Available online at www.elixirpublishers.com (Elixir International Journal)

Statistics

Elixir Statistics 69 (2014) 23111-23116

ISSN: 2229-712X

Comparison of similarity coefficients and clustering methods with amplified fragment length polymorphism markers in *Colletotrichum gloeosporioides* isolates from yam

Ojurongbe Taiwo Adetola[1*], Aduramigba-Modupe Adefoyeke Olufunmilayo[2], Schachtel Gabriel[3] and Frisch Matthias[3]

[1]Department of Mathematical and Physical Sciences, Osun State University, Osogbo Nigeria.

[2]Department of Crop Protection & Environmental Biology, University of Ibadan, Ibadan Nigeria.

[3]Department of Biometry and Population Genetics, Justus Liebig University, Giessen, Germany.

ARTICLE INFO

Article history:

Received: 27 February 2014;

Received in revised form:

29 March 2014;

Accepted: 10 April 2014;

Keywords

Similarity,
Clustering,
Dendrogram,
Polymorphism,
Markers,
Isolates.

ABSTRACT

The choice of the similarity coefficient used in clustering could have great impact on the resulting classification, there is need to study and understand these coefficients better to be able to make the right choice for specific situations. In this study, variations caused by three similarity coefficients: Dice, Jaccard and Simple matching with five clustering methods: (Unweighted Pair-Group Mean Arithmetic (UPGMA), Weighted Pair-Group Mean Arithmetic(WPGMA), complete linkage, single linkage and Neighbour-Joining with AFLP markers in *Colletotrichum gloeosporioides* isolates from yam were assessed. Comparison among the similarity coefficients and clustering methods were made using correlation analysis, multidimensional scaling and principal component analysis. Dendrogram topology was compared using consensus fork index (CFI) and node counts. The grouping of the pathogens by the markers is not related to their agro-ecological zones. The CFI results showed varying level of similarity for the cluster analysis CA methods. It was observed that high correlation does not necessarily imply similarity in the topology of a tree, therefore care should be taken in its interpretation. The cophenetic correlation with original distances suggests that the UPGMA method gives consistent results with respect to grouping irrespective of the similarity coefficient. The use of UPGMA method is therefore recommended for its consistency.

Introduction

The choice of an appropriate coefficient of similarity is a very important and decisive point to evaluate clustering, true genetic similarity between individuals, analyzing diversity within populations and studying relationship between populations, because different similarity coefficients may yield conflicting results (Kosman and Leonard, 2005). Methods, such as cluster analysis (CA), factor analysis (FA), discriminant analysis (DA) and principal component analysis (PCA) can be applied in studies of divergence and phylogenetic relationships between and within plant pathogen populations. Among these methods, CA stands out as it does not demand an initial hypothesis with respect to the probability distribution of the data and it provides easy interpretation (Meyer *et al*., 2004). CA helps to identify objects that are similar to one another, based on some specified criteria that define a population. It divides data into groups that are meaningful, useful or both (Tan et al., 2006). However, in some cases, CA is just a useful starting point for other purposes, such as summarization or multivariate analysis of data.

A prerequisite of CA for many methods is the construction of similarity/dissimilarity coefficients between the individuals or objects being considered. The choice of a similarity/dissimilarity coefficient for studying divergence depends on the marker system properties involved, the germplasm genealogy, the taxonomic operational unit involved, the study objectives and on the conditions that are necessary for multivariate analyses (Reif *et al*., 2005). Taking into consideration that the results of clustering can be influenced by the choice of a similarity/ dissimilarity coefficient (Duarte *et al*., 1999; Jackson *et al*., 1989; Meyer *et al*., 2004), it is needful that these coefficients be better understood, so that the most efficient ones can be applied in specific situations. Therefore the knowledge of the genetical and mathematical properties as well as the application of these coefficients in different situations is important.

CA has been applied to many practical problems such as generating similarity matrices for molecular markers based on the absence or presence of a band in other to confirm inherent groupings Molecular markers have been widely used for the purpose of characterizing genetic diversity within or between populations or groups of individuals because they typically detect high levels of polymorphism. Random amplified polymorphic DNA (RAPDs) and Amplified fragment length polymorphisms (AFLPs) are efficient markers that allow multiple loci to be analysed for each individual in a single gel run. AFLP analysis is useful in identifying genetic diversity and analysis of population structure within complex genera of fungi (Aduramigba-Modupe *et al*., 2012) such as *Colletotrichum gloeosporioides* responsible for Anthracnose disease of yam.

Yams (*Dioscorea spp*.) constitute an economically staple food for millions of people in the tropics and subtropics (Abang et al., 2003). West Africa accounts for about 95% of world production and 93% of the total yam production area (FAO, 2002). Nigeria leads with 75% of the world's yam production ((FAO, 1999; IITA, 2000) and the two most important cultivated edible yams are white Guinea yam (*D. rotundata* Poir) and

Tele:
E-mail addresses: taojurongbe@yahoo.com

water yam (*D. alata* L.). *D. rotundata* is indigenous to West Africa while D. *alata* that was introduced to Africa from Asia in the 16th century was regarded as the most widely cultivated species globally. However, its major drawback in the field is the susceptibility of most cultivars to anthracnose disease which has a great impact on its productivity. Anthracnose affects the leaves, petioles, stems and veins of the plant, causing leaf spots, leaf blotches, petiole blights, premature abscission, dieback and eventual death of the entire plant. The disease usually has a dramatic effect on infected plants, converting a field of initially healthy yam plants from 'green' to 'black' within a few weeks (Green and Simons, 1994).

This study therefore investigated the effect of different similarity coefficients and clustering methods on binary data generated from AFLP markers *Colletotrichum gloeosporioides* isolates from yam using AFLP markers

Materials and Method

Ten primers were used to determine the genetic variation among isolates of *Colletotrichum gloeosporioides* from yam. Three of which were polymorphic and the resulting data were used to form three data sets namely; ACMA, AAMG and AAMO. Each data set had pathogens of the anthracnose disease from two different geographical locations; the Forest and Guinea Savannah.

The AFLP marker was analysed using a modified method of Vos *et al.,* (Vos *et al*., 1995) with 10 enzyme-primer combinations out of which three were polymorphic: EAA/MO, EAC/MA and EAA/MG. Only the polymorphic bands were used for the construction of binary value matrices, where the absence and presence of bands were represented by 0 and 1 respectively. Each band was considered a locus and the three sets of data resulting from the polymorphic primer combinations were named: AAMO, ACMA and AAMG respectively. AAMO has 30 pathogens with 20 bands; ACMA has 32 pathogens with 17 bands while AAMG has 27 pathogens with 21 bands. Grouping of the pathogens based on AFLP marker analysis was on the basis of origin of the pathogens, whether from the Humid Forest or Guinea Savannah region in Nigeria.

Similarity estimates between each pair of pathogens (i,j) were obtained for three similarity coefficients: Dice, Jaccard and Simple Matching (Table 1). Three similarity matrices were constructed from the resulting data. For each sample generated, dendrograms (trees) were constructed using UPGMA, WPGMA, NJ, single linkage and complete linkage cluster analysis (CA) methods for the Dice, Jaccard and Simple matching coefficients using NTSYS software (Rohlf, 2002). The aim was to see if the agro-ecological zones of these pathogens will still be reflected in the groups formed by CA and to see the effect of these similarity measures and CA methods on the resulting groupings. Cophenetic matrices of the trees were also calculated. The Consensus Fork Index (CFI) (Colless, 1980) was calculated to measure the similarity of the corresponding pairs of Dice, Jaccard and Simple matching trees. The CFI is defined as

$CFI = c / (n - 2)$

Where c is the total number of clusters (partitions) in the consensus tree, with the exception of the total set, and the subsets where the elements are separate, n is the total number of objects in the clusters and n-2 is the maximum groupings or clusters possible. It is a measure of dendrogram similarity that expresses the proportion of sub-clusters shared by two dendrograms, ranging from zero, if no sub-clusters are shared, to one, if all sub-clusters are shared (Angielczyk and Fox, 2006). Therefore CFI was used to compare the similarity among the constructed dendrograms for the different similarity measures and CA methods. Multidimensional scaling (MDS) and Principal Component Analysis (PCA) were also carried out to compare the groupings.

Other Measures of Comparing Topology of Trees Used

(i). Pearson and Spearman Correlation coefficients were calculated for the cophenetic matrices of the data with respect to the afore-mentioned methods of clustering to compare the trees constructed using the Dice, Jaccard and Simple-Matching similarity measures.

(ii). Node count matrices were generated for the Dice, Jaccard and Simple-Matching trees experimental data sets. The different matrices for each data set were converted into a vector each and the Pearson and Spearman correlation coefficients were calculated for the UPGMA, WPGMA, Single and Complete linkage methods of clustering.

(iii). Node count values and cophenetic values for each similarity measure were combined and the Pearson and Spearman correlation coefficients calculated between the two measures for the different methods of clustering.

Table 1: Similarity coefficients used for the AFLP markers

Coefficient	Expression	Occurrence Interval	Source
Jaccard	a/ (a+b+c)	[0,1]	Jaccard, 1901
Dice	2a/ (2a+b+c)	[0,1]	Dice, 1945
Simple Matching	(a+d)/ (a+b+c+d)	[0,1]	Sokal and Michener, 1958

Results and Discussion

A visual inspection of the dendrograms revealed a high level of similarity among those generated using the Dice and Jaccard measures. However, those constructed using the Simple matching coefficient showed some distinct differences (Figure 1) corroborating the similarity differences between the three measures (Duarte *et al.,* 1999; Jackson *et al.,* 1989; Meyer *et al.,* 2004). These differences were revealed in the alterations in the levels in which the individuals are clustered (Figure 1). Previous works which had been carried out on the construction of dendrogram using binary data involving about eight similarity measures which were divided into different groups according to whether the similarity measure excludes or includes negative co-occurrences of the objects being compared in their calculations also confirmed the differences ((Balastre *et al.,* 2008; *Meyer et al.,* 2004). These studies have also shown the diversity in their conclusions about the comparison of similarity coefficients, leading to a general acceptance that the behavior of these coefficients is specific to data (Jackson *et al.,* 1989) which was also observed in this study. In previous studies, it was observed that the Dice and Jaccard coefficients are highly correlated and a visual inspection of the dendrograms obtained with the UPGMA method shows that the dendrograms constructed using the Dice and Jaccard coefficients present similar clustering structures (Duarte *et al.,* 1999; Meyer *et al.,* 2004).

One of the criteria for choosing the most appropriate coefficient of similarity depends on type of marker and ploidy of the organism under consideration (Kosman and Leonard, 2005). (Landry and Lapointe, 1996) suggested that the Dice or Jaccard coefficients might be a better choice to the Simple matching coefficient when RAPD analysis are used to compare groups of distantly related taxa. However, based on our result using AFLP markers, it was discovered that the Dice or Jaccard similarity coefficient could also be given a preference over the Simple matching coefficient for such markers. The Jaccard measure proved to be a better choice from the results in our study. Having observed that the Dice and Jaccard measure could be used interchangeably with little or no difference, the choice depends on the interest of the researcher. The Simple matching coefficient was suggested to be the more appropriate measure of

similarity when closely related taxa are considered (Hallden *et al*., 1994), but (Kosman and Leonard, 2005) believe that the choice of a similarity coefficient should be supported with estimates of DNA sequence identity between the taxa. If there are no supporting sequence identity estimates, then similarity values based on dominant markers data should be regarded as tentative (Dalirsefat *et al*., 2009) .

The dendrograms constructed showed there was a mixture of the pathogens from the different agro-ecological zones (figure 1) suggesting that the location of the pathogens were not preserved after classification and that the grouping of the pathogens by the markers is not perfectly related to their agro-ecological zones. In the ACMA primer data, UPGMA, complete and single linkage methods produced identical classifications for both Dice and Jaccard measures while WPGMA and NJ methods did not. However, in the AAMG primer data, only the NJ method did not result in identical classifications for Dice and Jaccard measures while in the AAMO primer data, NJ and WPGMA methods did not give identical classifications for the two measures. This observation supports the fact that different primers amplify markers differently which was also revealed in the resulting classifications. This result also reflected the fact that not all clustering methods will produce identical classification for Dice and Jaccard measures.

The comparison of the constructed dendrograms by the Consensus fork index (CFI), allows a refinement of what was observed through visual inspection. This is similar to the observations of previous authors (Balastre *et al.,* 2008; *Duarte et al.,* 1999; Dalirsefat *et al.,* 2009; Meyer *et al.,* 2004). By this index that ranges between 0 and 1, two dendrograms are considered identical when the CFI value equals one and otherwise if not.

The CFI comparing the topology of Dice and Jaccard dendrograms for all experimental data for the UPGMA method ranged between 0.89 and 1. For the WPGMA method, the range of the CFI was between 0.64 and 1; for single linkage method it was 0.21 and 1; complete linkage method had a range of 0.93 and 1 while for the NJ method, it ranged between 0.39 and 0.68. All the methods with the exception of the NJ had the highest value of 1 for the CFI. This might not be unconnected with the fact that the NJ method produces unrooted trees (Kumar and Gadagkar, 2000) while the others produced rooted trees (Knipe and Howley, 2007). However, among the rooted trees, the single linkage method produced the least similar trees. The single linkage method is well known for producing a long chain dendrogram with lots of singletons, small clusters or outliers (Stuetzle and Nuggent, 2007), this is well reflected in the CFI values. The complete linkage and the UPGMA methods tend to produce trees that are somehow similar, which was also reflected in the CFI values. In general, the UPGMA method produced the highest number of identical trees with the CFI value of 1, reflecting the usefulness of this method in detecting the similarity in the topology of trees. The single linkage and the NJ methods had least occurrences of identical trees. Based on our results, these two methods are therefore not advised to be used for classification for data of the type used in this study. However, because of the advantage of the NJ method in handling large data, it could be used when dealing with very large data and if the researcher has interest in unrooted trees. As previously reported, the NJ method is recommended when the branch length of objects are important (Saitou and Nei, 1987). However, the method has the disadvantage of producing only one type of tree.

The CFI values for the Dice and Simple matching dendrograms were very low. These CFI values for dendrograms between Dice and Simple matching as well as Jaccard and Simple matching also confirm the suggested similarity between the Jaccard and Dice measures. However, even though in cases where the classification produced by the single linkage method was identical for Dice and Jaccard measures, the CFI value was not 1. The numerous singletons produced by the single linkage method could be responsible for this, since the formula for calculating the CFI is the number of subsets found in the two trees being compared divided by the total number of objects minus 2. This suggests the single linkage method might not be recommended because the result is not completely reliable.

However, in the AAMG primer, only the NJ method did not result in identical classifications for Dice and Jaccard while in the AAMO primer data, NJ and WPGMA methods did not give identical classifications for the coefficients. WPGMA also gave three clusters for Simple matching and five clusters for Dice and Jaccard. The comparison of the constructed dendrograms by the CFI allows a refinement of what was observed through visual inspection. Similar results were obtained in previous studies (Balastre *et al.,* 2008; Duarte *et al.,* 1999; Meyer et al., 2004) However, none of these studies was on isolates from yam. In the ACMA and AAMO primer data, the UPGMA, complete linkage and single linkage methods gave the same classifications for both Dice and Jaccard measures while the WPGMA and NJ methods gave different classifications. In all the data, the classification for the Simple matching coefficient was different from that of Dice and Jaccard for all methods. Some level of closeness were also observed with dendrograms generated using the UPGMA, WPGMA and complete linkage methods. However, the dendrograms constructed using the single linkage and NJ methods were quite different.

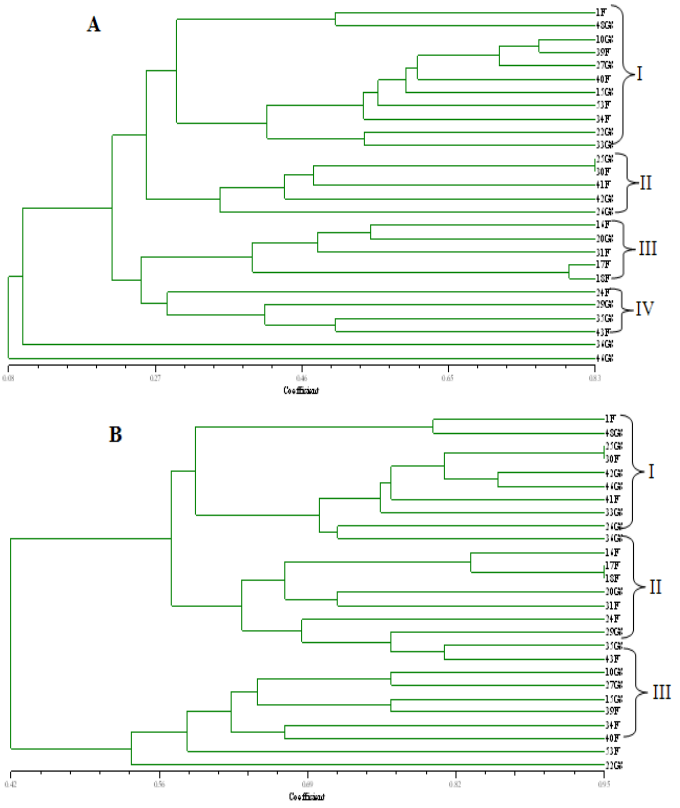

Figure 1: AAMG dendrogram for Jaccard and Simple matching coefficients (UPGMA).
A – Jaccard and B – Simple matching

Table 2: CFI summary for different methods of CA

A

Source	No of Isolates	No of differentials	UPGMA	WPGMA	SINGLE	COMPLETE	NJ
AAMG-primer	27	21	1.00	1.00	0.72	0.96	0.56
AAMO-primer	30	20	1.00	0.90	0.77	0.93	0.40
ACMA-primer	32	17	0.93	0.64	0.93	0.96	0.68

B

AAMG-primer	27	21	0.36	0.16	0.24	0.24	0.36
AAMO-primer	30	20	0.57	0.50	0.21	0.64	0.46
ACMA-primer	32	17	0.50	0.40	0.23	0.33	0.37

C

AAMG-primer	27	21	0.36	0.16	0.24	0.24	0.48
AAMO-primer	30	20	0.54	0.54	0.21	0.64	0.43
ACMA-primer	32	17	0.50	0.40	0.23	0.33	0.43

A - CFI values for Dice and Jaccard, **B** – CFI values for Dice and Simple matching and **C** – CFI values for Jaccard and Simple matching.

Table 3: Pearson correlation coefficients for Dice and Jaccard for different CA methods

A

Source	No of Isolates	No of differentials	UPGMA	WPGMA	Single linkage	Complete linkage
AAMG-primer	27	21	0.6634	0.5815	0.9586	0.6203
AAMO-primer	30	20	0.9858	0.9857	0.9928	0.9860
ACMA-primer	32	17	0.9867	0.9737	0.9964	0.9873

AAMG-primer	27	21	0.5811	0.5942	1.0	1.0
AAMO-primer	30	20	0.9834	1.0	1.0	1.0
ACMA-primer	32	17	1.0000	0.9353	1.0	1.0

AAMG-primer	27	21	0.8705	0.8949	0.9743	0.9961
AAMO-primer	30	20	0.9945	0.9999	0.9999	0.9999
ACMA-primer	32	17	0.9999	0.9814	0.9999	0.9999

A – Correlation coefficients for cophenetic distances, **B** – Correlation coefficients for node counts and **C** – Correlation coefficients for combination of cophenetic distances and node counts.

Table 4: Spearman correlation coefficients for Dice and Jaccard for different CA methods

Source	No of Isolates	No of differentials	UPGMA	WPGMA	Single linkage	Complete linkage
AAMG-primer	27	21	0.8749	0.7778	1.0	1.0
AAMO-primer	30	20	0.9741	1.0	1.0	1.0
ACMA-primer	32	17	1.0000	0.8634	1.0	1.0

AAMG-primer	27	21	0.5434	0.5335	1.0	1.0
AAMO-primer	30	20	0.9805	1.0	1.0	1.0
ACMA-primer	32	17	1.0000	0.9286	1.0	1.0

AAMG-primer	27	21	0.9281	0.9150	1.0	1.0
AAMO-primer	30	20	0.9944	1.0	1.0	1.0
ACMA-primer	32	17	1.0000	0.9746	1.0	1.0

A – Correlation coefficients for cophenetic distances, **B** – Correlation coefficients for node counts and **C** – Correlation coefficients for combination of cophenetic distances and node counts.

Table 5: Correlation coefficients from cophenetic matrices and original distances for all data

Data	Method/Similarity	Dice	Jaccard	SM
AAMG	UPGMA	0.72	0.77	0.76
	WPGMA	0.69	0.73	0.73
	Single linkage	0.59	0.63	0.69
	Complete linkage	0.56	0.62	0.68
	NJ	0.37	0.29	0.67
AAMO	UPGMA	0.91	0.93	0.75
	WPGMA	0.88	0.92	0.66
	Single linkage	0.87	0.89	0.44
	Complete linkage	0.82	0.84	0.67
	NJ	0.63	-0.05	0.62
ACMA	UPGMA	0.81	0.83	0.74
	WPGMA	0.74	0.80	0.56
	Single linkage	0.73	0.73	0.64
	Complete linkage	0.63	0.69	0.63
	NJ	0.48	0.78	0.33

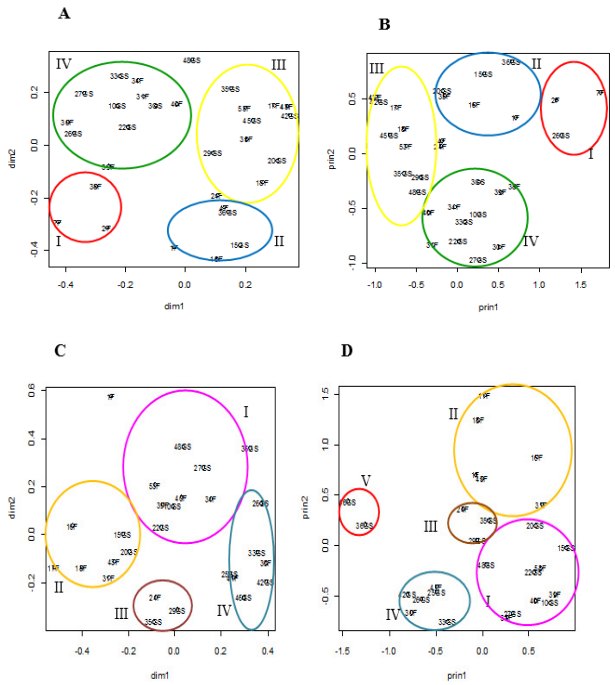

Figure 2: ACMA and AAMG MDS & PCA prin1 versus prin2 plot

A – Jaccard MDS plot for ACMA, B –Jaccard PCA plot for ACMA, C – Dice MDS plot for AAMG and D – Dice PCA plot for AAMG

Comparative results of the MDS and PCA for ACMA and AAMO primer data showed a similar grouping. The ACMA and AAMO primers produced four clusters in the MDS and PCA plots for all similarity coefficients. However, in the AAMG, the three coefficients had different groupings. The Dice had four groupings in the MDS plot and five in the PCA plot (Figure 2); the Jaccard had four groupings for both MDS and PCA plots while Simple matching had three groupings for both the MDS and PCA plots. The differences observed in the groupings for the Dice measure in the MDS plot and the PCA plot suggests that this measure could be unstable unlike the Jaccard measure that produced the same groupings for the same plots in all other experimental data. This could also be one of the reasons the Jaccard measure is more widely used among researchers for CA because of its stability and easy interpretation. The MDS plot and the first two PCA axes plots were also similar in this case as was seen in the results from other data except for the AAMG data for Dice coefficients. In general, the bi-dimensional plots indeed confirmed the classification observed in the dendrograms for all data sets (figure 2).

From the PCA results, in the ACMA primer data, the first three principal axes accounted for 80%, 72% and 80% for the Dice, Jaccard and Simple matching measures respectively; in the AAMG primer data, they accounted for 77%, 66% and 82% in a similar order and in the AAMO primer data, they accounted for 85%, 77% and 88%.

A summary of the CFI for the different CA methods is given in table 2. Comparing the values of the correlation coefficients in tables 3 and 4 with the CFI summary revealed that high correlation does not necessarily imply similarity in the topology of a tree. The Pearson correlation coefficients for Dice and Jaccard for the different CA methods (table 3) revealed that the AAMG primer had low values for the UPGMA and WPGMA methods but higher values for the Single and Complete linkage methods. However, the Spearman correlation coefficients (table 4) revealed that the Dice and Jaccard values for single linkage and complete linkage methods are perfectly monotonically related. Correlation coefficients from cophenetic matrices and original distances are shown in table 5. It was observed that only the NJ method gave a negative value in this case while the UPGMA consistently gave the highest CFI value. This could also serve as a note caution in the use of the NJ method in classification.

Conclusion

In all of the data sets, it was observed that high correlation does not necessarily imply similarity in the topology of a tree, therefore care should be taken in its interpretation. The cophenetic correlation with original distances suggests that the UPGMA method gives consistent results with respect to grouping irrespective of the similarity measure/coefficient. However, the combination of the Jaccard coefficient and the UPGMA method was observed to give a higher cophenetic correlation value for all data possibly explaining why many researchers prefer to use this combination more often especially in cases that relate to different types of markers. We will therefore recommend the use of UPGMA method because of its consistency. The Pair-wise comparison which measures similarity of two individuals and the clustering method, which measures the similarity of groups may both have big impact on the results of classification. Therefore there is need to carefully select these two options depending on the data and purpose of research.

References

Abang, M.M., Winter, S., and Mignouna, H.D. (2003). Molecular taxonomic, epidemiological and population genetic approaches to understanding yam anthracnose disease. Afr. J. Biotechnol. *2*, 486–496.

Aduramigba-Modupe, A.O., Asiedu, R., Odebode, A., and Owolade, O. (2012). Genetic diversity of Colletotrichum gloeosporioides in Nigeria using amplified fragment length polymorphism AFLP markers. African Journal of Biotechnology *11*, 8189–8195.

Angielczyk, K.D., and Fox, D.L. (2006). Exploring new uses for measures of fit of phylogenetic hypotheses to the fossil record. Paleobiology *32*, 147–165.

Balastre, M., Von Pinho, R.G., Souza, J.C., and Lima, J.L. (2008). Comparison of maize similarity and dissimilarity genetic coefficients based on microsatellite markers. Genetics and Molecular Research. Genet. Mol. Res. *7*.

Colless, D.H. (1980). Congruence between morphometric and alloyzyme data for Menidia species: A reappraisal. Syst Zool *29*, 288–299.

Dalirsefat, S.B., da Silva Meyer, A., and Mirhoseini, S.Z. (2009). Comparison of similarity coefficients used for cluster analysis with amplified fragment length polymorphism markers in the silkworm, Bombyx mori. J. Insect Sci. Online *9*, 1–8.

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

FAO (1999). FAOSTAT agriculture data. Food and Agriculture Organisation of the United Nations.

FAO (2002). FAOSTAT agriculture data. Food and Agriculture Organisation of the United Nations.

Green, K.R., and Simons, S.A. (1994). "Dead skin" on yams (Dioscorea alata) caused by Colletotrichum gloeosporioides. Plant Pathol. *43*, 1062–1065.

Hallden, C., Nilsson, N.O., Rading, I.M., and Sall, T. (1994). Evaluation of RFLP and RAPD markers in a comparison of Brassica napus breeding lines. Theor. Appl. Genet. *88*, 123–128.

IITA (2000). Annual report of project 5: Improvement of yam-based systems (Ibadan, Nigeria, International Institute of Tropical Agriculture).

Jackson, A.A., Somers, K.M., and Harvey, H.H. (1989). Similarity coefficients: measures for co-occurrence and association or simply measures of co-occurrence? Am Nat *133*, 436 – 453.

Knipe, D.M., and Howley, P.M. (2007). Fields virology (Lippincott Williams and Wilkins).

Kosman, E., and Leonard, K.J. (2005). Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. Mol. Ecol. *14*, 415–424.

Kumar, S., and Gadagkar, S.R. (2000). Efficiency of the Neighbor-Joining Method in Reconstructing Deep and Shallow Evolutionary Relationships in Large Phylogenies. J Mol Evol *51*, 544–553.

Landry, P.A., and Lapointe, F.J. (1996). Landry PA, Lapointe FJ. 1996. RAPD problems in phylogenetics. Zoologica Scripta *25*, 283–290.

Meyer, A., Garcia, A., Pereira de Souza, A., and Lopes de Souza Jr., C. (2004). Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (Zea mays L). Genetics and Molecular Biology *27*, 83–91.

Reif, J.C., Melchinger, A.E., and Frisch, M. (2005). Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management.(Review & Interpretation). Crop Sci.

Rohlf, F.J. (2002). NTSYS-pc numericcal taxonomy and multivariate analysis system. (Applied Biostatistics Inc., New York).

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. *4*, 406–425.

Stuetzle, W., and Nuggent, R. (2007). A generalized single linkage method for estimating the cluster tree of a density. (University of Washington).

Tan, P., Steinbach, M., and Kumar, V. (2006). Introduction to data mining (Addison-Wesley) (Addison-Wesley).

Vos, P.R., Hogers, M., Blecker, M., Reijans, T.V.D., Lee, M., Kuiper, M., and Zabeaus, M. (1995). AFLP a new technique for DNA fingerprinting. Nucleic Acid Res *23*, 4407–4414.