



# Comparison of maize similarity and dissimilarity genetic coefficients based on microsatellite markers

M. Balestre, R.G. Von Pinho, J.C. Souza and J.L. Lima

Departamento de Biologia, Universidade Federal de Lavras, Lavras, MG, Brasil

Corresponding author: M. Balestre  
E-mail: marciobalestre@hotmail.com

Genet. Mol. Res. 7 (3): 695-705 (2008)  
Received March 13, 2008  
Accepted June 27, 2008  
Published August 5, 2008

**ABSTRACT.** The present study compared different similarity and dissimilarity coefficients and their influence in maize inbred line clustering. Ninety maize  $S_{0,1}$  inbred lines were used and genotyped with 25 microsatellite markers (simple sequence repeat). The simple matching, Rogers and Tanimoto, Russel and Rao, Hamann, Jaccard, Sorensen-Dice, Ochiai, and Roger's modified distance coefficients were compared by consensus index, projection efficiency in a two-dimensional space and by Spearman's correlation. Changes were found in high genetic similarity groupings with different coefficients using the consensus index. Russel and Rao and Jaccard coefficients had the greatest stress values with 75.67 and 40.16%, respectively, indicating that these coefficients should not be used. Genotype ranking changed, mainly in the comparison of the Roger's modified distance in relation to some coefficients ( $r_s = 0.75$ ). Russel and Rao's and Jaccard's coefficients should be avoided for their low accuracy. Moreover, genotype clustering by different similarity coefficients, without a close consideration of these coefficients could affect the research results.

**Key words:** Genetic divergence; Cluster analysis; Heterotic groups

## INTRODUCTION

Several studies have been published in the last years using molecular markers to study genetic divergence and phylogenetic relations between species (Dias et al., 2004).

In maize crops, several researchers have made an effort to estimate hybrid performance allocating the inbred lines in heterotic groups, or even directing the crosses based on genotype genetic distances (Reif et al., 2003; Amorim et al., 2006; Guimarães et al., 2007). However, the results of these associations are inconsistent.

These studies usually employ different similarity coefficients for dominant markers, such as random amplified polymorphic DNA and amplified fragment length polymorphism, as well as for multiallele ones, such as simple sequence repeat (SSR). Some of these coefficients are specific for dichotomic variables, and their use is suggested for divergence studies based on molecular markers (Duarte et al., 1999). In general, the similarity coefficients are based on comparisons between the occurrence of common bands (indicated by ones in common in a data matrix), different bands (indicated by one and zero or zero and one), while other coefficients also consider the occurrence of zeros in common. Also, genetic dissimilarity coefficients are used, such as Roger's modified distance and Nei's distance, which explore information of allele frequency obtained by molecular markers, especially those of microsatellites.

According to Reif et al. (2005), the choice of a coefficient for studying divergence depends on the marker system properties involved, on the germplasm genealogy, on the taxonomic operational unit involved, on the study objectives, and on the conditions required for a multivariate analysis. According to these authors, several studies ignore these conditions, especially those related to the coefficient properties, which are connected to the study objective, and the conditions for multivariate analysis, which are very important for decision making considering the proper coefficient to be used.

Amorim et al. (2006) used microsatellite markers to estimate the performance of maize inbred lines based on genetic distances. These authors used partially endogamic maize lines ( $S_{0.1}$ ) and the genetic distance relations were estimated by Roger's modified distance. The authors found a high correlation between genetic distance and performance of the single cross hybrids.

In a similar study, with inbred lines and microsatellite and amplified fragment length polymorphism markers, Guimarães et al. (2007) used Jaccard's coefficient to estimate the genetic distance between the genotypes. These authors found a low correlation between genetic distance and hybrid performance, demonstrating that, for the same objectives and using the same markers, there is no simple explanation for the choice of genetic distance coefficients.

Considering that the results of genotype clustering can be affected by the choice of the similarity coefficient (Jackson et al., 1989; Duarte et al., 1999), these coefficients need to be better studied, since genotype ranking changes can affect the values in correlation studies. Moreover, there is a risk of locating genotypes in non-corresponding heterotic groups. Another aspect that needs to be highlighted is that, usually, papers published do not justify the choice of the coefficient employed. Jackson et al. (1989) state that this lack of justification can cause problems, jeopardizing the nature of the analysis.

Thus, the knowledge of the mathematical properties and, consequently, the application of these coefficients in genetic studies, are imperative. Moreover, some of these coefficients present Euclidian distance properties, allowing the application of multivariate analyses

as a factor for choosing the most adequate coefficient for a given study (Eymedio et al., 2003; Meyer et al., 2004; Reif et al., 2005).

This study compared, through multivariate analysis techniques, different similarity and genetic distance and their effect on clustering maize inbred lines.

## MATERIAL AND METHODS

The comparison of similarity and genetic distance coefficients was done with 90  $S_{0.1}$  partially endogamic progenies from three populations derived from single cross hybrids. These progenies were coded, with lines 1-30 belonging to population A, originating from the single cross hybrid 30F45 of Pioneer, lines 31-61 belonging to population B, from the simple hybrid DOW657 of DowAgroScience, and lines 62-90 belonging to population C, from the simple hybrid DKB333B of Monsanto.

Twenty seeds of each line were sown in 128-celled trays in a green house. After the emergence of the second leaf pair, leaf tissue from the twenty plants for each line was bulked. DNA was extracted according to Saghai-Marouf et al. (1984). Extracted DNA was quantified in a fluorimeter and its concentration adjusted to 10 ng/mL.

Subsequently, the material was genotyped with 25 microsatellite primers. Nine of these primers are QTL linked to kernel yield (Amorim et al., 2006). This number of SSR primers was obtained by bootstrap analysis with 49 SSR for the same populations according to Amorim et al. (2006). The reactions with SSR were prepared for a final volume of 11.06  $\mu$ L, consisting of 2.25  $\mu$ L genomic DNA, 1.96  $\mu$ L reaction buffer (50 mM Tris, 2.0 mM  $MgCl_2$ , 20 mM KCl, 250 mg/mL bovine serum albumin, 1% Ficoll 400, 1 mM tartrazine), 100  $\mu$ M dNTPs, 4.45  $\mu$ L pure water, 0.6  $\mu$ L Taq DNA polymerase, and 0.2  $\mu$ M of each primer (forward and reverse). The amplifications were done in a thermocycler (Thermocycler Mastercycler Gradient), using 0.2-mL microtubes.

Initial denaturation was done at 95°C for 2 min, followed by 32 cycles, each one consisting of a denaturation at 95°C for 20 s, annealing at 55°C for 20 s, and a final extension at 72°C for 20 s. The fragments were separated on a 3% agarose gel (Invitrogen), prepared with 0.5X TBE buffer. Electrophoresis was done at constant 100 V, for approximately 3 h in a horizontal tank. After electrophoresis, the gels were treated with ethidium bromide (0.5  $\mu$ g/mL), for 20 min and viewed under ultraviolet light. The images were captured with an EDAS 290 digital camera (Kodak Digital Science). Subsequently, a binary matrix was made and used for the comparison of similarity coefficients. The binary data were converted into allele frequencies with the program NTSYS-PC 2.1 to use Roger's modified distance coefficient (Rohlf, 2000).

Seven similarity coefficients ( $sg_{ij}$ ) were used with dichotomy analysis and one coefficient of genetic ( $d_{ij}$ ) dissimilarity, with allele analysis (Table 1). In the Appendix, there is one example of how to calculate similarity between two pair of genotypes by dichotomy analysis. The similarity data were, then, transformed into dissimilarity measures ( $dg_{ij}$ ) and subjected to Euclidian properties through the expression:  $dg_{ij} = \sqrt{(1 - sg_{ij})}$ , with the exception of Roger's modified distance, which already possesses this property in its expression (Reif et al., 2005). All similarity matrices met the assumptions for transformation into genetic dissimilarity, i.e., they were defined as non-negative (Johnson and Wichern, 1998). The genetic similarity and dissimilarity analyses were done with the program NTSYS-PC 2.1.

**Table 1.** Similarity coefficients and distances used for analysis of the genetic data obtained with SSR markers in 90 maize progenies.

| Coefficient                   | Expression <sup>1</sup>  |
|-------------------------------|--|
| Simple matching (SM)          | $\frac{a+d}{a+b+c+d}$  |
| Rogers and Tanimoto (RT)      | $\frac{a+d}{a+d+2(b+c)}$   |
| Russel and Rao (RR)           | $\frac{a}{a+b+c+d}$  |
| Hamann (H)                    | $\frac{(a+d)-(b+c)}{a+b+c+d}$                                    |
| Jaccard (J)                   | $\frac{a}{a+b+c}$  |
| Sorensen-Dice (SD)            | $\frac{2a}{2a+b+c}$  |
| Ochiai (O)                    | $\frac{a}{\sqrt{(a+b)(a+c)}}$                                    |
| Roger's modified distance (R) | $d_{ij} = \sqrt{\frac{1}{2n_{loci}} \sum_k (x_{ki} - x_{kj})^2}$ |

<sup>1</sup> $a = 1$  and  $l$ ;  $b = 1$  and  $0$ ;  $c = 0$  and  $l$ ;  $d = 0$  and  $0$ .

<sup>2</sup> $n$  = number of loci;  $x_{ki}$  and  $x_{kj}$  = frequency of k-esm allele for the lines  $i$  and  $j$ .

The dendrograms were made with the SAHN method (sequential, agglomerative, hierarchic, non-overlapping clustering methods) and with UPGMA (unweighted pair-group method using the arithmetic average) (Rohlf, 2000).

Different coefficients were compared by a co-phenetic correlation between the dissimilarity matrices with the co-phenetic matrices ( $r_c$ ). This correlation measures the adjustment degree between the original dissimilarity matrices and the matrix resulting from the simplification given by the grouping method. The consensus index ( $CI_c$ ), that provides a relative estimate of dissimilarities between cladograms was estimated, because it uses the proportion of common subgroups between two cladograms in a total number of subgroups formed by all progenies.

Genotype ranking changes were analyzed by Spearman's correlation ( $r_s$ ), using SAS, version 8.0 (SAS Institute, 2000).

Similarity projection in a bidimensional space was done using the software GENES, to determine the effect of choosing the similarity coefficient (Cruz, 1997; Meyer et al., 2004). This technique of multivariate analysis estimates the coordinates for each individual, from a dissimilarity matrix, by statistic procedures that minimize the differences between the original distances and the distances obtained in the bidimensional plane. This method allowed the comparison of the coefficients using the three parameters listed:

- a) Correlation between the original and the estimated distances ( $r$ );

b) Distortion degree ( $I-\alpha$ ), which is given by:

$$\alpha = \frac{\sum_{i < j} \sum d_{gij}^2}{\sum_{i < j} \sum d_{oij}^2};$$

c) Stress value ( $s$ ), which is given by:

$$s = \sqrt{\frac{\sum_{i < j} \sum (d_{oij} - d_{gij})^2}{\sum_{i < j} \sum d_{oij}^2}}$$

where:  $d_{gij}$  and  $d_{oij}$  are the graphic (bidimensional space) and original distances (non-dimensional space), respectively, for all pairs of genotypes  $i$  and  $j$ .

Stress value is a parameter that determines the quality of the graphic projection adjustment, which, according to Kruskal (1964), can be classified as 40, 20, 10, 5, or 0% stress, that is, adjustment will be non-satisfactory, regular, good, excellent and perfect, respectively.

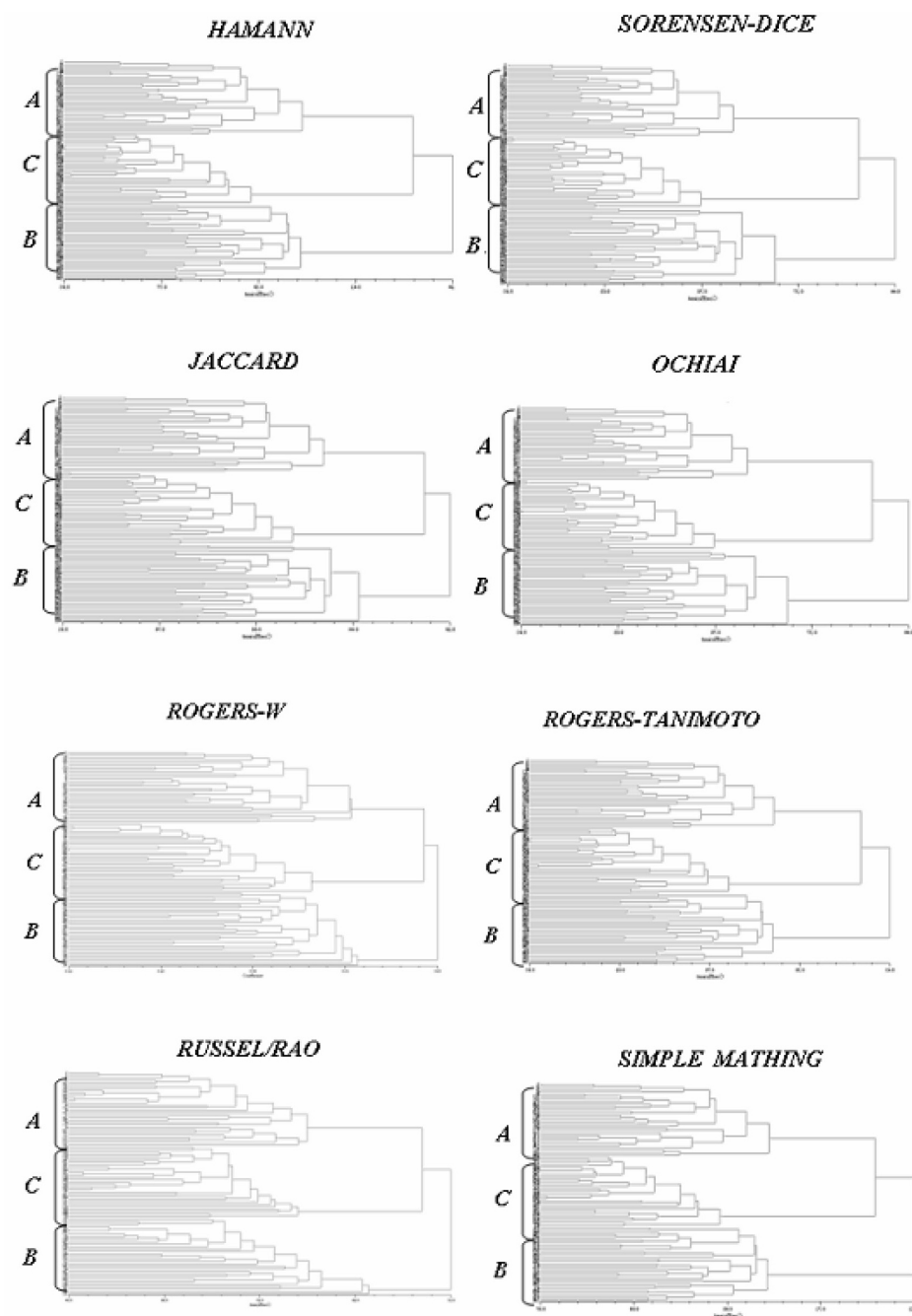
## RESULTS

Through 25 SSR markers 103 alleles were obtained, with 4.12 alleles on average per loci (Table 2). The dendrograms obtained by the eight coefficients presented the same population distribution structure (Figure 1), i.e., from the graph visual analysis it is possible to note that all coefficients were effective in distinguishing the three heterotic groups formed by the three populations (A, B and C). In group A, all coefficients grouped the progenies 1 to 30, as well as for progenies of group B (31-61) and those of group C (62-90). Meyer et al. (2004), studying maize lines originating from two populations (BR-105 and BR-106), found that the use of different similarity coefficients did not change line grouping within each group. Similar results were found by other authors in common beans (Johns et al., 1997; Duarte et al., 1999) and in potatoes (Rocha et al., 2001).

**Table 2.** Microsatellite markers (simple sequence repeat, SSR), genomic positions (Bin) and allele numbers obtained for 90 maize  $S_{0.1}$  inbred lines.

| SSR locus | Bin  | Alleles | SSR locus | Bin   | Alleles |
|-----------|------|---------|-----------|-------|---------|
| BNLG1484  | 1.03 | 5       | UMC1409   | 7.01  | 3       |
| BNLG615*  | 1.07 | 6       | BNLG434   | 7.03  | 3       |
| BNLG1737* | 1.10 | 5       | BNLG1666  | 7.04  | 6       |
| UMC1845   | 2.03 | 2       | BNLG240*  | 8.06  | 4       |
| BNLG2077* | 2.07 | 3       | BNLG1607  | 8.06  | 5       |
| BNLG2241  | 3.06 | 2       | UMC1069   | 8.08  | 5       |
| BNLG1951  | 3.06 | 3       | UMC1638   | 8.09  | 3       |
| BNLG1318  | 4.01 | 3       | BNLG2122  | 9.01  | 3       |
| UMC1088*  | 4.05 | 5       | UMC1893   | 9.02  | 4       |
| BNLG1738* | 4.10 | 5       | BNLG1250  | 10.05 | 3       |
| UMC1221   | 5.04 | 5       | Total     |       | 103     |
| UMC1792*  | 5.08 | 5       | Average   |       | 4.12    |
| BNLG1600* | 6.00 | 5       |           |       |         |
| UMC1014   | 6.04 | 4       |           |       |         |
| UMC1653*  | 6.07 | 6       |           |       |         |

\*QTL linked to kernel yield.



**Figure 1.** Dendrograms made from the genetic dissimilarity matrices obtained from the coefficients simple matching, Rogers and Tanimoto, Russel and Rao, Hamann, Jaccard, Sorensen-Dice, Ochiai, and Roger's modified distance. The groups A, B and C represent the progenies 1-30, 31-61, and 62-90, respectively.

The consensus index, comparing the dendrograms in a more refined manner since it uses the proportion of common subgroups among the dendrograms, has been suggested by some authors (Duarte et al., 1999; Meyer et al., 2004). There was great variation in the results, that is, the consensus values among the coefficients varied from 0 to 1 (Table 3). This demonstrates that there is a great variation in grouping from one coefficient to another, especially in the comparison of coefficients that do not take into account common absences, in relation to those that consider this information. Although, the changes were found only within groups, which is not of utmost importance in phylogeny studies, these changes can mask the results if the similarities of the genotypes used in the study are high.

**Table 3.** Comparison of dendrograms through the consensus index ( $I_c$ ), generated from the different genetic similarity and dissimilarity coefficients in 90 maize progenies.

| Coefficient              | D    | H    | J    | O    | R    | RR   | RT   | SM   |
|--------------------------|------|------|------|------|------|------|------|------|
| Dice (D)                 | 1.00 | 0.64 | 1.00 | 0.88 | 0.00 | 0.64 | 0.31 | 0.64 |
| Hamann (H)               | -    | 1.00 | 0.64 | 0.70 | 0.00 | 0.22 | 0.22 | 1.00 |
| Jaccard (J)              | -    | -    | 1.00 | 0.88 | 0.00 | 0.64 | 0.31 | 0.64 |
| Ochiai (O)               | -    | -    | -    | 1.00 | 0.00 | 0.70 | 0.29 | 0.70 |
| Roger's modified (R)     | -    | -    | -    | -    | 1.00 | 0.01 | 0.01 | 0.01 |
| Russel and Rao (RR)      | -    | -    | -    | -    | -    | 1.00 | 0.22 | 0.22 |
| Rogers and Tanimoto (RT) | -    | -    | -    | -    | -    | -    | 1.00 | 1.00 |
| Simple matching (SM)     | -    | -    | -    | -    | -    | -    | -    | 1.00 |

There was a high correlation between the co-phenetic matrix and the original one for all coefficients (Table 4) demonstrating that there is a good representation of the similarity matrices and distances in the form of dendrograms. However, due to similarities of co-phenetic values between the coefficients, this parameter was not useful in distinguishing the best coefficient.

**Table 4.** Efficacy of the projection of genetic distances in a bidimensional space, through the co-phenetic correlation ( $r_c$ ), distortion degree, correlation between the original and projected distances ( $r$ ), and stress values.

| Coefficients        | $r_c$ co-phenetic | Distortion (%) | Correlation ( $r$ ) | Stress (%) |
|---------------------|-------------------|----------------|---------------------|------------|
| Simple matching     | 0.89**            | 27.23          | 0.86                | 32.56      |
| Rogers and Tanimoto | 0.91**            | 35.35          | 0.85                | 39.53      |
| Russel and Rao      | 0.85**            | 73.25          | 0.53                | 75.69      |
| Hamann              | 0.89**            | 27.73          | 0.86                | 32.34      |
| Jaccard             | 0.90**            | 36.62          | 0.83                | 40.16      |
| Sorensen-Dice       | 0.88**            | 25.02          | 0.85                | 31.48      |
| Ochiai              | 0.89**            | 25.08          | 0.84                | 31.58      |
| Roger's modified    | 0.87**            | 29.84          | 0.83                | 31.97      |

\*\*Significant to the Mantel test at 1% probability based on 5000 simulations.

The projection efficacy of the genetic distances of the similarity coefficients, in a bidimensional space, according to different criteria (stress, distortion and correlation between original and estimated distances), presented little variation (Table 4). There were no major differences among most coefficients; however, the coefficients of Rogers and Tanimoto, Jaccard and Russel, and Rao deserve special note since they presented the greatest stress values



(Table 4). Emygdio et al. (2003) also observed that the coefficients of Rogers and Tanimoto, and Russel and Rao had the greatest stress estimates in common bean genotype grouping (61 and 78%, respectively); similar results were also found by Duarte et al. (1999), demonstrating that these coefficients should not be used.

It can be seen that, for most coefficients, there was a high correlation among the genetic dissimilarity matrices (Table 5). However, in some cases there was an inversion of the observation posts, that is, it was possible to observe changes in genotype ranking. This demonstrates that, depending on the coefficient choice, the researcher can find different correlation values due to change in line ranking, as can be seen with Roger's modified distance coefficient, which had matrix correlations varying from 0.75 to 0.81.

**Table 5.** Spearman's correlation between 4005 pairs of genetic distances obtained through different genetic similarity and dissimilarity coefficients.

| Coefficients             | D      | H      | J      | O      | R      | RR     | RT     | SM     |
|--------------------------|--------|--------|--------|--------|--------|--------|--------|--------|
| Dice (D)                 | 1.00** | 0.92** | 0.99** | 0.99** | 0.81** | 0.92** | 0.94** | 0.92** |
| Hamann (H)               | -      | 1.00** | 0.92** | 0.91** | 0.75** | 1.00** | 0.75** | 1.00** |
| Jaccard (J)              | -      | -      | 1.00** | 0.99** | 0.81** | 0.94** | 0.92** | 0.92** |
| Ochiai (O)               | -      | -      | -      | 1.00** | 0.81** | 0.94** | 0.92** | 0.92** |
| Roger's modified (R)     | -      | -      | -      | -      | 1.00** | 0.76** | 0.79** | 0.79** |
| Russel and Rao (RR)      | -      | -      | -      | -      | -      | 1.00** | 0.75** | 0.75** |
| Rogers and Tanimoto (RT) | -      | -      | -      | -      | -      | -      | 1.00** | 1.00** |
| Simple matching (SM)     | -      | -      | -      | -      | -      | -      | -      | 1.00** |

\*\*Significant to the *t* test at 1% probability.

## DISCUSSION

The visual analysis of Figure 1 indicates that the dendrograms were affected by the analysis nature of the coefficients, i.e., the coefficients Jaccard, Ochiai, and Sorensen-Dice, which do not consider the common absence of bands (Table 1), had the same intragroup distribution pattern. The same was found for the coefficients that consider a common absence (simple matching, Rogers and Tanimoto and Hamann). However, the coefficient of Russel and Rao presented a grouping distinct from all others, as well as Roger's modified distance, which presented a grouping quite different from all others, demonstrating that, for low dissimilarity values, there is a change in genotype grouping (Duarte et al., 1999; Meyer et al., 2004).

There was no consensus between the groupings from the expression of Roger's modified distance with any other coefficient. This fact could be due to the use of allele frequencies by the Roger's modified distance, and not the use of dichotomous variables, as in the other coefficients. In contrast, there were no differences between the coefficients of Jaccard and Sorensen-Dice, as well as between the coefficients simple matching and Hamann and Rogers and Tanimoto. The dendrograms obtained from Russel and Rao's coefficient, although considering the common absences (d), had low consensus with all other coefficients that explore the same information. However, this coefficient had greater consensus with coefficients that disregard the information of common absences (d). This fact could be due to the similarity of Russel and Rao's expression with the coefficients that do not explore the common information (Table 1), especially for using only the common occurrence of bands (a) in the numerator of the expression.

In this study, the high stress value observed for Jaccard's coefficient needs attention.



This coefficient had a stress value of 40.16%, which, according to Kruskal's scale, would be a non-satisfactory index. Similar results were found by other authors (Emygdio et al., 2003; Meyer et al., 2004). However, several published papers indicated few promising results in the study of genetic distance correlation with hybrid performance, and also in estimating maize hybrids through the methodology of mixed models (BLUP), using Jaccard's coefficient in grouping the lines (Iemma, 2003; Guimarães et al., 2007).

From the high stress values presented by Jaccard's coefficient in this study and in previous ones (Emygdio et al., 2003; Meyer et al., 2004), and, also, due to the lack of justification about the use of this coefficient (Iemma, 2003; Guimarães et al., 2007), its empirical choice could have affected the results, not adjusting the binary data to the estimated genetic distances. Moreover, it is recommended that the researcher considers the properties of the markers used, and the coefficients listed in Table 1 are indicated for dominant, or non-allelic, markers, with the exception of Roger's modified distance, which is recommended for multi-allele markers, such as microsatellites (Reif et al., 2005).

The coefficients that presented the smallest stress values were those that did not consider the common absences (Sorensen-Dice, Ochiai) and Roger's modified distance, which is calculated from the allele frequencies obtained by the transformation of the binary matrix. The use of co-dominant markers (SSR) in this study would lead to the assumption that the coefficients possessing (d) in its expression would have better adjustments, which was not observed (Table 4). The index (d) considers that the allele in question is not present in both pairs of genotypes compared.

Reif et al. (2005) report that the probability of non-amplification of a DNA fragment is high. Therefore, the coefficients that consider the common absence of bands should be avoided, since the lack of bands in both genotypes cannot always be interpreted as a common characteristic, that is, one cannot infer for a given locus that the genotypes are similar when they actually are not.

Correlation studies between genetic distances and maize hybrid performance have shown discrepancies, as shown by Dias et al. (2004). Since the correlation depends on the covariance between the genetic distances (X) and other variables (Y), such as yield, heterosis, specific combination capacity, if a ranking change occurs in genetic distances and ranking of variable Y remains constant, there will be a change of association value of the two variables. Thus, the choice of a coefficient that adjusts itself to molecular data could be a determining factor for the success of that given study.

Although the Spearman correlation values among the coefficients observed in this study are apparently high, i.e., the Roger's modified distance coefficient showed correlation values between 0.75 and 0.81, one should also consider that in regression studies, the linear determination coefficient ( $R^2$ ) for values similar to these, for example, would indicate that the coefficients of Hamann and Sorensen-Dice would explain only 56.25 and 65.61%, respectively, of the genetic distances of Roger's modified distance expression.

According to Melchinger (1999), heterosis can be described through the expression  $\Delta H = \sum DMR^2 d_i$ , that is, heterosis is a function of the dominance effect of the QTL and of Roger's modified distance. Reif et al. (2005), based on the genetic and mathematical properties of Roger's modified distance, concluded that this expression is linearly related with heterosis and suggests its use for multi-allele markers in estimation studies of heterosis and in establishing heterotic groups. Taking this coefficient as a reference, due to its genetic and mathematical properties, in comparison with the other coefficients (Table 5), it can be inferred that the

choice, without criteria, of similarity coefficients can change the research results since the correlation between Roger's modified distance and the other coefficients is not near 1.

This experiment did not allow to reach a definite position about the best coefficient to study genetic diversion through the bidimensional analysis technique. However, it was possible to demonstrate that the coefficients of Russel and Rao, Jaccard and Roger, and Tanimoto should be avoided.

Roger's modified distance can be used for the study of divergence and phylogeny, since it presents acceptable values of stress and distortion.

Due to changes in genotype grouping and ranking for the different similarity and dissimilarity coefficients, the choice, without criteria, of these coefficients can affect the results of research.

## ACKNOWLEDGMENTS

The authors thank Dr. João Bosco dos Santos and the Molecular Genetics Laboratory of the Department of Biology at the Universidade Federal de Lavras for making the reagents available, and Dr. Anete Pereira de Souza (CBMEG/UNICAMP), for supplying the microsatellite primers. Research supported by FAPEMIG.

## REFERENCES

- Amorim EP, Amorim VBO, dos Santos JB, Pereira deSouza A, et al. (2006). Genetic distance based on SSR and grain yield of inter- and intrapopulational maize single cross hybrids. *Maydica* 51: 507-513.
- Cruz CD (1997). Programa Genes: aplicativo computacional em genética e estatística. Universidade Federal de Viçosa, Viçosa.
- Dias LA, Picoli EA, Rocha RB and Alfenas AC (2004). *A priori* choice of hybrid parents in plants. *Genet. Mol. Res.* 3: 356-368.
- Duarte JM, dos Santos JB and Melo LC (1999). Comparison of similarity coefficients based on RAPD markers in the common bean. *Genet. Mol. Biol.* 22: 427-432.
- Emygdio BM, Antunes IF, Choer E and Nedel JL (2003). Eficiência de coeficientes de similaridade em genótipos de feijão mediante marcadores RAPD. *Pesq. Agropec. Bras.* 38: 243-250.
- Guimarães PS, Paterniani MEAGZ, Lüders RR, Souza AP, et al. (2007). Correlation between the heterosis of maize hybrids and genetic divergence among lines. *Pesq. Agropec. Bras.* 42: 811-816.
- Iemma M (2003). Use of best linear unbiased prediction in diallel analysis and in prediction of single crosses. Doctoral thesis, Escola Superior de Agricultura Luiz de Queiroz (ESALQ), Piracicaba.
- Jackson AA, Somers KM and Harvey HH (1989). Similarity coefficients: measures for co-occurrence and association or simply measures of occurrence? *Am. Nat.* 133: 436-453.
- Johns MA, Skroch PW, Nienhuis J, Kinrichsen P, et al. (1997). Gene pool classification of common bean landraces from Chile based on RAPD and morphological data. *Crop Sci.* 37: 605-613.
- Johnson RA and Wichern DW (1998). Applied Multivariate Statistical Analysis. Prentice-Hall, New Jersey.
- Kruskal JB (1964). Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrika* 29: 1-27.
- Melchinger AE (1999). Genetic diversity and heterosis. In: The Genetics and Exploitation of Heterosis in Crops (Coors JG and Pandey S, eds.). CCSA, ASA, and SSA, Madison.
- Meyer AS, Garcia AAF, Souza AP and 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.
- Reif JC, Melchinger AE, Xia XC, Warburton ML, et al. (2003). Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. *Crop Sci.* 43: 1275-1282.
- Reif JC, Melchinger AE and Frisch M (2005). Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop Sci.* 41: 1-7.

Rocha BHG, Augustin E, da Silva JB and Viegas J (2001). Isoenzymatic variability in wild potatoes. *Pesq. Agropec. Bras.* 36: 781-791.  
 Rohlf FJ (2000). NTSYSpc numerical taxonomy and multivariate analysis system. Version 2.1. Exeter Software, Setauket.  
 Saghai-Marooof MA, Soliman KM, Jorgensen RA and Allard RW (1984). Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. *Proc. Natl. Acad. Sci. U. S. A.* 81: 8014-8018.  
 SAS Institute (2000). SAS language and procedures: Usage. Version 8.1. SAS Institute, CD-ROM, Cary.

**Appendix.** Genetic similarity estimate by dichotomy analysis, using simple matching (SM) coefficient.

| Locus    | Inbred lines |   |   |
|----------|--------------|---|---|
|          | 1            | 2 | 3 |
| BLNG1318 | 0            | 0 | 0 |
| BLNG1318 | 0            | 1 | 0 |
| BLNG1318 | 1            | 1 | 1 |
| BLNG1250 | 0            | 0 | 0 |
| BLNG1250 | 1            | 0 | 1 |
| BLNG1250 | 1            | 1 | 1 |

Calculate by  
SM coefficient

$$SM = \frac{a + d}{a + b + c + d}$$

$s_{ij}$  (1 x 2)

a = 1 and 1  
c = 0 and 1

b = 1 and 0  
d = 0 and 0

so:

a = 2  
c = 1

b = 1  
d = 2

$$s_{ij} = \left( \frac{2+2}{2+1+1+2} \right) = 0.667$$