ratar. povrt.

original scientific paper / originalni naučni članak
Ratar. Povrt. 2017, 54(1): 25-30
doi:10.5937/ratpov54-12174

Comparison of Statistical Methods for Genetic Similarity Evaluation of Maize Inbred Lines

Ana Nikolić* · Marija Kostadinović · Jelena Vančetović · Goran Stanković · Dragana Ignjatović-Micić

Maize Research Institute, Zemun Polje, Slobodana Bajića 1, 11185 Belgrade-Zemun, Serbia

Summary: Conventional breeding methods have been aided by molecular genetic techniques giving the chance for efficient improvement in creation of maize hybrids. Proper choice of statistical methods for data analysis is very important because it ensures greater reliability. The aim of this study was to determine the most suitable statistical approach for molecular marker data analysis. SSR markers were used for the analysis of 10 maize inbreds. Genetic similarity/distance was calculated using three types of data: binary, allele frequency based on densitometry and allele frequency according to band size data applying Simple matching, Jaccard's and Rogers' coefficient. Cluster analysis was performed in NTSYS, 2.11a software. The highest value for Spearman's rank of correlation (0.95) was detected between distance matrices based on binary data. The results showed that binary data (Jaccard's coefficient) and allele frequency data based on fragment sizes (Rogers' coefficient) gave identical clusters by visual inspection and according to CI_c index.
Key words: coefficient of similarity/distance, genetic similarity, maize, SSR

Introduction

Characterization of different economically important plant species provides information which could be successfully exploited in breeding programs. Description of plant germplasm has been done using different type of data (morphological, biochemical, molecular) throughout the past decades.

Although molecular marker techniques did not completely fulfil the expectations regarding their assumed potential in accurate deciphering relationships among genotypes, they play indispensable role in genetic diversity assessment and there is a constant striving for their improvement. The choice of a proper molecular marker type (Jones et al., 2007; Abdellatif et al., 2010; Singh et al., 2013) as well as selection of the most adequate approach for calculating genetic distance and thus for genetic divergence determination has been the subject of discussion in many studies in different plant species (Duarte et al., 1999; Meyer et al., 2004).

Corresponding author:
anikolic@mrizp.rs

Acknowledgement:
This work was supported by the Ministry of Education, Science and Technological Development, Republic of Serbia, through the project TR31028 "Exploitation of maize diversity to improve grain quality and drought tolerance".

Maize is one of the most important crops worldwide and great attention has been dedicated to description of its genetics. In maize breeding, data about genetic distance represent valuable information for suggesting pairs of genotypes which would potentially express high level of heterosis, the phenomenon which is the basis of maize hybrids creation.

SSR markers were widely applied in maize germplasm characterization and different approaches in scoring and processing marker data were used (Liu et al., 2003; Souza et al., 2008; Ignjatović-Micić et al., 2013). SSRs (*Simple Sequence Repeat*) are still frequently in use in spite of the existence and development of third generation molecular markers – SNPs (*Single Nucleotide Polymorphism*), which have many advantages over all the other marker systems. Moreover, different genetic diversity studies on maize discovered higher polymorphism with SSR compared to SNP markers (Hamblin et al., 2007; Yang et al., 2011).

Type of coefficient of similarity/dissimilarity chosen for genetic distance calculation varies widely in different studies. Reif et al. (2005) stated that genetic and mathematical properties of coefficients should be considered before their use, because they could influence classification of genotypes when multivariate analyses such as clustering and principal component analysis are applied. Balestre et al. (2008) compared seven different coefficients for genetic similarity/

distance calculation in maize inbred lines analysed using SSR markers. They suggested that some of them should be avoided, yet not giving the best solution for the problem under study.

The objective of the study presented herein was to identify the most suitable data processing approach for genetic divergence determination with SSR markers. For this purpose, SSR analysis was performed on ten maize inbreds using different modes of data scoring and different types of coefficients for calculating genetic distance.

Materials and Methods

Ten dent maize inbred lines were analysed with 24 SSR markers. Nine of the inbreds were of the Western Balkan origin, and one was US inbred B37. Two lines (R348 and R59) were selected from variety Ruma dent, one (V395) from Vukovar dent and one (Š144) from Šid dent. These three varieties were grown in close geographical regions. The remaining five inbreds have similar origin. Inbreds i2/29 and i32/1157 derived from a variety Istra large-kernel dent. The first line originated directly from the indicated variety, while the second one was created from a cross between Istra dent and lines W32 and B1157. On the other hand, lines i171/37-121, i172/16-3 and i172/348-142 originated from Istra dent variety marked as Number 17: i171 (17/1) as first ear and i172 (17/2) as second ear. Inbred i171/37-121 was created by reselection from cross between first ear of population 17 and US line B37. Line i172/348-142 is the result of a cross between second ear of population 17 and line R348, while line i172/16-3 was created by reselection also from the second ear of population 17. Pedigree data and line abbreviations are shown in Table 1.

Genomic DNA was isolated from maize seeds following the procedure of Rogers and Bendich (1988). Amplification reaction was performed according to the modified method of Edwards et al. (1991). Fragments were separated using agarose gel electrophoresis, stained with ethidium bromide and gels were photographed under UV light using Biometra BioDocAnalyze Live gel documentation system. Fragment sizes were determined in comparison with band sizes of 20bp DNA ladder (ThermoScientific).

Alleles were recorded both as binary scores for presence/absence (1,0) and as frequencies of the amplified fragments. Furthermore, allele frequency was calculated in two ways – densitometrically by UN-SCAN–IT gel 6.1 program package and according to band sizes data using PowerMarker version 3.25. Genetic similarities/distances were established for all three types of data. Simple matching - SM (Sokal and Michener, 1958) and Jaccard coefficient (Jaccard, 1908) were used for genetic similarity calculation according to binary data. Data transformed into genetic distance matrices were subjected to cluster analysis in NTSYS 2.11a using UPGMA method in SAHN option. On the other hand, allele frequency data (frequencies based on densitometric analysis and frequencies according to band sizes) were processed for genetic distance determination and cluster construction in the same software using Rogers' genetic distance (Rogers, 1972). Similarity/dissimilarity coefficients used for analysis of SSR data are shown in Table 2.

Spearman's rank of correlation between genetic distances determined with different coefficients was calculated in Excel (Microsoft Office, 2010). Results of cluster analysis for all three ways of data scoring were compared with pedigree data by visual inspection. The consensus index (CI_c) implemented in NTSYS 2.11a was applied for cluster comparison. Mantel test was done for cophenetic correlation determination between dissimilarity and cophenetic matrices (r_c) using the same software.

Table 1. Pedigree data for the analysed genotypes

Genotype	Abbreviation	Origin
V395/31	V395	Vukovar dent
R59		Ruma dent
R348		Ruma dent
Š144		Šid dent
i2/29	i2	Istra large kernel dent
i171/37-121	i171	Istra dent
i172/16-3	il172a	Istra dent
i172/348-142	il172b	Istra dent
i32/1157	i32	Istra large kernel dent
B37		US inbred line

Table 2. Coefficients of similarity/dissimilarity

Coefficient	Expression
SM (1958)	$\frac{a+d}{a+b+c+d}$
Jaccard (1908)	$\frac{a}{a+b+c}$
Rogers (1972)	$\frac{1}{L}\sqrt{\sum_{u}\frac{(Xu-Yu)^2}{2}}$

a = 1 and 1; b = 1 and 0; c = 0 and 1; d = 0 and 0.
L = number of loci; Xu and Yu frequency of u-th allele for the lines i and j.

Results and Discussion

Genetic distance between the analysed lines was in the range from 0.21 to 0.86 for all data processing approaches (data not presented). The lowest values for genetic distance were detected between lines i171and i172a in all cases (0.21, 0.43, 0.39 for SM, Jaccard and Rogers *bs* respectively) except for frequencies based on densitometric data for which inbreds i2 and i32 were the least genetically distant (0.48). These results are in accordance with pedigree data, as i171 and i172a originate from Istra dent variety, and i2 and i32 from Istra large-kernel dent. The highest distance was calculated between different pairs of lines. Thus, R59 / i32 (0.51) and R348 / i32 (0.51) for SM, V395 / B37 (0.73) for Rogers *bs* and R59 / B37 (0.86) for both Jaccard and Rogers *dens* coefficients were the most divergent lines. In three out of four cases, the highest distance was recorded between the lines of Western Balkan origin and US B37 line, while only for SM coefficient the most divergent lines were from Istrian dent and Ruma dent origin.

The highest correlation value (0.95) was calculated between Jaccard and SM coefficients based on binary data (Table 3). Somewhat lower values were detected between binary data based coefficients (Jaccard, SM) and those based on allele frequency determined by densitometry (Rogers, *dens*). Correlation values between the two types of frequency data (Rogers *dens* and Rogers *bs*) were lower than for the above data, but higher than between binary (Jaccard and SM) and band size frequency data (Rogers *bs*). Numerous studies also showed high correlation for the coefficients implemented for dichotomic variables (Duarte et al, 1999; Balestre et al., 2008; Denčić et al., 2016). Different types of data used for the analysis (two different approaches for calculating allele frequencies) are the reason for different Spearman's coefficient values when the same coefficient of similarity/dissimilarity was applied (Rogers' coefficient).

Resulting dendrograms according to all analysed types of data are shown in Figure 1. General structure of all four dendrograms was the same. By visual inspection, similar structure of clusters can be observed. Inbred lines were grouped in two main clusters (I and II) in each dendrogram with small modifications in classification of genotypes within clusters. Clustering results were generally in agreement with pedigree data and all individuals were well separated. Cluster I was comprised of Vukovarski, Rumski and Šidski dent lines, while cluster II encompassed lines of Istarski dent origin and US line B37. The best congruence with pedigree data and identical structure by visual inspection was detected for dendrograms constructed using Jaccard (binary data) and Rogers *bs* coefficient. In these two dendrograms the US line B37 clustered separately from the lines of Istarski dent origin, while its position in dendrograms derived using SM and Rogers' *dens* coefficient was not in agreement with pedigree data. Several authors (Duarte et al., 1999; Meyer et al., 2004; Balestre et al., 2008) proposed CI_c index as a more accurate tool for cluster comparison. Values of this index are in a range from 0 to 1. Identical clusters were those constructed using Jaccard's (binary data) and Roger's *bs* coefficients - CI_c index value was 1 (Table 4). This index was lower for other pairs of clusters, although identical value was calculated for several of them. The lowest value was detected between dendrograms constructed using SM and Rogers *bs* coefficients.

Based on cophenetic correlation data (r_c), better representation of distance matrices by dendrograms was detected for binary data based coefficients (Table 5). Good fit for r_c ($0{,}8 \leq r_c \leq 0{,}9$) was detected for Jaccard and SM, while very poor fit ($r_c < 0.7$) was found for frequency data processed with Rogers coefficient. The goodness of fit was defined according to Rohlf (1992). On the contrary, Balestre et al. (2008) found uniform and high ($r_c > 0.9$) values for r_c for seven different coefficients of similarity/dissimilarity in the analysis of maize inbred lines with SSRs. They stated that according to their results this parameter was not useful in determination of the best coefficient because no significant difference in values was detected. However, our results indicate that coefficients with higher values for r_c (good fit), might be a better choice for data processing than the coefficients with low values (very poor fit).

Table 3. Spearman's coefficient between genetic distances calculated using different coefficients/different type of data

	Jaccard	SM[1]	Rogers *dens*[2]	Rogers *bs*[3]
Jaccard	-			
SM[1]	0.95	-		
Rogers *dens*[2]	0.83	0.841	-	
Rogers *bs*[3]	0.67	0.67	0.75	-

[1]SM – Simple matching, [2]*dens* – densitometry, [3]*bs* – band sizes
All values are significantly different from zero ($p < 0.001$)

Table 4. CI_c indices for different pairs of coefficients/data

CI_c index	Jaccard	SM[1]	Rogers *dens*[2]	Rogers *bs*[3]
Jaccard	-			
SM[1]	0.75	-		
Rogers *dens*[2]	0.75	0.75	-	
Rogers *bs*[3]	1	0.63	0.88	-

[1, 2, 3] - abbreviations as per Table 3

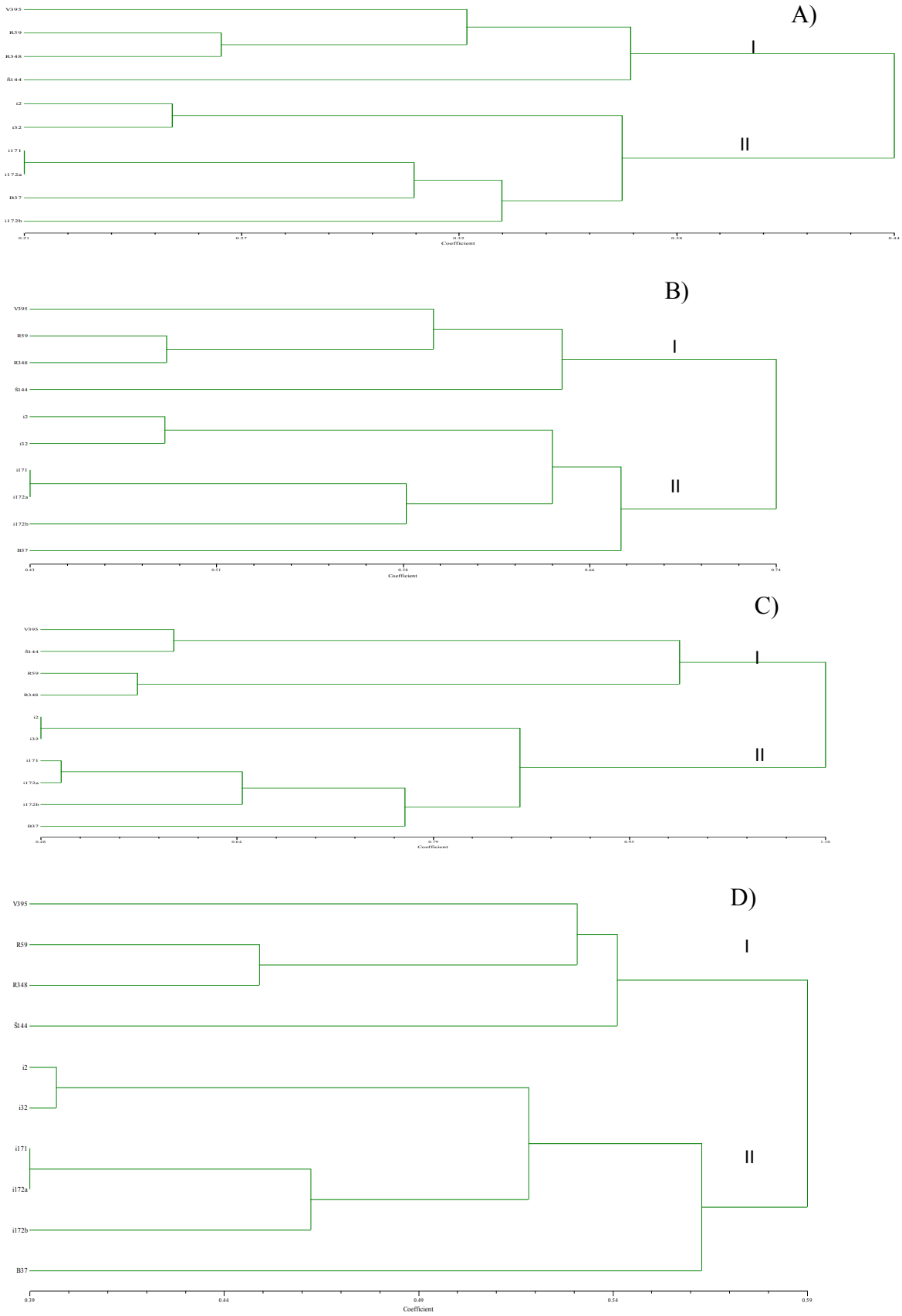

Figure 1. Dendrograms constructed from distance matrices using: A) Simle matching (SM), B) Jaccard, C) Rogers *dens* and D) Rogers *bs* data. I and II – two main clusters.

Table 5. Cophenetic correlation data

Cophenetic corr	r_c
Jaccard	0.80
SM[1]	0.85
Rogers *dens*[2]	0.67
Rogers *bs*[3]	0.62

[1, 2, 3] -abbreviations as per Table 3

Different authors studied convenience of various coefficients of similarity/dissimilarity for genetic diversity determination in diverse plant and animal species. Their research was focused predominantly on dominant marker data (RAPD, AFLP) and on coefficients suitable for dichotomic variables (Duarte et. al, 1999; Meyer et al., 2004; Dalirsefat et al., 2009; Sesli et al., 2010). Reif et al. (2005) gave detailed analysis of ten various coefficients of similarity/distance applied for germplasm surveys with molecular markers for different purposes in maize breeding pointing out several factors that should be considered prior to decision which one will be used: properties of markers used, genealogy of germplasm, objects under study (lines, populations, etc.), topic of research (for example in maize: genetic similarity of inbred lines, classification in heterotic groups, etc.) and conditions needed for multivariate analysis. They stated that Rogers' dissimilarity coefficient is convenient for pedigree determination in maize inbred lines because it is linearly correlated to coefficient of co-ancestry and also suitable for allelic informative data such as SSR. The same result was obtained in our study, as application of Rogers' *bs* dissimilarity coefficient presented dendrograms with good concurrence to pedigree data, which could recommend it for further use. However, the same agreement was obtained with Jaccard coefficient. Visual inspection of clusters and value of 1 for CI_c index suggested the same suitability of Jaccard's and Rogers *bs* coefficients. Jaccard coefficient was also favoured by a high r_c value. However, a high r_c value was also noted for SM coefficient, although it showed a poor agreement with pedigree data. This could be explained by the fact that SM coefficient considers absence of alleles in both lines under comparison and that allele absence may not be the consequence of similarity between lines yet the reason could be absence of amplification or „alleles can be identical by state but not identical by descent" (Senior et al., 1998; Li et al., 2002).

Conclusion

Presented results indicate that final decision which coefficient/type of data is the most appropriate could be a challenge, because of contradictory results related to the parameters which should help in unravelling their advantages/disadvantages. The choice of coefficient without adequate criteria could affect the results of research, taking into account the fact that application of different coefficients influences classification of genotypes. Our results showed that Rogers *dens* and SM are not a good choice for inbred lines genetic similarity determination. Rogers *bs* could be recommended in pedigree analysis. Finally, Jaccard coefficient was found to be the best choice for this kind of analysis in spite of the fact that it uses binary data while SSR data are allelic informative. These results could be the consequence of a small sample size (i.e. smaller number of detected alleles) and sometimes there is a necessity for studying samples of such extent. Thus, the choice of similarity/ dissimilarity coefficient should be done individually for each experimental design.

References

Abdellatif, K.F., & Khidr, Y.A. (2010). Genetic diversity of new maize hybrids based on SSR markers as compared with other molecular and biochemical markers. *J Crop Sci Biotechnol,* 13(3), 139-145.

Balestre, M., Von Pinho, R.G., Souza, J.C., & Lima, J.L. (2008). Comparison of maize similarity and dissimilarity genetic coefficients based on microsatellite markers. *Genetics and Molecular Research,* 7(3), 695-705. pmid:18752197

Dalirsefat, S., Meyer, A., & Mirhoseini, S. (2009). Comparison of similarity coefficients used for cluster analysis with amplified fragment length polymorphism markers in the silkworm, Bombyx mori. *Journal of Insect Science,* 9, 1-8.

Denčić, S., Depauw, R., Momčilović, V., & Aćin, V. (2016). Comparison of similarity coefficients used for cluster analysis based on SSR markers in sister line wheat cultivars. *Genetika,* 48 (1), 219-232.

Duarte, J.M., dos Santos, J.B., & Melo, L.C. (1999). Comparison of similarity coefficients based on RAPD markers in the common bean. *Genet. Mol. Biol.,* 22, 427-432.

Ignjatović-Micić, D., Ristić, D., Babić, V., Andjelković, V., Marković, K., & Vančetović, J. (2013). Genetic assessment of maize landraces from former Yugoslavia. *Genetika,* 45(2), 405-417.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.,* 44, 223-270.

Jones, E.S., Sullivan, H., Bhattramakki, D., & Smith, J.S.C. (2007). A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (Zea mays L.). *Theor. Appl. Genet.,* 115(3), 361-71. pmid:17639299

Edwards, A., Civitello, A., Hammond, H.A., & Caskey, C.T. (1991). DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *Am. J. Hum. Genet.,* 49(4), 746-56. pmid:1897522

Hamblin, M.T., Warburton, M.L., & Buckler, E.S. (2007). Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. *PLoSONE,* 12, 1367.

Li, Y., Du, J., Wang, T., Shi, Y., Song, Y., & Jia, J. (2002). Genetic diversity and relationships among Chinese maize inbred lines revealed by SSR markers. *Maydica,* 47, 93-101.

Liu, K., Goodman, M., Muse, S., Smith, S.J., Buckler, E., & Doebley, J. (2003). Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics,* 165(4), 2117 -28. pmid:14704191

Meyer, A.S., Garcia, A.A.F., Souza, A.P., & Souza, C.L. (2004). Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (Zea mays L). *Genet. Mol. Biol.,* 27, 83-91.

Reif, J.C., Melchinger, A.E., & Frisch, M. (2005). Genetical and Mathematical Properties of Similarity and Dissimilarity Coefficients Applied in Plant Breeding and Seed Bank Management. *Crop Sci.,* 45, 1-7.

Rogers, J.S. (1972). Measures of genetic similarity and genetic distance. In *Studies in genetics VII*.. (pp. 145-153). Austin: Univ. of Texas. Publ. 7213.

Rogers, S.O., & Bendich, A.J. (1988). Extraction of DNA from plant tissues. *Plant Molecular Biology Manual,* A6, 1-10.

Rohlf, F.J. (1992). *NTSYS-pc (Numerical Taxonomy and Multivariate Analysis System). Version 1. 70*. Setauket, NY: Exeter.

Singh, N., Choudhury, D.R., Singh, A.K., Kumar, S., Srinivasan, K., Tyagi, R.K., Singh, R. (2013). Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of Indian rice varieties. *PLoS ONE,* 8(12), 84136. pmid:24367635

Sokal, R.R., & Michener, C.D. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas. Sci. Bull.,* 38, 1409-1438.

Senior, M.L., Murphy, J.P., Goodman, M.M., & Stuber, C.W. (1998). Utility of SSRs for determining genetic similarities and relationships in maize using and agarose gel system. *Crop Sci.,* 38, 1088-1098.

Sesli, M., & Yegenoglu, E.D. (2010). Comparison of similarity coefficients used for cluster analysis based on RAPD markers in wild olives. *Genetics and molecular research,* 9(4), 2248-53. pmid:21086261

Souza, S.G.H., Carpentieri-Pípolo, V., Ruas, C.F., Carvalho, V.P., & et al., (2008). Comparative analysis of genetic diversity among the maize inbred lines (Zea mays L.) obtained by RAPD and SSR markers. *Braz. Arch. Biol. Technol,* 51, 183-192.

Yang, X., Xu, Y., Shah, T., Li, H., Han, Z., Li, J., & Yan, J. (2011). Comparison of SSRs and SNPs in assessment of genetic relatedness in maize. *Genetica,* 139(8), 1045-54. pmid:21904888

Poređenje statističkih metoda za određivanje genetičke srodnosti samooplodnih linija kukuruza

Ana Nikolić · Marija Kostadinović · Jelena Vančetović · Goran Stanković · Dragana Ignjatović-Micić

Sažetak: Metode klasične selekcije kukuruza se dopunjuju tehnikama molekularne genetike u cilju efikasnijeg dobijanja pouzdanih rezultata, pri čemu pristupi u obradi podataka imaju veliki značaj u ostvarivanju ovog cilja. Primenom SSR molekularnih markera analizirano je 10 samooplodnih linija kukuruza. Različite statističke metode su upoređene sa ciljem da se utvrdi najpogodnija za određivanje genetičke srodnosti ispitivanih genotipova. Genetička sličnost/distanca je izračunata korišćenjem tri tipa podataka: binarni podaci (1,0), frekvencija alela izračunata pomoću denzitometrije i frekvencija alela izračunata na osnovu veličina umnoženih fragmenata u baznim parovima primenom Simple matching, Jaccard i Rogers koeficijenata. Klaster analiza je urađena u NTSYS, 2.11a softveru. Najveća vrednost Spirmanovog koeficijenta (0.95) je utvrđena između matrica genetičkih sličnosti/distanci izračunatih na osnovu binarnih podataka. Rezultati ukazuju da se identični dendrogrami dobijaju korišćenjem Jaccard-ovog koeficijenta izračunatog za binarni zapis i Rogers-ovog koeficijenta izračunatog na osnovu frekvencija alela određenih prema veličini umnoženih fragmenata, kako vizuelnom ocenom tako i na osnovu CI_c indeksa.

Ključne reči: genetička srodnost, koeficijenti sličnosti/distance, kukuruz, SSR

Received: 24 October 2016, Accepted: 3 February 2017
Published online: 27 February 2017