Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Comparison of similarity coefficients used for cluster analysis based on RAPD markers in wild olives

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

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

Corresponding author: M. Sesli
E-mail: meltem.sesli@bayar.edu.tr

Genet. Mol. Res. 9 (4): 2248-2253 (2010)
Received July 5, 2010
Accepted August 8, 2010
Published November 16, 2010
DOI 10.4238/vol9-4gmr966

ABSTRACT. Five different similarity coefficients (Jaccard, Sorensen-Dice, simple matching, Rogers and Tanimoto, and Russel and Rao) were evaluated and 10 wild olives analyzed with RAPD markers. The influence of the similarity coefficients on wild olives clustering was investigated. Forty-five primers were used on samples from 10 wild olives (Wild 1 and 2 obtained from Mugla province; Wild 3, 4, 5, 6, 7, and 8 from Manisa province and Wild 9 and 10 from Izmir province of Turkey). The similarity matrices obtained from RAPD markers were compared by the Mantel test. Cluster analysis was made with UPGMA dendrograms, and the consensus fork indexes between all pairs of dendrograms were calculated. The Jaccard and Sorensen-Dice coefficients gave the same results, due to the fact that both exclude negative co-occurrences. The dendrograms using the simple matching and Rogers and Tanimoto coefficients were similar; Wild 4 (Akhisar, Manisa) and Wild 9 (Bornova, Izmir) olives had the closest genetic similarities. This occurred because these coefficients include negative co-occurrences. The Russel and Rao coefficients produced different results, because they include negative co-occurrences in the denominator. We concluded that the coefficients that

do not include negative co-occurrences are more efficient for studies of wild olives clustering based on RAPD markers.

Key words: Wild olives; RAPD; PCR; Genetic similarity coefficients; UPGMA

INTRODUCTION

The results of studies performed with molecular markers are evaluated using different statistical methods. One such method is cluster analysis (Duarte et al., 1999); this method is preferred because it is easily explainable. A matrix should be formed in order to exhibit the similarities or distances between the genotypes before applying the method. Such matrices can be calculated with various coefficients (Weir, 1996); many different similarity coefficients are suggested for use in their determination (Johnson and Wichern, 1988). Such coefficients are calculated by double comparison based on the bands obtained from samples, scored as 0 or 1 (Skroch et al., 1992). The consequence of selecting different coefficients affects the clusters obtained (Gower and Legendre, 1986; Jackson et al., 1989). Meyer Da Silva et al. (2004) recommended the Jaccard, Sorensen-Dice, Anderberg, and Ochiai coefficients, because they are easy to explain and since they do not consider negative co-occurrences. Care should be taken in selecting the coefficients.

The aim of this study was to investigate the influence of the choice among five different similarity coefficients on unweighted pair-group method with arithmetic mean (UPGMA) cluster analysis, based on data taken from the dominant molecular marker analysis (random amplified polymorphic DNA, RAPD) of 10 wild olives.

MATERIAL AND METHODS

The wild olives used in this study were supplied from villages in Manisa, Izmir and Mugla provinces, and Wild 9 was supplied from the Olive Research Institute of Turkey. Fresh leaves were taken from a total of 10 wild olives and they were kept in liquid nitrogen until DNA extraction. Table 1 shows the wild olives used in this study and their origin.

Table 1. Provinces where wild olives were obtained.

Type of olive	Provinces
Wild 1	Pinarcik 1, Milas, Mugla, Turkey
Wild 2	Yatagan, Muğla, Turkey
Wild 3	Caglak 1, Akhisar, Manisa, Turkey
Wild 4	Harlak, Akhisar, Manisa, Turkey
Wild 5	Sabancilar 1, Akhisar, Manisa, Turkey
Wild 6	Haskoy, Akhisar, Manisa, Turkey
Wild 7	Yayakirildik 2, Akhisar, Manisa, Turkey
Wild 8	Karacakas 2, Soma, Manisa, Turkey
Wild 9	Bornova 2, Bornova, İzmir, Turkey
Wild 10	Dikili, Bademli, İzmir, Turkey

Genomic DNA was extracted from young leaves using the Doyle and Doyle method (1987). Forty-five different decamer primers were used for RAPD analyses of *Olea europaea oleasters*. A total of 45 primers from the OP-A, OP-I, and OP-Z (7, 8, 9, 19, 20) kits (Operon Technologies, Alameda, CA, USA) were used for RAPD-polymerase chain reaction

(PCR) analysis. The data obtained with RAPD markers were evaluated using 5 similarity coefficients (Table 2), and the similarity coefficients were calculated with NTSYS pc 2.01 (Rohlf, 1998).

Table 2. Similarity coefficients used among 10 wild olives based on RAPD markers.

Coefficient	Expression	Interval	References
Jaccard	$\frac{A}{A+B+C}$	(0,1)	Jaccard, 1901
Sorensen-Dice	$\frac{2A}{2A+B+C}$	(0,1)	Dice, 1945; Sørensen, 1948
Simple matching	$\frac{A+D}{A+B+C+D}$	(0,1)	Sokal and Michener, 1958
Rogers and Tanimoto	$\frac{A+D}{A+D+2(B+C)}$	(0,1)	Rogers and Tanimoto, 1960
Russel and Rao	$\frac{A}{A+B+C+D}$	(0,1)	Russel and Rao,1940

Five similarity coefficients were compared using the Mantel test for RAPD markers (Rohlf, 1998). Dendrograms were constructed according to the UPGMA, using NTSYS pc 2.01 (Rohlf, 1998). The dendrograms based on different coefficients were compared by the consensus fork index (CI_C) (Rohlf, 2000). CI_C provides a relative estimate of dendrogram similarities and was calculated using NTSYS pc 2.01 (Rohlf, 1998).

RESULTS AND DISCUSSION

The Mantel test correlation coefficients between the five similarity coefficients based on RAPD markers are shown in Table 3.

Table 3. The Mantel test of similarity coefficients based on RAPD markers.

Coefficient	J	SD	SM	RT	RR
J	**********				
SD	0.9984[1]	**********			
SM	0.1891	0.1815	**********		
RT	0.1908	0.1828	0.9980[1]	**********	
RR	0.9617[1]	0.9620[1]	0.0363	0.0326	**********

J = Jaccard; SD = Sorensen-Dice; SM = simple matching; RT = Rogers and Tanimoto; RR = Russel and Rao. [1]The correlation between matrices is significant ($P < 0.05$).

The correlation between Jaccard (1901) and Sorensen (1948)-Dice (1945) values were 0.9984; between Jaccard (1901) and Russel and Rao (1940), 0.9617; between Sorensen (1948)-Dice (1945) and Russel and Rao (1940), 0.9620, and between simple matching and Rogers and Tanimoto (1960), 0.9980, and all were found to be significant.

In accordance with these results, Jaccard (1901) and Sorensen (1948)-Dice (1945) were highly correlated. The same situation occurred between simple matching and the Rogers and Tanimoto (1960) coefficients. Considering the dendrograms obtained with different coefficients using UPGMA (Figures 1-3), the same results were obtained for Jaccard (1901) and Sorensen (1948)-Dice (1945) (Wild 4 and Wild 9), (Wild 1 and Wild 3), (Wild 7 and Wild 8)

olives were found to be the closest samples in terms of genetic similarities. The closest samples in terms of genetic similarities obtained from Rogers and Tanimoto (1960) and simple matching were found (Wild 4 and Wild 9), (Wild 1 and Wild 3). The results from Russel and Rao (1940), (Wild 1 and Wild 3), (Wild 7 and Wild 8) as well as (Wild 6 and Wild 4) and (Wild 2 and Wild 9) olives, which were not observed with any of the other coefficients, were found to be the ones closest to each other in terms of genetic similarities. When viewed in general, the dendrograms provided similar results except for Russel and Rao (1940). Similar results have been observed also in a study with the maize plant (Meyer Da Silva et al., 2004).

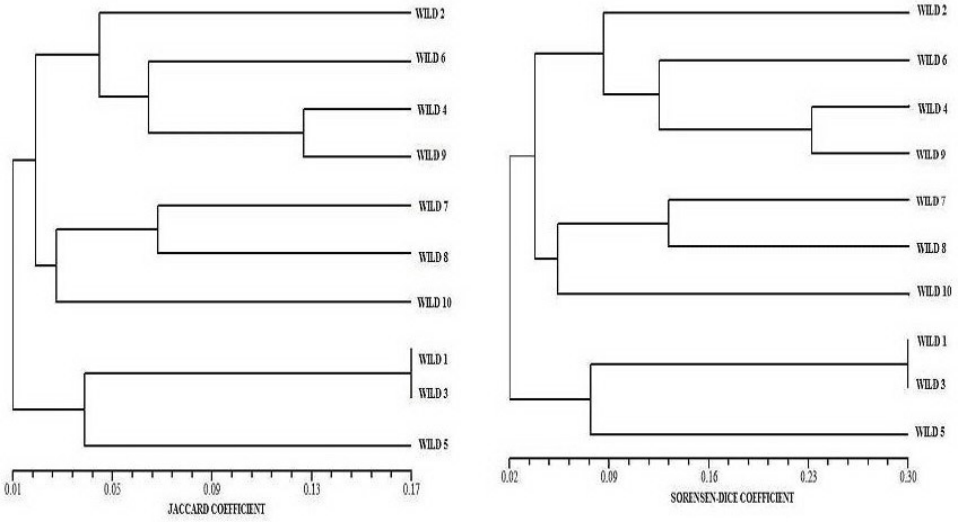

Figure 1. Dendrograms showing genetic similarities among 10 wild olives, constructed using the Jaccard and Sorensen-Dice similarity coefficients based on RAPD markers (UPGMA).

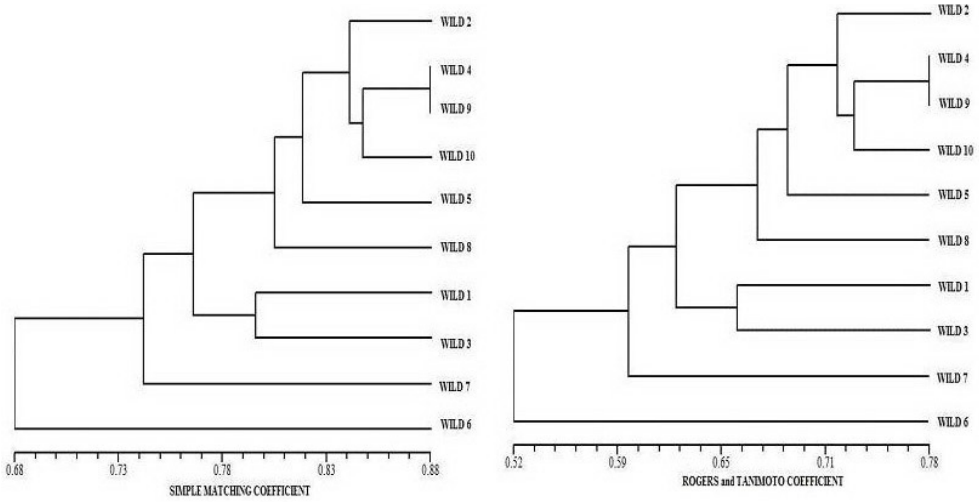

Figure 2. Dendrograms showing genetic similarities among 10 wild olives, constructed using the simple matching and Rogers and Tanimoto similarity coefficients based on RAPD markers (UPGMA).

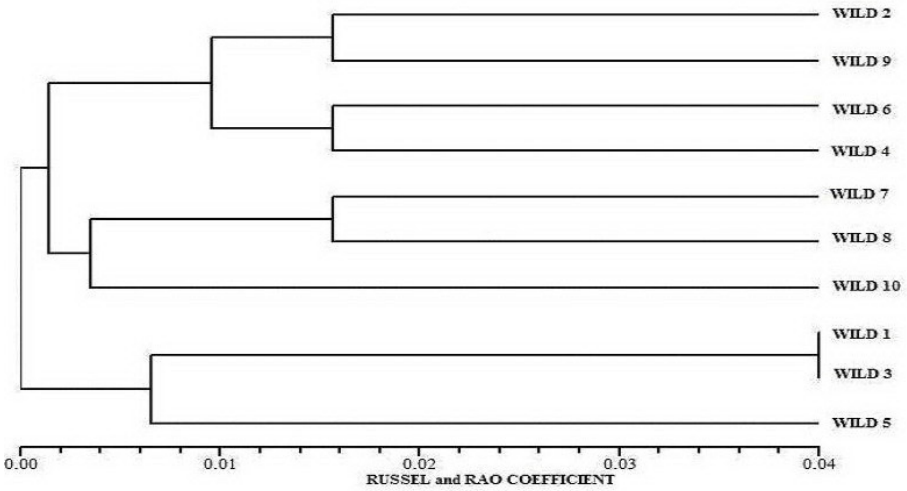

Figure 3. Dendrograms showing genetic similarities among 10 wild olives, constructed using the Russell and Rao similarity coefficient based on RAPD markers (UPGMA).

The dendrograms were controlled using the CI_C. This index is between 0 and 1. The dendrograms are identical when the CI_C value is 1 (Table 4).

Table 4. Consensus fork index between the dendrograms (UPGMA) produced by similarity coefficients for 10 wild olives, based on RAPD markers.

Coefficient	J	SD	SM	RT	RR
J	**********	1.00	0.25	0.25	0.75
SD	0.75	**********	0.25	0.25	0.75
SM	0.10	0.10	**********	1.00	0.125
RT	0.10	0.10	0.90	**********	0.125
RR	0.60	0.60	0.05	0.005	**********

J = Jaccard; SD = Sorensen-Dice; SM = simple matching; RT = Rogers and Tanimoto; RR = Russel and Rao.

In accordance with the results obtained, CI_C = 1 was found for the Jaccard (1901) and Sorensen (1948)-Dice (1945) coefficients and Rogers and Tanimoto (1960) and simple matching coefficients. The Russel and Rao (1940) coefficient showed very low values with the simple matching and Rogers and Tanimoto (1960) coefficients (CI_C = 0.125). This is in line with the findings from the dendrograms and shows why different results obtained from the dendrogram using the Russel and Rao (1940) coefficient as compared to the others.

The Jaccard (1901) and Sorensen (1948)-Dice (1945) coefficients provide different results compared to the Rogers and Tanimoto (1960) and simple matching coefficients, because these do not consider the negative co-occurrences (Dalirsefat et al., 2009). In comparisons performed based on RAPD markers, the Jaccard (1901) and Sorensen (1948) or Dice (1945) coefficients may be preferable to the simple matching coefficient (Landry and Lapointe, 1996). The reason why the Russel and Rao (1940) coefficient differs from the other coefficients is that the negative co-occurrences are only available in the denominator. In this respect, such coefficients should be used in specific situations (Meyer Da Silva et al., 2004).

In conclusion, both the Jaccard (1901) and Sorensen (1948)-Dice (1945) coefficients,

which were used for evaluating the results obtained based on RAPD markers in this study using 5 different similarity coefficients, determined that the genetic similarities between Wild 4 and Wild 9, Wild 1 and Wild 3, and Wild 7 and Wild 8 olives were close. However, simple matching, Rogers and Tanimoto (1960), and Russel and Rao (1940) coefficients, which include negative co-occurrences, produced different results from their dendrograms.

Accordingly, these findings in wild olives indicate that similarity coefficients that do not include negative co-occurrences are more efficient for the studies performed with dominant markers.

ACKNOWLEDGMENTS

Research supported by the Turkish Republic State Planning Organization.

REFERENCES

Dalirsefat SB, da Silva MA and Mirhoseini SZ (2009). Comparison of similarity coefficients used for cluster analysis with amplified fragment length polymorphism markers in the silkworm, *Bombyx mori*. *J. Insect Sci.* 9: 1-8.

Dice LR (1945). Measures of the amount of ecologic association between species. *Ecology* 26: 297-302.

Doyle JJ and Doyle JL (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 9: 11-15.

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

Gower JC and Legendre P (1986). Metric and Euclidean properties of dissimilarity coefficients. *J. Classif.* 3: 5-48.

Jaccard P (1901). Étude comparative de la distribuition florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat.* 37: 547-579.

Jackson AA, Somers KM and Harvey HH (1989). Similarity coefficients: measures for co-occurrence and association or simply measures of occurrence. *Am. Natur.* 133: 436-453.

Johnson RA and Wichern DW (1988). Applied Multivariate Statistical Analysis. Prentice Hall, New Jersey, 607.

Landry PA and Lapointe FJ (1996). RAPD problems in phylogenetics. *Zoolog. Scripta* 25: 283-290.

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

Rogers JS and Tanimoto TT (1960). A computer program for classifying plants. *Science* 132: 1115-1118.

Rohlf FJ (1998). NTSYSpc: Numerical Taxonomy and Multivariate Analysis System. Version 2.02. Exeter Software, Setauket, New York.

Rohlf FJ (2000). NTSYSpc: Numerical Taxonomy and Multivariate Analysis System. Version 2.02. Exeter Software, Setauket, New York.

Russel PF and Rao TR (1940). On habitat and association of species of anophelinae larvae in south-eastern Madras. *J. Malar. Inst. India* 3: 153-178.

Skroch P, Tivang J and Nienhuis J (1992). Analysis of Genetic Relationships Using RAPD Marker Data. In: Applications of RAPD Technology to Plant Breeding. Symposia series, Madison, CCSA, ASHS and AGMA, Minneapolis, 26-30.

Sokal RR and Michener CD (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* 38: 1409-1438.

Sørensen T (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Videnski Selsk. Biol. Skr.* 5: 1-34.

Weir BS (1996). Genetic Data Analysis II: Methods for Discrete Population Genetic Data. Sinauer Associates Inc., Sunderland, 445.