Tanzania Journal of Science 44(3): 93-102, 2018 ISSN 0856-1761, e-ISSN 2507-7961
© College of Natural and Applied Sciences, University of Dar es Salaam, 2018

SIMILARITY COEFFICIENTS INFLUENCE THE DELIMITATION OF SPECIES IN THE GENUS *ALOE* L. (XANTHORRHOEACEAE)

Halima M Amir[1] and Mkabwa LK Manoko[2]
1. Department of Biological Sciences, Dar es Salaam University College of Education
2. , Department of Botany, University of Dar es Salaam
halima_amir@yahoo.com

ABSTRACT

The present study was designed to establish the influence of coefficients indices in delimiting species using phenetic approach based on morphological data. Data were collected from thirty nine Aloe species described in Flora of Tropical East Africa. A total of forty two qualitative and quantitative characters were compiled from 83 specimens of aloes. Ten coefficient indices were tested. Data were analysed using UPMA approach of PAST software. The analyses confirmed that, truly coefficient of indices influenced the resulting classification. Of the 10 coefficients used only three produced 36-38 of the 39 species. Almost a third produced less than ten species-specific clusters another third producing less than 25 species–specific clusters. The best coefficients were Gower, Hamming and Rho whereas Chord, Correlation and Euclidean were the worst. These findings are comparable to other similar studies.

Key words: Phenetics, morphological characteristics, cluster analysis, coefficient index, phenogram, *Aloe*

INTRODUCTION

Aloe species are succulent, perennial plants that are either herbs, shrubs or small trees. They belong to the family Xanthorrhoeaceae, sub family Asphodeloideae order Asparagales, of the monocot clade (Judd et al. 2002, APG III 2009). Globally, over 560 species of *Aloe* have been described, 83 of which occur in East Africa (Carter 1994, Grace 2013). Carter (1994) reported 39 species of *Aloe* in Tanzania, 33% of which are endemic. However, field surveys in Pangani, Arusha and Mbeya recently recorded species previously not known to occur in Tanzania (Wabuyele 2006, McCoy and Lavranos 2007). These authors also described novel species. This suggests a possibility of existence of undiscovered species.

Many *Aloe* species are widely used locally as medicine and as ingredients in pharmaceutical and cosmetic industries indicating difference in their chemical composition (Wabuyele 2006, Grace et al. 2009). The wide use of *Aloe* species necessitates their proper classification and identification.

Taxonomy of *Aloe* based on morphological data is considered problematic due to close morphological similarities among species coupled with hybridization (Carter 1994). As a result, the use of DNA markers to delimit species boundaries is growing (Chase et al. 2000, Fikre 2006, Wabuyele 2006). Some morphological characteristics however, have supported the circumscription of the maculate species and the sectional evolutionary relationships between tropical and subtropical species (Grace 2009).

In practice, morphologically defined groups have been used to provide baselines for

molecular variation studies (Mishler 2000). In fact, many taxa established on the basis of morphological data have been supported by molecular data and vice versa (Hillis and Wiens 2000, Furin and Wunder 2004, Sun and Downie 2010, Manoko 2018). In the later study, morphological data of *Solanum* species analyzed using phenetic approach recovered species that were previously recognized based on AFLP markers by Manoko (2007). Some taxonomists blame morphological data when complexity in delimiting species occurs (Carter 1994, van der Bank and van Wyk 1996). However, some studies (e.g. Manoko 2018) have shown shown that it was the selection of similarity coefficients and method of coding that mattered and not the morphological data themselves. In the phenetic classification of *Solanum* sect. *Solanum*, Manoko (2018), showed that the resulting classification matched the one obtained using AFLP markers only when Gower or Hamming coefficients indices were used. In fact, these two coefficients were not influenced with coding. Similarly, Jackson et al. (1989) concluded that the choice of measures of similarity in cluster analysis greatly affected the results of analysis. This is not only when morphological data are used; different similarity coefficient used with molecular data reported to influence the results of cluster analysis too (Duarte et al. 1999, Meyer et al. 2004). Although some workers have called for comparative studies on the consequences of choosing particular similarity coefficient (Hubalek 1984, Gower and Legendre 1986), similar studies are lacking in the genus *Aloe*. The current study was designed thus to access clustering patterns of individuals in the genus *Aloe* under the influence of the different similarity coefficient and consequently the delimitation of species.

MATERIAL AND METHODS

Table 1 presents a list of individuals and character states used in the study. In total 83 individuals belonging to thirty nine known species of *Aloe* were included in the present study. Forty two characteristics both qualitative and quantitative were included in the data matrix. Characters and characters states of the respective species were extracted from descriptions in the Flora of Tropical East Africa (Carter 1994). Since character states are manifestation of character themselves, the number of character states for qualitative characteristics reflected the manifestation of each character. Coding depended on whether character had two or more character states. Characteristics with only two character states were coded as binary and multistate characters were coded using Conventional coding method. For example, perianth type had four character states which were coded as follows: Cylindrical trigonous (0), Cylindrical (1), Slightly trigonous (2) and Trigonous (3). The maximum number of character states for each species determined the number of individuals per species to be included in the analysis. This made at least two individuals per species. For quantitative characteristics upper and lower limits of the character were considered to represent two character states. In a situation where three individuals were desired a mean between the lower and the upper limit of each value represented a character state of the third individual, and the code of the quantitative character was the value recorded. All Tanzanian species described by Carter (1994) were included in the study. On Table 1, column two presents the name of the species and the authority based on Newton and Rowley (2001), column three is the code used in the phenograms and column four provides for the number of individuals included in the study from each species.

Table 1: List of *Aloe* species and individual used

	Name of species	Code	Number
1	*A. myriacantha* (Haw.) Schult. & Schult.f.	A. myr	2
2	*A. nuttii* Baker	A. nut	2
3	*A. leedalii* S. Carter	A. lee	2
4	*A. richardsiae* Reynolds	A. ric	2
5	*A. bullockii* Reynolds	A. bul	2
6	*A. bulbicaulis* Christian	A. bulb	3
7	*A. wollastonii* Rendle	A. wol	3
8	*A. kilifiensis* Christian	A. kil	2
9	*A. lateritia* Engl.	A. lat	2
10	*A. duckeri* Christian	A. duc	2
11	*A. mzimbana* Christian	A. muz	2
12	*A. congdonii* S. Carter	A. con	2
13	*A. chabaudii* Schönland	A. cha	2
14	*A. veseyi* Reynolds	A. ves	3
15	*A. bukobana* Reynolds	A. buk	2
16	*A. christanii* Reynolds	A. chr	3
17	*A. dorotheae* A. Berger	A. dor	2
18	*A. bussei* A. Berger	A. bus	2
19	*A. leptosiphon* A. Berger	A. lep	3
20	*A. massawana* Reynolds	A. mas	2
21	*A. mawii* Christian	A. maw	2
22	*A. bicomitum* L.C. Leach	A. bic	2
23	*A. macrosiphon* Baker	A. mac	2
24	*A. secundiflora* Engl.	A. sec	2
25	*A. leachii* Reynolds	A. lea	2
26	*A. brandhamii* S. Carter	A. bran	2
27	*A. confusa* Engl.	A. conf	2
28	*A. flexilifolia* Christian	A. fle	2
29	*A. boscawenii* Christian	A. bos	2
30	*A. rabaiensis* Rendle	A. rab	2
31	*A. ngongensis* Christain	A. ngo	2
32	*A. brachystachys* Baker	A. bra	2
33	*A. babatiensis* Christian & I. Verd.	A. bab	2
34	*A. fibrosa* Lavranos & L.E. Newton	A. fib	2
35	*A. morijensis* S. Carter & Brandham	A. mor	2
36	*A. volkensii* Engl.	A. vol	2
37	*A. ballyi* Reynolds	A. bal	2
38	*A. elata* S. Carter & Newton	A. ela	2
39	*A. deserti* A. Berger	A. des	2

Data Analysis

The code of the species name in the phenogram is made from the first three letters of the species name, where more than one species ended having the same code a fourth letter was added. Each data matrix was analysed by Unweighted Pair Group Method with Arithmetic mean (UPGMA)

clustering technique using PAST 2.16 software. The ten coefficient indices used are: Euclidean, Hamming, Gower, Chord, Jaccard's, Dice, Rho, Kulczynski, Correlation, and Simpson. The first four coefficient indices were distance coefficient whereas the last six were similarity coefficient. The coefficient used in the analysis cover both frequently and rarely used ones. Cophenetic values were recorded for each of the phenogram generated and used as a measure of the phenogram to matrix match.

RESULTS

Table 2 presents a summary of the results from the cluster analyses using the ten selected coefficient indices. A group was considered a species-specific cluster only when the two or three individuals from the same species used in the analysis clustered together first before they clustered with individuals from other species.

The Phenogram obtained by using Hamming coefficient recovered 38 of the 39 expected species-specific clusters (Fig. 1) and Gower and Rho coefficients produced 36 species-specific clusters each (Table 2). Thus, phenograms produced by Hamming and Gower distance coefficients and Rho similarity coefficient produced the highest number of species-specific clusters (36-38 out 39 expected). Phenograms produced using Chord, Correlation and Euclidean coefficients recovered the lowest number of species-specific clusters (8-9 out of the 39 expected) (Fig. 2). All others coefficients produced less than 25 species specific clusters. Other representative Figures are appended.

Table 2: Summary of the Results of Cluster Analysis Using Different Coefficient Indices

Coefficient indices	Coefficient index type	Cophenetic correlation value	Clusters recovered
Euclidean	Distance	0.7708	9
Hamming	Distance	0.7541	38
Jaccard’s	Similarity	0.7347	22
Dice	Similarity	0.7329	22
Kulczynski	Similarity	0.7186	23
Gower	Distance	0.6966	36
Rho	Similarity	0.6778	36
Chord	Distance	0.6168	8
Correlation	Similarity	0.5154	8
Simpson	Similarity	0.4762	13

Generally, many phenogram had comparable cophenetic values that were above 0.6 except the phenogram produced by Correlation coefficient that had its cophenetic value at 0.5154 and a phenogram produced by Simpson which had a cophenetic value of 0.4762.

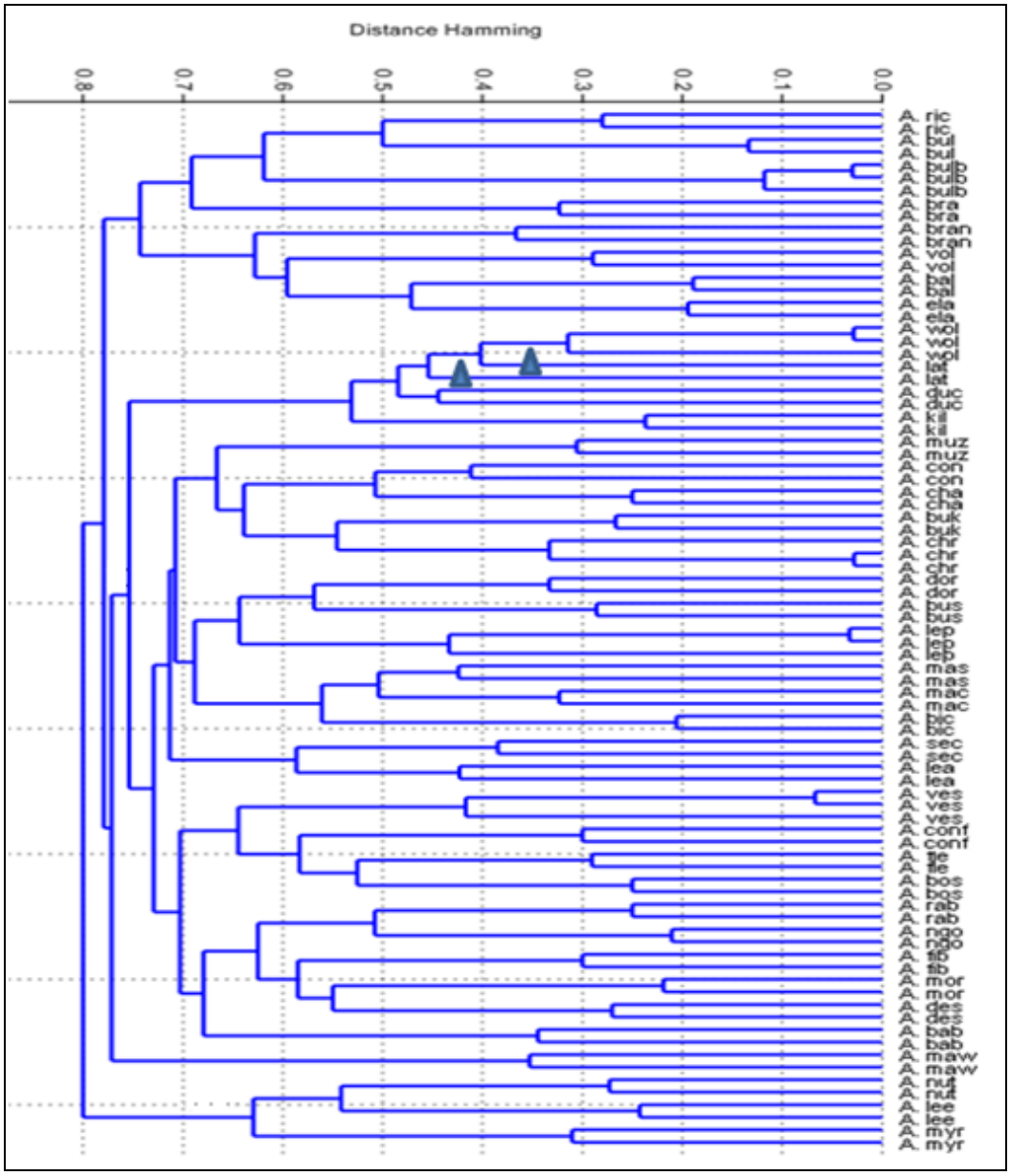

Figure 1: A phenogram produced from the cluster analysis using Hamming coefficient. Each cluster represents one species with exception of two individuals of *A. lateritia* denoted by a greyish triangular mark on branches which did not fall in a specific cluster.

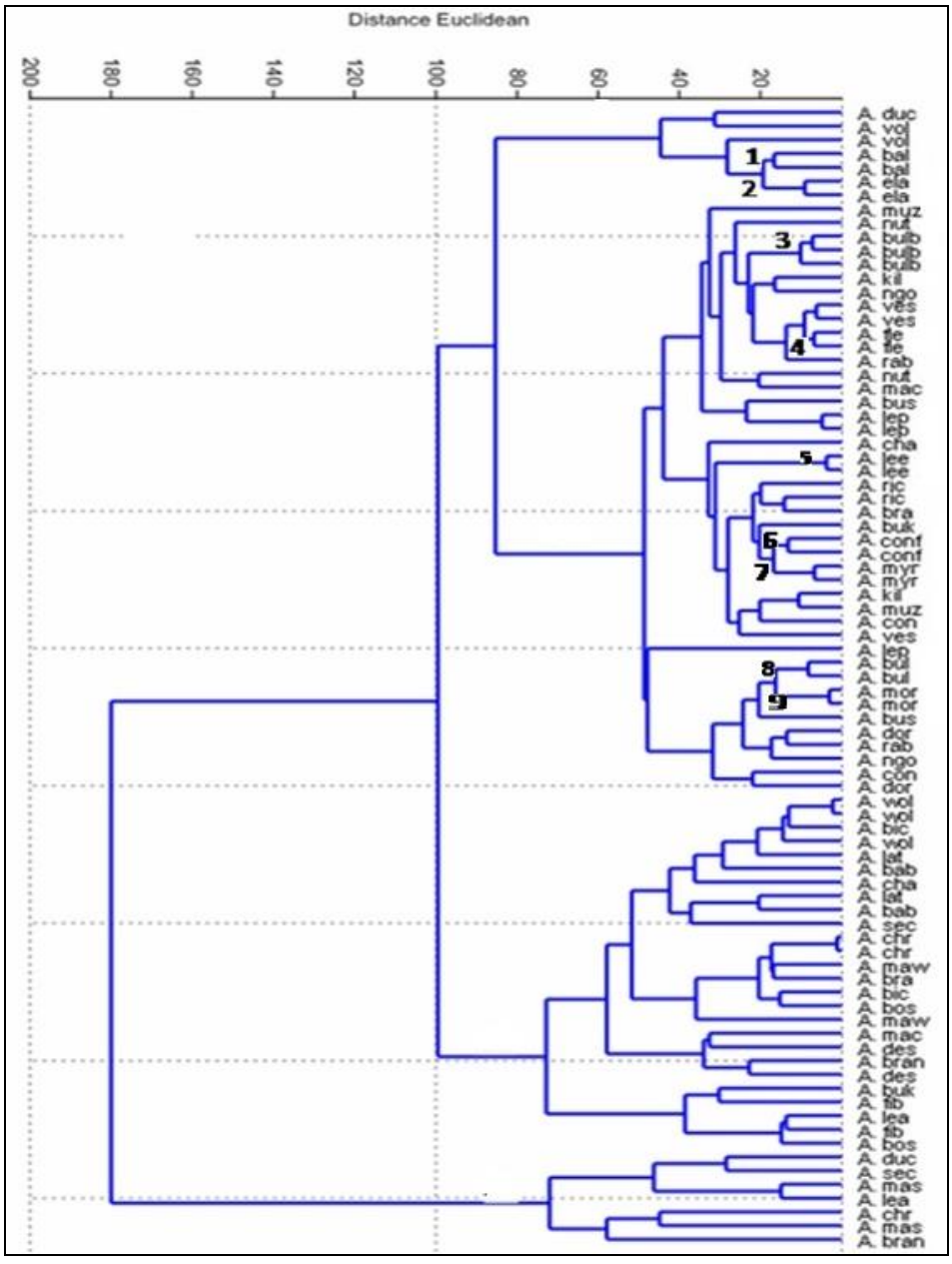

Figure 2: A phenogram produced from the cluster analysis using Euclidean coefficient. Arabic numbers 1-9 denote the only species-specific clusters recovered.

DISCUSSION

Cluster analysis of morphological characteristics of *Aloe* species by different coefficient indices provided phenograms with different clustering patterns and cophenetic correlation values. The assumption in this case was that, since 39 species were analysed, 39 species specific clusters were expected in each analyses. Out of the expected 39 clusters, different numbers of species-specific clusters ranging from 8 to 38 were produced (Table 2). This is an indication that coefficient indices influenced the clustering patterns thus the resulting classification. In phenetic classification, clusters are equated to species at species level taxonomy i.e. the number of species-specific clusters formed are thus equal to the number of species to be recognized in the group. It can therefore be concluded that coefficients influences classification. The analysis was carried out without the removal of outliers. In the study on impact of similarity measure on web-age clustering by Strehl et al. (2000), coefficient used influenced the clustering too. Several other studies have reported the reliance of clustering on the coefficient index (Duarte et al. 1999, Murguia and Villasenor 2003, Meyer et al. 2004, Naseem et al. 2010, Manoko 2018).

Clustering pattern exhibited by Hamming and Gower coefficient indices compares to results obtained by Manoko (2018) where ten species previously delimited by AFLP markers were recovered. However, the pattern shown by Rho coefficient in the present study cannot be compared. Reason for this observation can only be speculated but the power of Gower coefficient has been demonstrated in several other studies. In a study on comparison of multivariate statistical algorithms to cluster heirloom accessions, Gonclaves et al. (2008) showed that Gower coefficient was more effective than other coefficients used too. The good performance of Gower coefficient index is probably attributed to its wider range of application domains i.e. it can be applied for binary, multistate and quantitative characters (Gower 1971).

Based on the present study it can thus be said that successful delimitation of species in complex groups like *Aloe* using phenetics approach depends on the selection of coefficient. Euclidean coefficient though often used and despite recording the highest cophenetic correlation value recovered only 9 out of 39 expected species. Obviously this applies for Chord and Correlation which if used would recover only 8 species of the 39 expected species but splitting individuals of same species to other unrelated species. Performance of Jaccards and Dice in the present study was expected because according to Manoko (2018) the two coefficients produced species specific cluster only with binary coded data. In the present study data were coded using conventional method only.

Species are delimited for different purposes but correct identification of any species for use being it in pest control, medicinal purpose, epidemiology or biodiversity conservation will only save mankind if classification is predictable. The later entirely depends on proper delimitation of species.

Results of the current study demonstrate comparable cophenetic correlation value which range from 0.61 to 0.77 with exception of Correlation and Simpson coefficient which had 0.52 and 0.48 cophenetic correlation values respectively. Cophenetic correlation coefficient is a measure of degree of fit of a classification to a data set or the efficiency of various clustering techniques (Farris 1969). In this

case majority of the phenogram match well with the data set with the exception of Correlation, Euclidean and Simpson coefficients. Thus where cophenetic correlation value matches the produced phenogram the difference in phenograms in producing species–specific clusters therefore can only be explained by the factor that was changing across the analysis that is the coefficient differences. For Correlation and Simpson coefficients the two recorded the lowest cophenetic correlation values indicating probably that the pattern of clustering produced actually did not match the data set. Although this may be taken to signify the importance of using cophenetic correlation value as a basis of selecting best phenogram based on the present this does not apply. Euclidean coefficient produced a phenogram with the highest coefficient but out of 39 expected species recovered only 9 species. This conclusion is also shared by Holgersson (1978) who recommended use of cophenetic correlation value as a clustering criterion with care because could be misleading. In fact, it has been argued that cophenetic correlation value be used only when different selecting between phenograms produced using the same coefficient by different clustering methods (Goncalves et al. 2008).

CONCLUSION

From the results of the current study, Hamming coefficient recovered 38 out of 39 species-specific clusters where Gower and Rho coefficients both recovered 36 out of 39 species specific clusters. Recovering 38 out of 39 species whereas others coefficients less than 25 and others less than 10 species indicates choice of a coefficient is critical in species delimitation when using phenetic approach and UPGMA method to classify plant species. Though often people have chosen to ignore and blamed morphological characteristics in favour of molecular markers, based on this study it may be concluded that when using phenetic approach and UPGMA to delimit species it is the choice of the coefficient to use which matters and not the type of data. In the present study it was possible to delimit *Aloe* species from Tanzania described in the Flora of tropical East Africa using morphological data.

REFFERENCES

Angiosperm Phylogeny Group III 2009 An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants APG III. *Bot. J. Linn. Soc.* **161**: 105–21.

Carter S 1994 *Aloaceae*, In Polhill, RM (ed) *Flora of Tropical East Africa*, AA Balkema, Rotterdam, 1-60.

Chase MW, De Bruijn, AY, Cox A, Reeves G, Rudall PJ, Johnson MAT and Eguiarte IE 2000 Phylogenetics of Asphodelaceae (Asparagales): an analysis of plastid *rbcLtrnL*-F DNA sequences. *Annals Bot.* **86**:935-951.

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

Farris JS 1969 On the cophenetic correlation coefficient. *Syst. Biol.* **18**(**3**): 279-285.

Fikre D 2007 *Taxonomic and Demographic studies on three species complexes within the Genus Aloe* L. (Aloaceae) in Ethiopia. PhD Thesis, Addis Ababa University.

Furin A and Wunder J 2004 Analysis of eggplant (*Solanum melongena*)-related germplasm: morphological and AFLP data contribute to phylogenetic interpretations and germplasm utilization. *Theor.. Appl. Genet.* **108(2):**197-208.

Goncalves LSA, Rodrigues R, Amaral Junior AT, Karasawa M and Sudre C P 2008 Comparison of multivariate statistical algorithms to cluster tomato

heiloom accessions. *Genet. Molec. Res.* **7(4):**1289-1297.

Gower JC and Legendre P 1986 Metric and Euclidean properties and dissimilarity coefficients. *J. Classific.* **3**:5-45.

Gower JC 1971 A general coefficient of similarity and some of its properties. *Biometrics*, **27**:857-871.

Grace OM 2009 Contributions to the systematics and biocultural value of *Aloe* L. (Asphodelaceae). a PhD Dissertation, Pretoria University.

Grace OM, Klopper RR, Smith GF, Crouch NR, Figueiredo E, Rønsted, N., van Wyk, AE 2013 A revised generic classification for Aloe (Xanthorrhoeaceae subfam. Asphodeloideae). *Phytotaxa* **76**:7–14.

Hillis DM and Wiens JJ 2000 Molecules versus morphology in systematics. in Wiens, (ed.) *Phylogenetic analysis of morphological data.* Smithsonian Institution Press, Washington, D.C. 1-19.

Holgersson M 1978 The limited value of cophenetic correlation as a clustering criterion. *Patt. Recogn..*, **10**(4):287-295.

Hubalek Z 1982 Coefficients of association and similarity, based on binary (presence-absence) data: an evolution. *Biol. Rev.* **57(54**)669–689.

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

Judd WS, Campbell CS, Kellogg EA, Stevens PF and Donoghue MJ 2002 *Plants Systematic*, a phylogenetic approach, 2nd Ed, Sunderl and Massachusetts U.S.A.

Manoko MLK 2007 A systemstic study of African *Solanum* L. Section *Solunum* (*Solanaceae*), A Phd thesis, Radboud University Nijmegen.

Manoko M 2018 The power of coefficients and methods of coding in delimiting species using phenetic approach: the case of African *Solanum* section *Solanum* sensu Edmonds. *Tanz. J. Sci.* **44(1**)37-41.

McCoy T and Lavranos J 2007 Four interesting new species of Tanzanian aloes**.** *Aloe*,**44(2):** 50-53.

Meyer ADS, Garcia AAF, Souza APD and Souza JCLD 2004 Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (Zea mays L). *Genet. Molec. Biol.* **27**(**1**):83-91.

Mishler BD 2000 Deep phylogenetic relationships among "plants" and their implications for classification. *Taxon* **49**: 661–683.

Murguia M and Villaseno JL 2003 Estimating the effect of the similarity coefficient and the cluster algorithm on biogeographic classification. *Annals Bot.* **40**:415-421.

Naseem R, Maqbool O and Muhammad S 2010 An Improved Similarity Measure for Binary Features in Software Clustering. *IEEE Explore.*

Newton LE and Rowley GD in: Eggli U (2001). Cites Aloe and Pachypodium checklist. The trustees of Royal Botanical Garden Kew, Sukklenten-Sammlung Zurich.

Strehl A, Ghosh J and Mooney R 2000 Impact of Similarity Measures on Web-page clustering, Workshop of Artificial for Web search, University of Texas U.S.A.

Sun F and Downie SR 2010 Phylogenetic analyses of morphological and molecular data reveal major clades within the perennial endemic western North American Apiaceae subfamily Apioideae, *J. Torr. Bot. Soc.* **137(2-3):**133-156.

Wabuyele E 2006 Studies on East African *Aloe* species: Aspects of Taxonomy,

conservation and Ethnobotany. PhD Dissertation, University of Oslo.

Wiens JJ 2004 The Role of Morphological Data in Phylogeny Reconstruction, *Syst. Biol*.**53 (4)**:653–661.

http://www.computer.org/csdl/proceedings/csmr/2011"Improved Similarity Measures for Software Clustering," csmr, 2011 15th European Conference on Software Maintenance and Reengineering, retrieved on Thursday, 6th