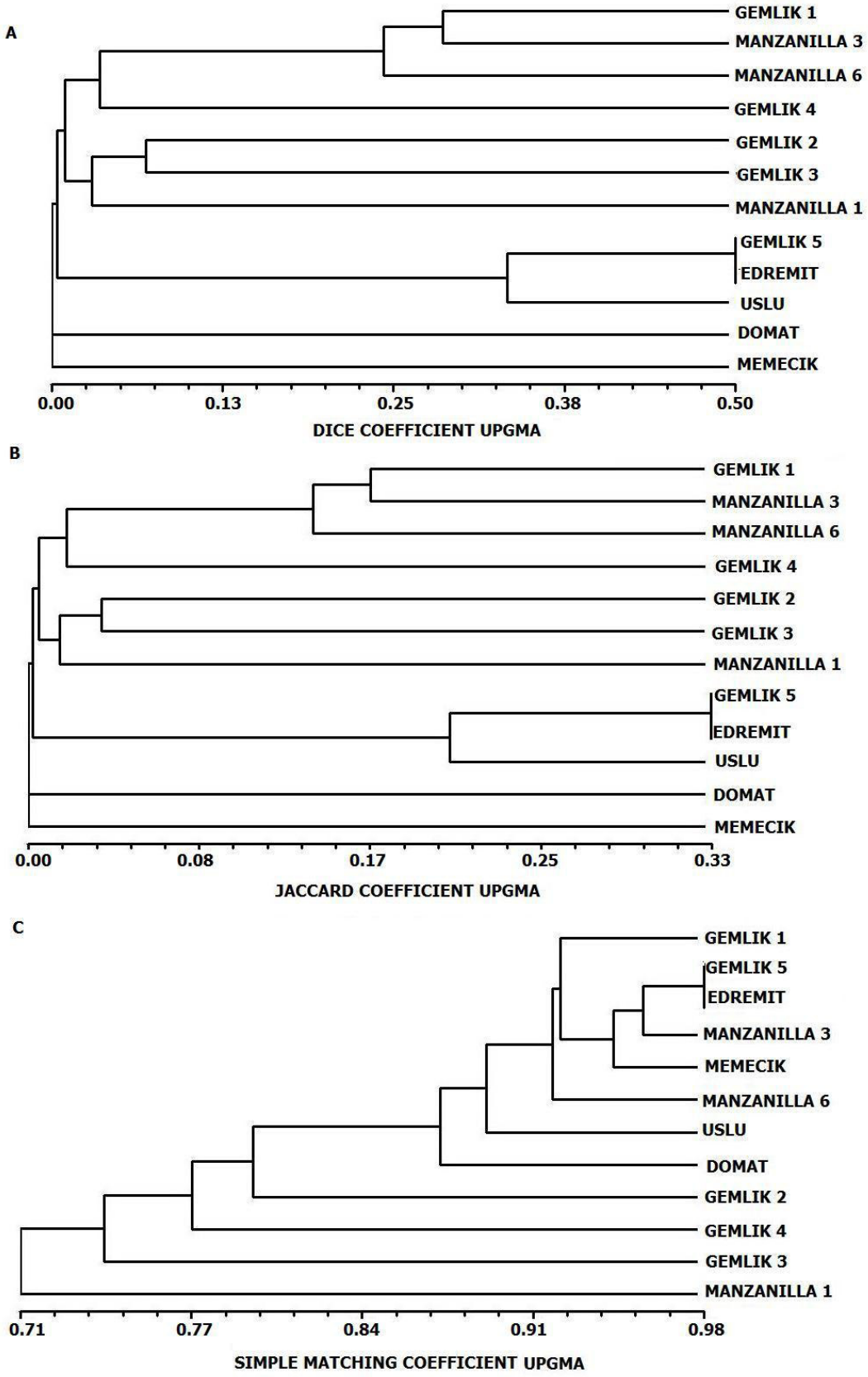

Figure 1. UPGMA dendrograms with different genetic similarity coefficients. (a) Dice coefficient UPGMA. (b) Jaccard coefficient UPGMA. (c) Simple matching coefficient UPGMA.

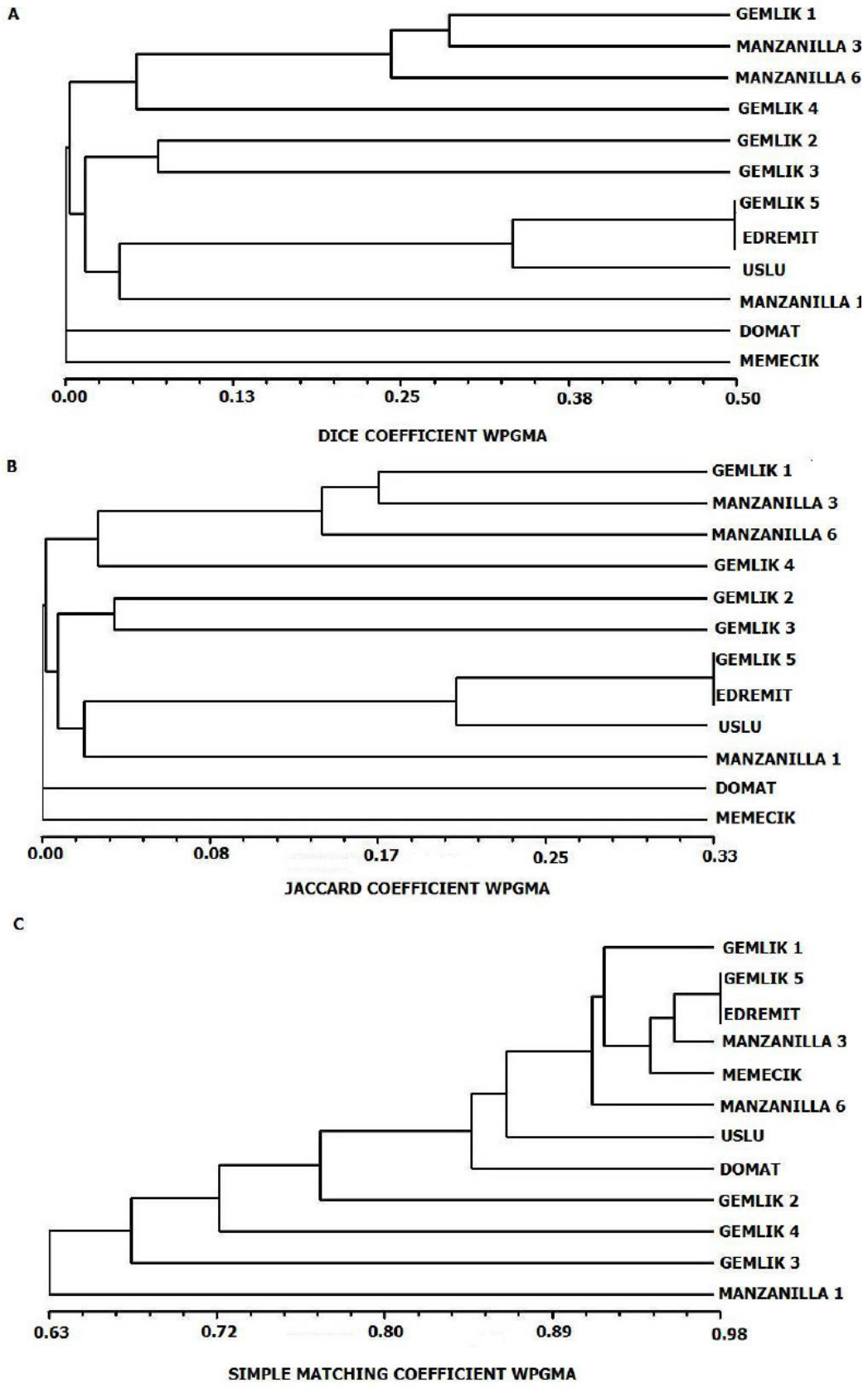

Figure 2. WPGMA dendrograms with different genetic similarity coefficients.(a) Dice coefficient WPGMA. (b) Jaccard coefficient WPGMA. (c) Simple matching coefficient WPGMA.

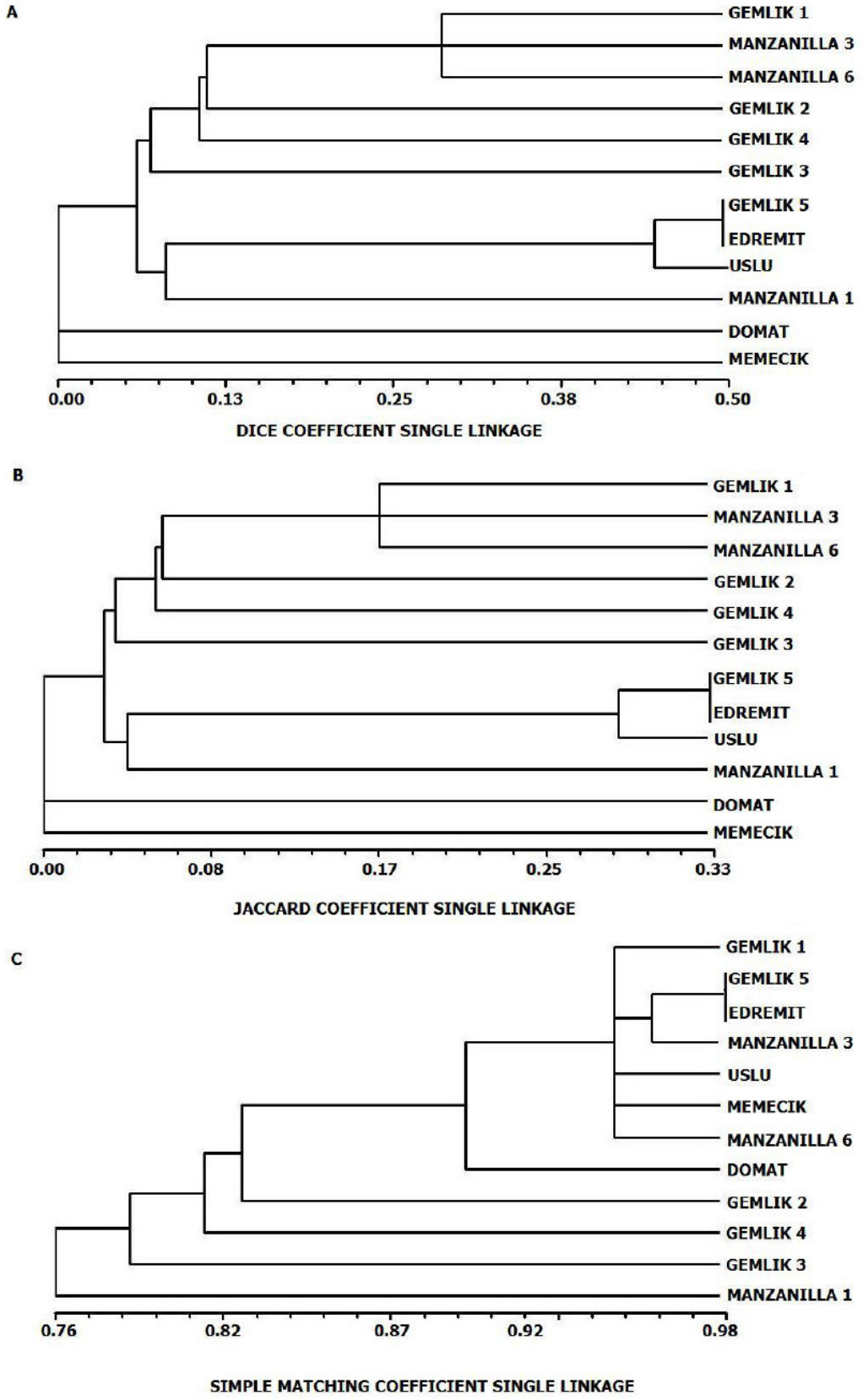

Figure 3. Single Linkage dendrograms with different genetic similarity coefficients. (a) Dice coefficient single linkage. (b) Jaccard coefficient single linkage. (c) Simple matching coefficient single linkage.

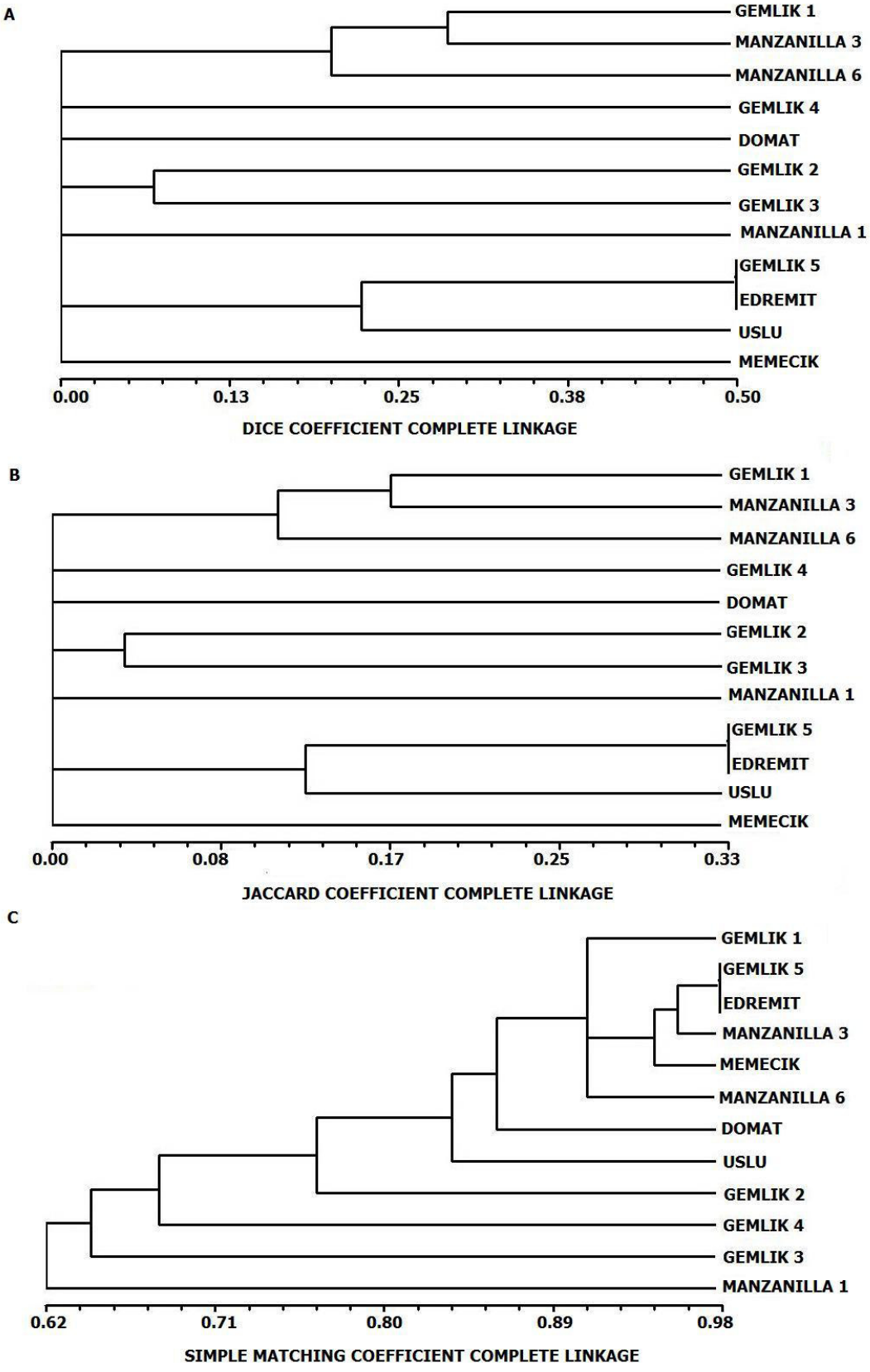

Figure 4. Complete linkages dendrograms with different genetic similarity coefficients. (a) Dice coefficient complete linkage. (b) Jaccard coefficient complete linkage. (c) Simple matching coefficient complete linkage.

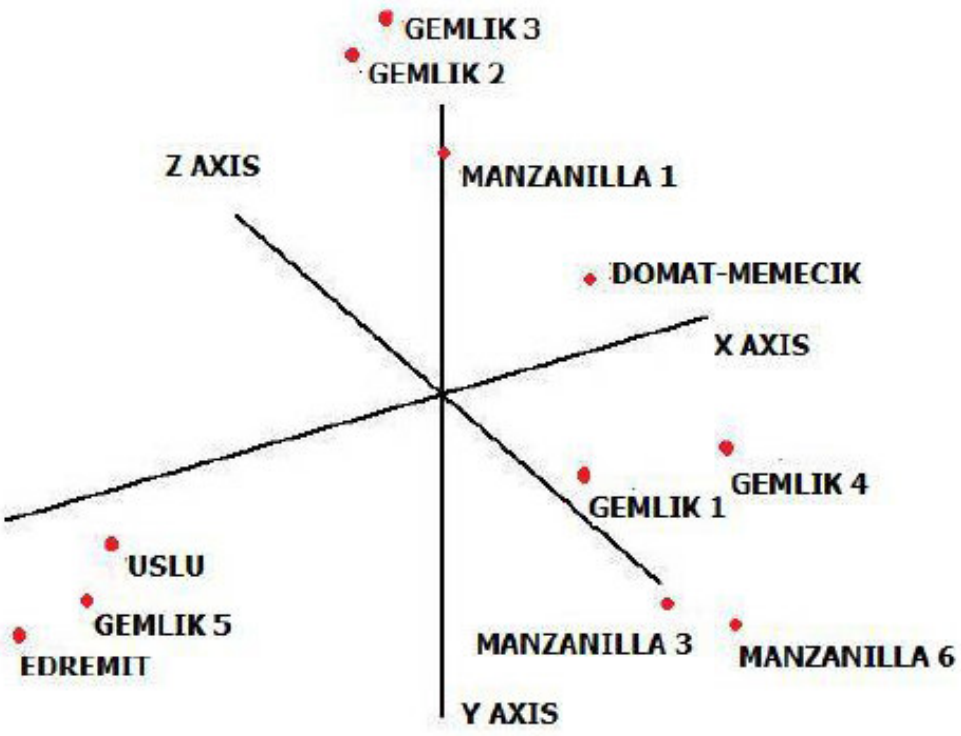

Figure 5. PCO analysis of cultivated olives.

A total of 38.97% was found. Accordingly, it was determined that Edremit and Gemlik 5 samples found as the closest ones based on their genetic similarity values are in the same plot. It is observed that the most distant samples based on their genetic similarity values, Manzanilla 1 and Gemlik 3 are in the same plot. When the data obtained from PCO and all clustering methods were compared, it was determined that the dendrograms obtained from Dice and Jaccard coefficients with UPGMA clustering method provided same results one-to-one with PCO.

In conclusion, it was suggested to use Dice and Jaccard genetic coefficients with UPGMA clustering method in the determination of genetic relations of cultivated olive because of the high correlation value they show; whereas, Simple Matching coefficient with Complete Linkage clustering method is not suggested because they show low correlation coefficient. In addition, it was concluded that Simple Matching coefficient is not suitable for the studies with RAPD since it causes change in the results due to negative co-occurrences.

REFERENCES

Baldoni L, Tosti N, Ricciolini C, Belaj A, Arcioni S, Panelli G, Germana MA, Mulas M, Porceddu A (2006). Genetic structure of wild and cultivated olives in the central Mediterranean Basin. Ann. Bot., 98(5): 935-942.

Belaj A, Trujillo I, De La Rosa R, Rallo L (2001). Polymorphism and discrimination capacity of randomly amplified polymorphic markers in an olive germplasm bank. J. Amer. Soc. Hort. Sci., 126(1): 64-71.

Besnard G, Baradat P, Bervillé A (2001). Genetic relationships in the olive (*Olea europaea* L.) reflects multilocal selection of cultivars. Theor. Appl. Genet., 102(2-3): 251-258.

Da Silva MA, Garcia A, Augusto F, De Souza AP, De Souza Jr. CL (2004). Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (*Zea mays* L). Genet. Mol. Biol., 27(1): 83-91.

Duarte JM, Dos Santos JB, Melo LC (1999). Comparison of similarity coefficients based on RAPD markers in the common bean. Genet. Mol. Biol., 22(3): 427-432.

Doyle JJ, Doyle JL (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull., 19: 11-15.

Gemas VJV, Rijo-Johansen MJ, Tenreiro R, Fevereiro P (2000). Inter- and intra- varietal analysis of three *Olea europaea* L. cultivars using the RAPD technique. J. Hortic. Sci. Biotech., 75: 312-319.

Jackson AD, Somers MK, Harvey HH (1989). Similarity coefficients: measures for co-occurrence and association or simply measures of occurrence?. Am. Natur., 133(3): 436- 453.

Koopman WJM, Zevenbergen MJ, Van Den Berg Ronald G (2001). Species relationships in *Lactuca* S.L. (*Lactuceae, Asteraceae*) inferred from AFLP fingerprints. Amer. J. Bot., 88(10): 1881-1887.

Mantel NA (1967). The detection of disease clustering and a generalized regression approach. Cancer Res., 27(2): 209-220.

Martins-Lopes P, Lima-Brito J, Gomes S, Meirinhios J, Santos L, Guedes-Pinto H (2007). RAPD and ISSR molecular markers in *Olea europaea* L.: Genetic variability and molecular cultivar identification. Genet. Resour. Crop Evol., 54(1): 117-128.

Mekuria TG, Collins G, Sedgley, M (2002). Genetic diversity within an isolated olive (*Olea europaea* L.) population in relation to feral spread. Sci. Hort., 94(1): 91-105.

Rohlf FJ (2000). *NTSYS-pc:* Numerical Taxonomy System. V*er. 2.1.* Exeter Software, Setauket, NY, USA. pp. 29-34.

Sesli M, Tokmakoglu A (2006). Olive Existence in Akhisar District in Manisa Province in Turkey, J. Appl. Sci., 6(13): 2849-2852.
Schlüter MP, Harris AS (2006). Analysis of multilocus fingerprinting data sets containing missing data. Mol. Ecol. Notes., 6(2): 569-572.
Wu SB, Collins G, Sedgley MA (2004). Molecular linkage map of olive (*Olea europaea* L.) based on RAPD, microsatellites and SCAR markers. Genome, 47(1): 26-35.