Justus Liebig University Gießen
Institute of Agronomy and Plant Breeding II
Biometry and Population Genetics
Prof. Dr. Matthias Frisch

Comparison of different proximity measures and classification methods for binary data

Dissertation
Submitted for the degree of Doctor of Agricultural Science
Faculty of Agricultural Sciences, Nutritional Sciences and Environmental Management
Justus Liebig University Gießen

Submitted by
Taiwo Adetola Ojurongbe
From Nigeria

Gießen, 2012

Dean: Prof. Dr. Dr.-Ing. Peter Kaempfer

Supervisor: Prof. Dr. Matthias Frisch

Second Supervisor: Prof. Dr. Dr. Wolfgang Friedt

Date of oral examination: 20 February 2012

1 Introduction

1.1 Background

Statistical 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. FA is a collection of methods used to examine how underlying concepts influence the responses on a number of measured variables. Basically there are two types of FA termed explanatory and confirmatory factor analysis. Explanatory and confirmatory factor analyses are based on the common factor model which proposes that each observed response is influenced partially by underlying common factors and partially by underlying unique factors. FA is performed by examining the pattern of correlations (or covariances) between the observed measures and it helps to reduce a vast number of variables to a meaningful, interpretable and manageable set of factors (DeCoster, 1998; Hatcher and Stepanski, 1994). PCA is a way of identifying patterns in data and expressing the data with the purpose of highlighting their similarities and differences. It is a common technique in finding patterns in data of high dimension (Smith, 2002). DA on the other hand is used for classifying a set of observations into predefined classes.

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. CA divides data into groups that are meaningful, useful or both. For meaningful groups, the natural structure of the data should be revealed in the groups (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. CA has been applied to many practical problems depending on whether the purpose is for understanding or utility. For example in biology, it is used to analyze large amount of genetic data and also in the study of the earth's climate, where it is used to find patterns in atmospheric pressure of polar regions and areas of the ocean that have a significant impact on land climate. Similarly, in psychology and medicine, CA has been used to identify different types of depression and in the detection of spatial and temporal patterns in the distribution of a disease (Tan et al., 2006).

In population genetics and plant breeding, quantifying the degree of dissimilarity among genera, species, subspecies, populations and elite breeding materials is of primary concern (Reif et al., 2005). Molecular markers have been widely used for this purpose to characterize 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. A prerequisite of CA for many methods is the construction of similarity/dissimilarity coefficients between the individuals or objects being considered. Several studies have been published in the past years using molecular markers to study genetic divergence and phylogenetic relations between species (Dias et al., 2004).

As suggested by Reif et al. (2005), 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. 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. It has also been observed that the choice of the coefficients used by many authors is not justified and this may cause problems, jeopardizing the nature of the analysis (Duarte et al., 1999; Jackson et al., 1989). Therefore the knowledge of the genetical and mathematical properties as well as the application of these coefficients in different situations is important. This study will therefore attempt to investigate and justify the effect of the use of different similarity coefficients on binary data.

1.2 Literature Review

In this section, a description of CA, its strategies, different linkage methods and CA prerequisites are discussed. Consensus trees and methods are also discussed, giving a basis for the comparison that follows later in the study.

1.2.1 Cluster Analysis

CA is a technique used to classify objects or individuals into mutually exclusive and collectively exhaustive groups with high homogeneity within clusters and low homogeneity between clusters. It is used to classify observations into a finite and, ideally, small number of groups based upon two or more variables. In some cases there are hypotheses regarding the number and make up of such groups, but more often there is little or no prior information concerning which individuals will be grouped together, making CA an exploratory analysis. In contrast to DA, CA operates on data sets for which pre-specified well-defined groups do not exist but are suspected and could be applied to a similarity/dissimilarity matrix. There are many measures used in calculating these matrices which include the Dice, Jaccard, Simple matching and so on. There are a number of clustering algorithms available, all having as their primary purpose the measurement of mathematical distance between individual observations, and groups of observations (Finch, 2005). CA techniques have been used to provide solutions to a large variety of research problems which includes archeology where researchers have made efforts to establish taxonomies of stone tools, funeral objects etc (Hartigan, 1975). Also in the field of medicine, clustering diseases, cures for diseases and symptoms can lead to very useful taxonomies (Hartigan, 1975; Hill and Lewicki, 2008). In plant and animal ecology, CA is useful in describing spatial and temporal comparisons of communities of organisms in heterogeneous environments (Jongman et al., 1995). It is also used in plant systematic to generate artificial phylogenies or clusters of organisms at the species, genus or higher level that have a number of common attributes.

1.2.2 Similarity/Dissimilarity Coefficients

The calculation of similarity/dissimilarity coefficient is a prerequisite for CA and different similarity coefficients are used based on specific types of data. A similarity coefficient (S) can be converted into dissimilarity (D) by taking the complement of the similarity coefficient i.e. D = 1-S. The choice of similarity coefficient to be used in CA has a strong impact on the results from clustering (Duarte et al., 1999; Jackson et al., 1989). The choice of an appropriate coefficient of similarity is a very important and decisive point to evaluate clustering, true genetic similarity between individuals, analysing

Table 1: Similarity coefficients for clustering binary variables (Johnson and Wichern, 1988).

Name	Coefficient	Rationale	Range
Anderberg	a/(a+2(b+c))	No 0-0 matches in numerator or denominator. Double weight for 1-1 matches	**[0, 1]**
Dice	2a/(2a+b+c)	No 0-0 matches in numerator or denominator. Double weight for unpaired matches	**[0, 1]**
Hamann	((a+d)-(b+c)) /(a+b+c+d)	Mismatches subtracted from matches	**[0, 1]**
Jaccard	a/(a+b+c)	Zero weight for 0-0 matches in numerator	**[0, 1]**
Kulcynski	a/(b+c)	Ratio of matches to mismatches with 0-0 matches excluded	**[0, ∞]**
Ochiai	$a/((a+b)(c+d))^{0.5}$	Ratio of 1-1 matches to Square root of product of sum of matches and mismatches	**[0, 1]**
Roger and Tanimoto	(a+d)/(a+d+2(b+c))	Double weight for unmatched pairs	**[0, 1]**
Russel & Rao	a/(a+b+c+d)	Zero weight for 0-0 matches in numerator	**[0, 1]**
Simple Matching	(a+d)(a+b+c+d)	Equal weight for 1-1 & 0-0 matches	**[0, 1]**
Sokal & Sneath 1	2(a+d)/(2(a+d)+b+c)	Double weight for 1-1 & 0-0 matches	**[0, 1]**

within populations and studying relationships between populations, because different similarity coefficients may yield conflicting results (Kosman and Leonard, 2005). These coefficients need to be better understood, so that the most appropriate ones are used in each specific situation. In a situation where there are two isolates observed for the presence (1) or absence (0) of different attributes, the similarity between the two objects/individuals can be calculated using the formulas in Table 1, derived from a two by two contingency table of one and zero, where "a" represents a 1:1 occurrence, "b" stands for 1:0 occurrence, "c" for a 0:1 occurrence and "d" for a 0:0 occurrence. The choice of the

similarity coefficient to be used is based on either the importance of a 1:1 occurrence or a 0:0 occurrence of the attributes in the two isolates being compared. Most often, authors do not justify the choice of the coefficients used, thus showing the necessity of studies on this subject. The list of similarity coefficients for binary data is given in Table 1.

1.2.3 Clustering Strategies

Clustering strategies may be grouped into the following categories: hierarchical or non-hierarchical, divisive or agglomerative and polythetic or monothetic (Lambert et al., 1973; Orloci, 1978; Sneath and Sokal, 1973). Non-hierarchical clustering techniques partition samples into a number of clusters but specify no structure about the relationship between the clusters. Hierarchical clustering techniques define relationships among the clusters and they show, for example that cluster A is more similar to B than it is to C. A single hierarchical analysis allows one to choose the final number of groups by selecting an appropriate level in the hierarchy, and this choice can be made based on the structure of the data set. The non-hierarchical technique is recommended if the only requirement in a clustering application is that a given number of clusters be formed, but are not related to one another (Gaugh and Whittaker, 1981).

Divisive hierarchical clustering strategies begin with all samples in a single cluster and divide them, usually into two clusters; these clusters are then further subdivided until each cluster contains no more than a specified number of samples. Agglomerative clustering strategies however, begin with the individual samples, and fuse these into successively larger clusters until finally a single cluster containing all samples is formed. This choice of strategy has important implications for the properties of clustering techniques, affecting which aspects of the data are emphasized and what criteria are optimized (Orloci, 1978; Sneath and Sokal, 1973).

Monothetic techniques partition data on the basis of presence or absence of a single character. An important example of monothetic technique in earlier times was association-analysis but it had an undesirably high rate of miss-classification (Hill et al., 1975; Orloci, 1978; Williams and Lambert, 1959). However, polythetic agglomerative hierarchical clustering (PAHC) techniques use all the information contained in all the variables. First, each entity is assigned as an individual cluster.

Subsequently, PAHC agglomerates these clusters in a hierarchy of larger and larger clusters until finally a single cluster contains all entities. This family of technique is also known as Sequential Agglomerative Hierarchical and Non-overlapping (SAHN) (McGarigal et al., 2000) and was used for the clustering analyses carried out in this study.

CA relies on similarity measures and CA methods and any combination of these two criteria are possible. However, the choice depends on the situation at hand or questions to be answered by the researcher. Generally, CA finishes with the obtention of the dendrogram (tree) and its respective analysis and interpretation (Restrepo and Villaveces, 2005).

1.2.4 Cluster Analysis Methods or Linkage Rules

Once several objects have been linked together as in CA, a linkage or amalgamation rule is needed to determine when two clusters are sufficiently similar to be linked together. There are various possibilities: for example, two clusters could be linked together when any two objects in the two clusters are closer together than the respective linkage distance. The different clustering methods as described in Hill and Lewicki, (2008) are as follows:

(i). Single linkage (nearest neighbour): In this method, the distance between two clusters is determined by the distance of the two closest objects (nearest neighbours) in the different clusters. This rule will, in a sense, “string” objects together to form clusters, and the resulting clusters tend to represent long "chains" as shown in Figure 1A.

(ii). Complete linkage (farthest neighbour): In this method, the distances between clusters are determined by the greatest distance between any two objects in the different clusters (i.e., by the "farthest neighbours"), Figure 1B. This method usually performs quite well in cases when the objects actually form naturally distinct cluster. It is not appropriate for clusters that tend to be somehow elongated or of a "chain" type nature.

(iii). Unweighted Pair-Group Mean Arithmetic method (UPGMA): This method is also very efficient when the objects form natural distinct clusters, however, it performs equally well with elongated,

"chain" type clusters. In UPGMA, the distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters, Figure 1C.

(iv). Weighted Pair-Group Mean Arithmetic method (WPGMA): This method is identical to the UPGMA method, except that in the computations, the size of the respective clusters (i.e., the number of objects contained in them) is used as a weight. Therefore, this method is recommended to be used instead of the UPGMA method when the cluster sizes are suspected to be very uneven.

(v). Unweighted Pair-Group Method using Centroid Average (UPGMC): The centroid of a cluster is the average point in the multidimensional space defined by the dimensions. It is the "center of gravity", in a way for the respective cluster. In this method, the distance between two clusters is determined as the difference between centroids.

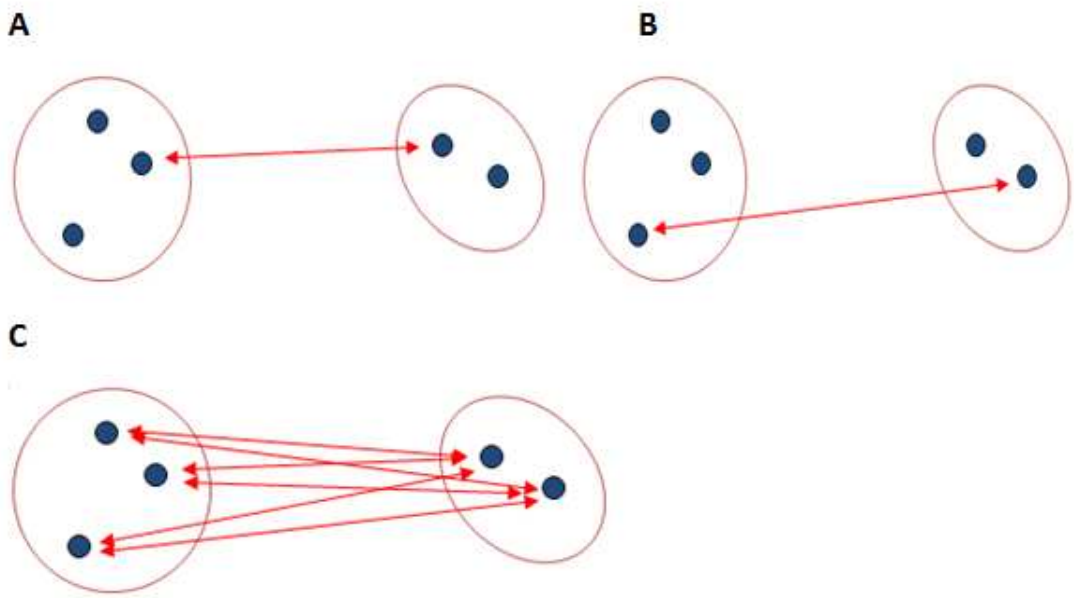


Figure 1: Distances between clusters.

A - An example of minimum distance between clusters. B - An example of maximum distance between clusters. C - An example of average distance between clusters.

(vi). Weighted Pair-Group Method using Centroid Average (WPGMC): When there are (or we suspect there is likely to be) considerable differences in cluster sizes, then, the WPGMC method is preferable to the UPGMC. WPGMC method is identical to the UPGMC, except that weighting is introduced into the computations to take into consideration differences in cluster sizes (i.e., the number of objects contained in them) (Hill and Lewicki, 2008).

The Neighbour-Joining method (NJ) is related to the clustering method but does not require the data to be ultrametric. It is a distance based method that requires a distance matrix and uses the star decomposition method. It is especially suited for data sets comprising lineages with largely varying rates of evolution. NJ keeps track of nodes on a tree rather than objects or clusters of objects. To use the NJ method, the initial assumption is that there is just one internal node from which branches leading to all the individuals radiate in a star-like pattern. The separation between each pair of nodes is adjusted on the basis of their average divergence from all other nodes. The principle is to find pairs of individuals (i.e. neighbors) that minimize the total branch length at each stage of clustering of the individuals starting with the star-like tree (Saitou and Nei, 1987).

NJ is a special case of the star decomposition method (Figure 2), in which the raw data are provided as a distance matrix and the initial tree is a star tree. A modified distance matrix is then constructed where the separation between each pair of nodes is adjusted on the basis of their average divergence from all other nodes. The tree is constructed by linking the least-distant pair of nodes in the newly constructed matrix. When two nodes are linked, their common ancestral node is added to the tree and the terminal nodes with their respective branches are removed from the tree. By so doing, the newly added common ancestor is converted into a terminal node on a tree of reduced size. At each stage in the process two terminal nodes are replaced by one new node. The process is complete when two nodes remain separated by a single branch (Saitou and Nei, 1987).

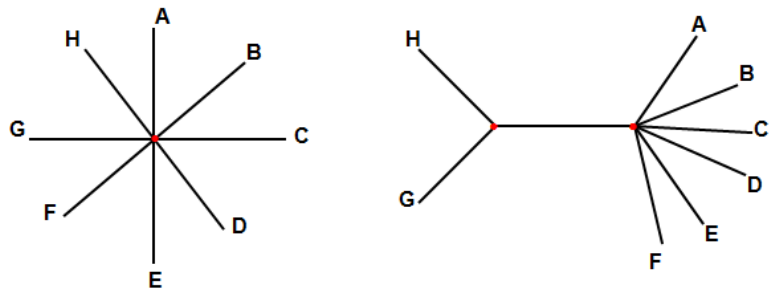


Figure 2: An example of the star decomposition method for NJ.

In the NJ method, there is the possibility of assigning a negative length to the branches in the tree known as negative branch lengths. This is because the NJ algorithm seeks to represent the data in the form of an additive tree. This makes the interpretation of branch lengths as an estimated number of substitutions to be very difficult. A way out of this difficulty is to set the branch length to zero and then transfer the difference to the adjacent branch length so that the total distance between an adjacent pair of terminal nodes remains unaffected. This does not in any way affect the overall topology of the tree (Kuhner and Felsenstein, 1994).

Advantages of the NJ method

1. It is fast and thus suitable for large data sets and for bootstrap analysis
2. It permits lineages with largely different branch lengths
3. It permits correction for multiple substitutions

Disadvantages of the NJ method

1. Sequence information is reduced
2. It gives only one possible tree
3. It is strongly dependent on the model of evolution used

1.2.5 Consensus Trees and Methods

A consensus tree is a tree that represents the consensus topology (subset of relationships) of two or more trees being compared. A consensus index is a numerical value that indicates the degree to which the consensus tree is resolved, i.e. fully bifurcating. If the original trees are fully resolved then the

degree to which the consensus tree is fully resolved is a measure of the similarity of the original trees (Rohlf, 1992). Consensus tree methods aim at a tree that represents the joint information or consensus of two or more trees; consensus-index methods furnish a numerical measure of the agreement among trees (Shao and Sokal, 1986).

Some of the methods for constructing consensus trees include strict consensus, majority rule consensus, stinebrickner consensus and loose consensus.

(i). Strict consensus: The strict consensus method is reported to be the simplest of the all the consensus methods. Given a collection of unrooted trees, the strict consensus tree contains exactly those splits common to all the trees in the collection. When the collection consists of rooted trees the strict consensus tree contains those clusters common to all the input trees.

For example, let T be the collection of rooted trees {((a, (b, c)), d), (((a, b), c), d). The clusters {a, b, c, d and {a, b, c} appear in both trees, so the strict consensus tree is ((a, b, c), d). Strict consensus trees have a natural generalisation to weighted trees. The strict consensus is computed as if for an unweighted tree, and then the minimum weight of each of the corresponding splits (clusters) is assigned to each branch in each of the input trees (Bryant, 2003).

(ii). Majority rule consensus: The majority rule tree contains exactly those clusters or splits that appear in more than half of the input trees. Thus every cluster (split) of the strict consensus tree will also be a cluster (split) of the majority rule tree. The majority rule tree refines the strict consensus tree. Example, let T be the collection of three rooted trees {((a, (b, c)), d), (((a, b), c), d), (((a, b), d), c)}. The clusters {a, b}, {a, b, c} and {a, b, c, d} appear in two out of three trees, so the majority rule tree is (((a, b), c), d). In another example, let T1 and T2 be unrooted trees ((a, b, c), (d, e, f)) and ((a, d), (b, e), (c, f)). If T contains three copies of T1 and two copies of T2 then the majority rule tree of T equals T1 (Bryant, 2003). When the number of input trees (m) = 2, then majority rule consensus and strict consensus are the same (Rohlf, 2002).

(iii). Stinebrickner consensus: A Stinebrickner consensus tree is more complex. For each cluster of size p (“cardinality p”), containing a member i, the intersection and the union are taken through all clusters of the m trees that have p or fewer members and also containing member i. The cardinality of

the intersection is divided by the cardinality of the union. If this value is greater than the parameter S_c, the index of stringency, then that intersection is included as a cluster in the consensus tree. The S_c parameter can be varied from 0 to 1. It allows a more flexible approach to the construction of consensus trees than does the strict consensus method. If $S_c = 1$ then this yields strict consensus trees. As the S_c value decreases to 0, additional clusters will be included in the consensus tree.

(iv). Loose consensus tree: The loose consensus tree was originally called the combinable component tree or semi-strict consensus tree (Bremer, 1990; Swofford, 1991). For a collection of rooted trees T, the loose consensus tree contains exactly those clusters that are compatible with every tree in T. Similarly, the loose consensus of a collection of unrooted trees T contains exactly those splits that are compatible with every unrooted tree in T. The loose consensus tree also refines the strict consensus tree (Bryant, 2003). Example: Let T be the collection of rooted trees {((a, b), (c, d)), ((a, b, c), d)}. The cluster {a, b} is compatible with both trees; however the cluster {c, d} is not compatible with the cluster {a, b, c}. Hence the loose consensus tree for T is ((a, b), c, d). The strict consensus tree for this collection equals (a, b, c, d).

1.3 Aims and Objectives

This work is aimed at studying the influence of the choice of similarity coefficient and clustering methods in cluster analysis with respect to different populations. The specific objectives of this study are to:

1. Investigate the impact of the underlying (chosen) similarity (dissimilarity) measure and the CA algorithms on the resulting classifications.
2. Find a good measure of comparing topology and to determine how consistent the topology of the constructed trees is.
3. Compare the quality of the classification with respect to CA.
4. Compare using multivariate techniques three similarity coefficients and their effect on clustering on yam pathogen, powdery mildew and plantain production constraints.

2 Materials and Methods

In chapter one, a description was given of CA and the different methods as well as cluster strategies. Different similarity measures used for binary data before carrying out CA were also discussed. In this chapter, details of the simulated data are outlined and description of the experimental data used in this study is presented. Data collection methods as well as data analyses procedures are also discussed. In all, a total of three experimental data with different scenarios, based on different plants and simulated binary data were used.

2.1 Simulation for Binary Data

Binary data for presence (1) or absence (0) of some characteristics (for example, degree of infection) describing different isolates with varying properties were generated using R software. Samples with different number of rows (r) and columns (c) per sample were generated in order to see whether the dimension of the binary data generated would have an effect on the resulting classification. For example virulence or marker data could sometimes have very long number of differentials or bands. The effect of an increase or decrease in these parameters on the resulting classification was observed. Two known groups (A and B) identified by the first half of the number of rows and the second half respectively were created per simulation with each group divided into three sections where C_{left}, C_{middle} and C_{right} represent the left, middle and right columns of the data respectively. The first ten columns (C_{left}) and the last ten columns C_{right}, referred to as the two "outer sections" contained the determining characteristics of each group, that is, a distinctly different (0, 1) composition. The middle section on the other hand was designed such that a "1" occurred with probability p = 0.7 and a "0" with q = 0.3, resulting in a 49 percent chance of having 1:1 occurrence between two objects, 21 percent chance each of having 1:0 and 0:1 and a 9 percent chance of having 0:0 occurrence. The strength of the (A, B) - grouping was relaxed by elongating C_{middle}. The two edges of each group ($C_{left,}$ C_{right}) were the determining characteristics of the group while C_{middle} was random.

The two groups (A and B) were created such that in Group A, C_{left} had a 100 percent chance of having 1:1 occurrence and the C_{middle} was as discussed above, while C_{right} was divided into equal halves of 5 columns each. The first 5 columns had a 100 percent chance of having a 1:1 occurrence and the last 5 columns had a 100 percent chance of having a 0:0 occurrence. In group B, C_{left} was also divided into

equal halves of 5 columns each: the first 5 columns had a 100 percent chance of having a 1:1 occurrence and the next 5 columns a 100 percent chance of having 0:0 occurrences. C_{middle} remained as it was and C_{right} had a 100 percent chance of having a 1:1 occurrence and this is shown in the example in Table 2. Suppose there are six isolates with 19 columns, the first six columns are designated as C_{left} and the last six columns as C_{right} and the seven columns in the middle as C_{middle}. Therefore, data with two groupings and the C_{middle} varying were simulated starting with r = 30 for rows and c = 20 for columns with increment of 10 up to 100 columns and 130 to 200 columns with increment of 20, 20, and 30 respectively. The number of rows was later increased to 40 for each round of simulation. That is, all pairs with r = 30, 40 and c_{middle} = 20, 30, 40, 50, 60, 70, 80, 90, 100, 130, 150, 170, 200, giving altogether 2 x 13 = 26 (r, c)-pairs. 1000 samples per (r, c)-pair were randomly generated with R software.

Table 2: An example of the simulated data showing the 3 sections: C_{left}, C_{middle} and C_{right}.

1 1 1 1 1 1	1 1 1 1 1 0 1	1 1 1 0 0 0
1 1 1 1 1 1	1 0 1 1 0 1 1	1 1 1 0 0 0
1 1 1 1 1 1	1 1 1 1 1 1 1	1 1 1 0 0 0
1 1 1 0 0 0	1 1 0 1 1 1 0	1 1 1 1 1 1
1 1 1 0 0 0	1 0 0 1 1 1 1	1 1 1 1 1 1
1 1 1 0 0 0	1 1 1 1 0 0 1	1 1 1 1 1 1
C_{left}	C_{middle}	C_{right}

2.2 Experimental Data from Field Trials

A description of the 3 different experimental data, namely:

1. Plantain data set from Nigeria
2. Powdery mildew data set from Germany and
3. Yam anthracnose disease amplified fragment length polymorphism (AFLP) marker data from Nigeria.

2.2.1 Background Information on Plantain

Plantain (*Musa* spp., AAB-group) is an important staple food in the humid forest zones of Western and Central Africa ((Flinn and Hoyoux, 1976; Guillemot, 1976; Melin and Djomo, 1972; Naku, 1983; Wilson, 1987; Wilson, 1983). About one-third of the population in the region derives more than 25% of their carbohydrates from this crop (Wilson, 1987). Plantain production in Western and Central Africa accounts for about 70% of world production and Nigeria is considered the largest producer in Western Africa (Lescot, 1998) and in the world (Akinyemi et al., 2010; FAO, 2006). In Nigeria, plantain is mainly cultivated on small-scale farms or in backyard gardens (Swennen and Vuylsteke, 1988). Apart from its importance as food, plantain earns cash for the small scale holders who sell their products in the rapidly growing urban areas (Speijer et al., 2001).

A reduction in the production of plantain in Nigeria and other Western and Central African countries is caused by several abiotic and biotic constraints (Fongeyn, 1976; Wilson, 1983). Major constraints to its production are declining soil fertility and acid soils as well as leaf diseases, mainly Black Sigatoka caused by the fungus *Mycosphaerella fijiensis*, the banana weevil Cosmopolites *sordidus* and plantain parasitic nematodes (Akinyemi et al., 2009; Mobambo et al., 1993; Schill et al., 1996; Wilson, 1987)

2.2.2 Data Collection Method and Analysis for Plantain Data Set

The plantain data set consist of Dichotomized Production Constraints (DPC). The constraints were categorized into two groups based on Plant Growth and Disease Evaluation as well as Root Health Assessment. A survey was carried out by the International Institute of Tropical Agriculture (IITA) between 1994 and 1995 in the plantain growing areas of Southern Nigeria and was reported by Speijer et al. (2001) as follows. Data were collected on plantain root health assessment as well as plant growth

and disease evaluation from nine states in Nigeria, which were further divided into three regions Western, Mid-Western and Eastern parts of the country (Table 3). The variables include % virus infection, % number of stands harvested, % of stands with toppled plant, % of stands with snapped plants, % of plants that are normal, average plant height in the plot, average circumference of the plant at the base, average number of suckers produced per plant in the plot, Cordana leaf disease index, speckle disease index, average percentage of banana leaf streak virus damage on leaves per plot, average number of leaves per plant in the plot, youngest leaf with spot (which is an index of Black Sigatoka leaf disease), yellow leaf streak (an index of yellow Sigatoka leaf infection), number of dead roots per plant in the plot, number of roots that appear healthy, root knox index, index of health of feeder roots per plant, root necrosis index on the main/primary roots of the main plant, root necrosis index on the main/primary roots of a sucker removed from the plant, weevil damage index, and height of sucker detached from the main plant. These variables were quantitative variables that were converted to binary data.

The areas for the survey were stratified using Geographic Information Systems (GIS) techniques on the basis of soil fertility. The GIS data file was similar to the construction of the Ugandan GIS data file described by Jagtap (1993). The GIS data file gave 220 possible sample cells, each approximately 300km^2. Stratified for the number of cells with a specific combination of identifiers, a total of 80 cells were chosen on the basis of presence of suitable farms and accessibility. For nematode root damage sampling, after possible sites within these cells were explored on a first visit, Speijer et al. (2001) retained a total of 73 survey sites on a second visit. However, for our study, a total of 70 sites were used due to missing values and for consistency. In each survey site, two farms located within 100 to 175 m of one another were selected. On each farm, ten recently flowered plants were chosen for sampling and data collection. A recently flowered plant has either just produced the inflorescence at the leaf axis or is still in the fruit filling stage.

In each farm, the number of recently flowered plants that had toppled, snapped or broken within one month prior to the site visit was recorded. A plant was defined as toppled when roots and corm were out of the soil, snapped when the corm was broken but remained partly in the soil and broken when the pseudo-stem was broken but the corm remained completely in the soil (Speijer and De Waele, 1997).

Plant toppling is an indicator of severe root and corm destruction that is often associated with nematode attack; snapping is also the result of severe corm destruction but is usually associated with banana weevil attack (Speijer and Gold, 1995); breakage is most often caused by wind (Stover, 1972). Plant height from soil level to the point where the inflorescence protruded from the leaf sheath and pseudo-stem circumference at one meter above ground was measured while the number of suckers and functional leaves was counted (Swennen and De Langhe, 1985). A leaf was considered functional when 70% or more of the leaf surface was not affected by necrosis or senescence (Craenen, 1998; Speijer et al., 2001). For this study, in all, data were collected on 23 variables that are related to plant growth and disease evaluation as well as root health assessment from 70 sites. These variables were transformed to binary data by finding the median and values less than the median represented by 0 while values above the median were represented by 1. The sites were the rows and the collected variables were the columns. Grouping of the data was based on the location of the sites and on the states in Nigeria where they were located. The states are identified by the pointed arrows in Figure 3.

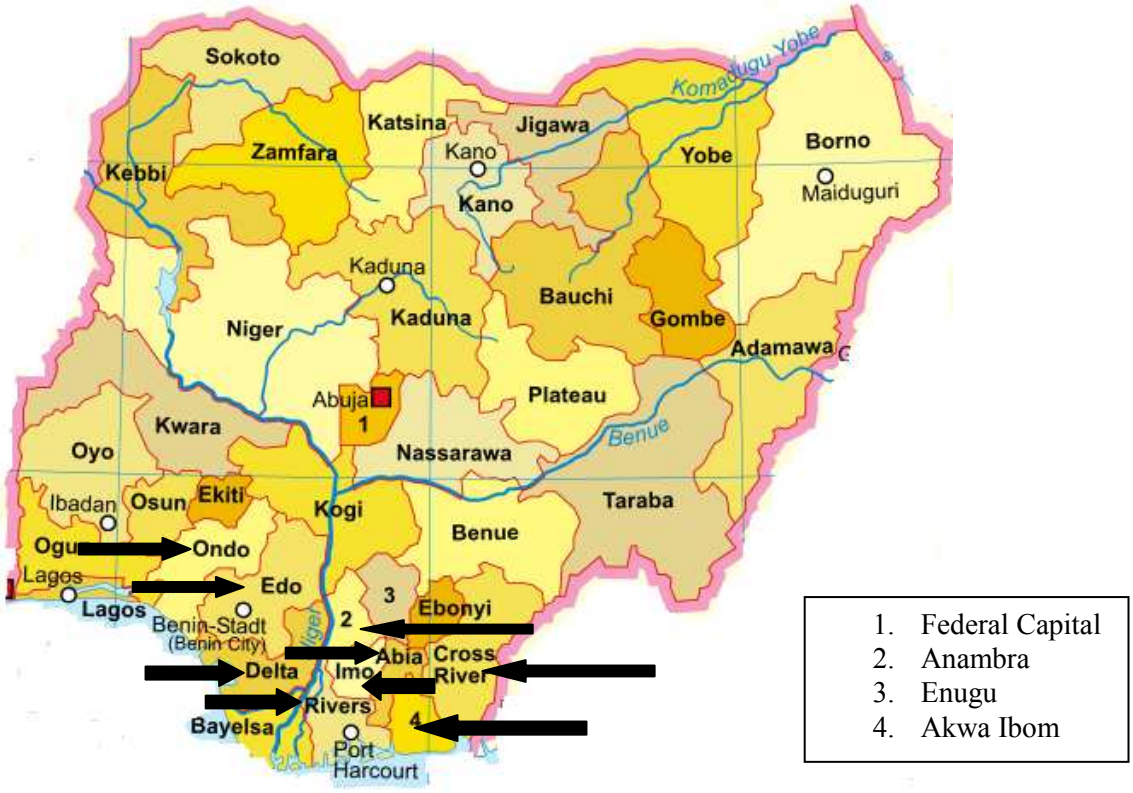


Figure 3: Map of Nigeria showing the nine states involved in the plantain production survey.

Table 3: Classification of states into three regions.

State	No. of farms	Grouping
Abia	5	East
Akwa Ibom	6	East
Anambra	1	East
Cross-River	16	East
Delta	10	Mid-West
Edo	10	Mid-West
Imo	5	East
Ondo	6	West
Rivers	11	East

Similarity estimates between each pair of locations (i,j) were obtained for three similarity coefficients: Dice, Jaccard and Simple Matching, dendrograms were produced for all similarity coefficients using five clustering methods (UPGMA, WPGMA, complete linkage, single linkage and NJ. The different dendrograms were compared by visual inspection and using the CFI. The CFI provides a relative estimate of the dendrogram similarity; it ranges between 0 and 1. Cophenetic matrices were constructed for all dendrograms for the three coefficients. Correlation analyses were carried out between the original similarity values and the cophenetic values. Spearman as well as Pearson correlation coefficients were calculated for the similarity coefficients. Multi Dimensional Scaling (MDS) and PCA were carried out to view the locations in two-dimensional plots and their distribution in space using R (R, 2008). Node counts matrices were constructed and the correlation coefficients calculated to compare node counts of the similarity coefficients for the purpose of comparing the topology of the trees.

2.2.3 Background Information on Powdery Mildew

Powdery mildew fungi are pathogens which belong to the Erysiphales (Ascomycota) and infect a wide range of angiosperm plants. About 650 powdery mildew species are known that occur on almost 10,000 host species (Glawe, 2008). These pathogens are obligate biotrophs, that is, they depend on living plant cells for survival and reproduction. By forming a haustorium that invaginates the epidermal cell of the host plant, the fungus establishes a specific feeding structure that enables the uptake of host nutrients (Oberhaensli et al., 2011).

Wheat and barley powdery mildew disease is a major problem in the crop producing regions of Asia, northern Europe, north and east Africa as well as in north and south America (Curtis et al., 2002). It is usually found on the leaf surface appearing as white fluffy patches, which turn grey when they mature, ranging from small isolated spots to complete leaf coverage and sometimes on the head. Leaves turn yellow-brown as the disease progresses (http://www.hannafords.com/ disease.php?id=10). Infected plants have reduced growth and vigor resulting in impairment on the head and seed filling. Heavily infected leaves and even whole plants can be killed prematurely. It has negative effects on yield quality (Everts et al., 2001) and quantity (Conner et al., 2003) and consequently leads to large economic damage. Yield losses are proportionate to the level of attack. Losses of up to 40% have been recorded and are greatest when the plants are infected in the seedling stage. The spores germinate and infect the leaf surface, where they use available nutrients, thereby reducing photosynthesis and increasing the energy requirements of the host plant. The causal agents, *Blumeria graminis* f. sp. *tritici* (*B.g. tritici*) and *Blumeria graminis* f. sp. *hordei* (*B.g. hordei*), respectively, belong to the cereal powdery mildews (*Blumeria graminis* (*DC*) *Speer*), a single species that comprises eight *formae speciales* (ff. spp.) (Inuma et al., 2007). They can be distinguished by their host specialization because they are restricted to a single host.

2.2.4 Data Collection Method and Analysis for Powdery Mildew Data Set

For this aspect of the study, an excerpt of a data set from a field experiment on evolution of powdery mildew populations in different selection regimes was used. The selection regimes were generated by the application of host resistance genes and fungicide used in the four treatments described as:

Treatment 1–Susceptible host,

Treatment 2–Susceptible host + fungicide,

Treatment 3–Resistant host,

Treatment 4– Resistant host + fungicide.

Samples of mildew isolates were taken out of the mildew populations in the field plots of the four treatments at different time points (1 - 5). For this study, 40 mildew isolates from time point 5 were selected to evaluate the effect of different similarity measures and clustering methods on these isolates. Isolate characteristics were virulence, detected through the 22 differentials. The data were divided into

2 sets, (no fungicide) with treatments 1 and 3 and called Mildewtrt13 while (the fungicide treated) with treatments 2 and 4 were combined and called Mildewtrt24. For Mildewtrt13 data, the treatments were used to group the data into 2 categories, A and B. Isolates with treatment 1 fall into the A category while those with treatment 3 fall into the B category. However, for Mildewtrt24 data, 2 categories C and D were formed, isolates with treatment 2 fall into category C while those with treatment 4 fall into the D category. For each data and for each category, the isolate numbers and the category code were used to identify the different isolates. The aim was to see how the choice of a similarity measure and clustering method affects classification and different analyses were carried out to confirm this. Therefore, genetic similarity estimates between each pair of isolates (i,j) were obtained for three similarity coefficients: Dice, Jaccard and Simple Matching. Dendrograms were produced for all similarity coefficients using five clustering methods (UPGMA, WPGMA, Complete linkage, Single linkage & NJ). The different dendrograms were compared by visual inspection and using the CFI. Cophenetic matrices were also constructed for all dendrograms for the three coefficients and correlation analyses were carried out between the original similarity values and their cophenetic values. Spearman and Pearson correlation coefficients were calculated for the similarity coefficients. Node counts matrices were constructed and correlation coefficients calculated to compare the node counts for the three similarity coefficients. MDS and PCA were carried out to view the isolates in two-dimensional plots and their distribution in space.

2.2.5 Background Information on Yam

Yams (*Dioscorea spp*.) constitute an economically staple food for millions of people in the tropics & 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 water yam (*D. alata* L.). *D. rotundata* is indigenous to West Africa while D. *alata* that was introduced to Africa from Asia in the 16^{th} century was regarded as the most widely cultivated species globally. *D. alata* has better characteristics for sustainable production, with high yield potential (especially under low to average soil fertility). It can be easily propagated, has early vigor for weed suppression and storability of tubers. However, its major drawback in the field is the susceptibility of most cultivars to anthracnose disease which has a great impact on its productivity.

The use of durable host plant resistance in D. *alata* against yam anthracnose disease will contribute significantly to an increased level and stability of field performance.

Anthracnose (*Colletotrichum gloeoporioides*) attacks all plant parts at any growth stage appearing first on leaves as small and irregular yellow, brown, dark-brown, or black spots. The spots can expand and merge to cover the whole affected area. The color of the infected part darkens as it ages and the symptoms are most visible on leaves. It causes leaf necrosis and dieback of yam vines, resulting in a reduction in the effective photosynthetic surface area of the crop with a concomitant reduction in the ability of the yam tuber to store food reserves. Epidemics that commence before or during tuber formation can have a great effect on tuber yield. Successful control of anthracnose disease would encourage greater widespread cultivation and significant increases in overall production to meet the high local and overseas demand for yam (Abang et al., 2003).

2.2.6 Collection Method and Analysis of Yam Anthracnose Disease Data

The AFLP marker was analysed using a modified method of 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, dendrograms were produced for all similarity coefficients using five clustering methods (UPGMA, WPGMA, Complete, Single and NJ) as previously explained for the Powdery Mildew Data Set (Page 22).

2.3 Analysis and Comparison of Data

The simulated and experimental data were analysed using different methods and their results were compared using different methods like CFI, PCA, MDS and correlation coefficients. In the simulated data, only Dice and Jaccard measures were used. However, a third measure, Simple matching, was introduced in the analyses of the experimental data.

2.3.1 Trees Based on Dice, Jaccard and Simple Matching

For each sample generated, dendrograms (trees) were constructed using UPGMA, WPGMA, NJ, single linkage and complete linkage CA methods for the Dice, Jaccard and Simple matching coefficients. Cophenetic matrices of the trees were also calculated.

2.3.2 Consensus Fork Index

In this study, the strict consensus method was used. The 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). It's worthy of note that care should be taken in the calculation of CFI. It is therefore advised that proper pruning of the data should be carried out to avoid unnecessary repetition of the objects to be compared using the construction of dendrogram and CFI values. The CFI is appealing due to its easy and simple interpretation as the proportion of the possible subsets for n objects that are in the consensus for two classifications (Rohlf, 1982).

2.3.3 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 for all data sets.

(ii). Node count matrices were generated for the Dice, Jaccard and Simple-Matching trees for all 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.

2.3.4 Multidimensional Scaling

MDS is a statistical technique used to visualize dissimilarity data. It is an ordination technique for representing the dissimilarity among n objects or variables by n points in a k-dimensional space so that the inter-point distances in the k-dimensional space correspond as well as possible to the observed distances between the objects (Groenen and Van de Velden, 2004). The major assumption in MDS is that responses can be described by values along a set of dimensions that places these responses as points in a multidimensional space and that the similarity between the responses is inversely related to the distances of the corresponding points in the multidimensional space (Steyvers, 2002). The aim of MDS is to arrange the investigated objects on a line or on a plane, or in a space of higher dimension, so that their mutual location would reflect, as far as possible, the degree of likeness or unlikeness between them.

MDS can be applied with different purposes. One of them is explanatory data analysis which can be achieved by placing objects as points in a low dimensional space, the observed complexity in the original data matrix can often be reduced while the essential information in the data is still preserved. Researchers are able to visually study the structure in the data by a representation of the pattern of

proximities in two or three dimensions (Steyvers, 2002). The mental representation of responses that explains how similarity judgments are generated was also discovered through MDS. Sometimes, it reveals the psychological dimensions hidden in the data that can meaningfully describe the data.

The multidimensional representations resulting from MDS have also been seen to be often useful as the representational basis for various mathematical models of categorization, identification, and/or recognition memory (Nosofsky, 1992) or generalization (Shepard, 1987; Steyvers, 2002). It is an alternative method of cluster analysis in the sense that from the resultant final configuration of points in two- or three-dimensional space, one could obtain information about the structure of corresponding set of objects (Vandev and Tsvetanova, 1995).

2.3.5 Principal Component Analyses

PCA is a classical statistical method; it is a linear transform that has been widely used in data analysis and compression. It involves the transformation of a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible and each succeeding component accounts for as much of the remaining variability as possible (Erkki, 1989). The purpose is to determine the class of an observation based on a set of variables known as predictors or input variables. The model is built based on a set of observations for which the classes are known. It is useful in the identification of the independent variables that discriminate a nominally scaled dependent variable of interest. PCA and FA are ordination techniques while CA and DA are classification techniques. These methods are useful tools in multivariate analysis especially in finding groups and pattern in data. The ordination methods are graphically used to display data in two or more dimensions. Among these methods, CA differs in that it does not involve any a priori hypotheses and provides easy interpretation (Meyer et al., 2004).

Two major objectives of PCA are to discover or to reduce the dimensionality of the data set and to identify new meaningful underlying variables. PCA is an eigen-analysis-based method. It is the simplest and oldest eigen-analysis-based method. It is a rigid rotation of the original data matrix, and can be defined as a projection of samples onto a new set of axes. The maximum variance in the data is

projected along the first axis, the maximum variation uncorrelated with the first axis is projected on the second axis, the maximum variation uncorrelated with the first and second axis is projected on the third axis and so on (Palmer, 2008). It is a way of identifying patterns in data, and expressing the data in such a way so as to highlight their similarities and differences. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analyzing data. Another main advantage of PCA is that once these patterns have been found in the data, one can compress the data, that is, by reducing the number of dimensions, without much loss of information (Smith, 2002).

Technically, a principal component can be defined as a linear combination of optimally-weighted observed variables. In order to understand the meaning of this definition, it is necessary to first describe how subject scores on a principal component are computed. In the course of performing a principal component analysis, it is possible to calculate a score for each subject on a given principal component. For example, if there are 10 variables in a data set, each subject in the data would have scores on ten components. The subject's actual scores on the ten variables would be optimally weighted and then summed to compute their scores on a given component.

In reality, the number of components extracted in a principal component analysis is equal to the number of observed variables being analysed. However, in most analyses, only the first few components account for meaningful amounts of variance, so only these first few components are retained, interpreted, and used in subsequent analyses. For instance, in the example given above with ten variables in a given data set, it is likely that only the first two components would account for a meaningful amount of variance; therefore only these would be retained for interpretation. It is usually assumed that the remaining eight components accounted for only trivial amounts of variance. These latter components would therefore not be retained, interpreted, or further analysed (SAS, 2011). Therefore PCA was carried out on the samples to be able to see maximum variability and pattern in the data as well as to compare the grouping of these objects with those from MDS.

3 Results

3.1 Results from Simulated Data

Completely randomly generated data and data with specific defining properties representing two sets of data were simulated. Each sample had two groups that were compared using the Dice and Jaccard similarity measures and the UPGMA clustering method. The members of the two groups showed some mingling in some samples while there was no mingling in some. The generated dendrograms were also compared using the CFI values that ranged between 0 and 1, where 0 depicts no similarity and 1 depicts complete similarity. The groupings in each sample were also compared through MDS and PCA.

3.1.1 Mingling of Objects from Two Different Groups

The dendrograms for samples with C_{middle}–length above 100 columns showed some mingling (i.e. the objects within the two groups were not well separated). Whereby, some objects from group A are mixed with objects from group B (Figures 5, 6 and 7). However, there were no mingling of objects for samples with C_{middle}–length that ranged between 10 and 100 columns. This suggests that the higher the number of non–discriminating factors or characteristics being measured, the higher the possibility of mixing of objects from the two groups A and B. A summary of the simulation is given in Tables 4 and 5.

3.1.2 Consensus Fork Index (CFI) values

The CFI results for samples with in-built grouping revealed that samples with C_{middle} varying between 10 and 100 showed no mingling of objects and the percentage of the total samples that had CFI values less than 1 were lower than in samples where there were mingling (Table 4). According to the summary of the simulations given in Table 4, even though there was no mingling of objects from the two groups, the minimum CFI value for this set of simulations ranged between 0.43 and 0.75. Out of a thousand samples per simulation for the different parameters given, for r = 30, and length of C_{middle} varying between 10 and 100, the number of samples that had CFI value less than 0.8 ranged between 0.9% and 3.2% while for length of C_{middle} above 100 (Table 5), for r = 30, the value is between 2.2% and 3.0% and for r = 40, is between 2.0% and 3.7%. This percentage increased as the middle section increases, and then dropped, to increase again. This suggests that no linear relation exists between the

number of samples with CFI less than 0.8 and the length of C_{middle}. However, samples with C_{middle} above 100 and with low CFI showed some mingling in the separation of the members of the two groups (Figures 5, 6 and 7).

The similarity in the dendrograms generated using the two measures is not surprising; taking into consideration the fact that there is just a slight difference in their formulas. Although most of the dendrograms generated were similar, contrasting them by the CFI result (Table 5), revealed some differences among them as seen in the percentage of the samples that had CFI values that are less than 1. Based on the general belief that Dice and Jaccard measures produce similar results from cluster analysis, the low CFI values for comparing dendrograms from both Dice and Jaccard measure for some of these samples suggest that this is not always so. A CFI as low as 0.393 or any value less than 0.5 implies that the structure of the two trees being compared are not similar since a CFI of 1 is associated with topologically identical trees. In order to clarify the similarity between the trees, matrix correlation coefficient was calculated between Dice and Jaccard similarity matrices. It was observed that for some of the samples that had low CFIs (even as low as 0.393 and 0.47), the matrix correlation coefficient between the similarity matrices of the two measures was as high as 0.99. This shows that despite high correlation, topology could be considerably different. Therefore, correlation alone cannot be used to measure topology. Though the CFI of some of the samples were not too low (higher than 0.5), the structure of the trees differ and the expected thorough separation of the members of the two groups was not observed. The objects in the two groups still mixed together which shows that the Consensus Fork Index alone cannot be used to determine topology.

Comparing samples with r = 30 and r = 40 and the same length of C_{middle} the number of CFI < 0.8 tend to be unstable. It increased and then decreased, to increase again suggesting that increasing the number of rows does not have a linear relationship with the number of samples with CFI < 0.8. The summary of the simulation (Table 5) shows that the minimum CFI ranges between 0.39 and 0.57. The mean CFI for all simulation runs ranged between 0.972 and 0.980 and the standard deviation ranged between 0.050 and 0.061. The increase in the number of rows did not affect the mean CFI or the standard deviation, however, an increase was observed in the percentage of samples with CFI values of less than 1.

Table 4: Summary of simulation parameters and CFI distribution for C_{middle} from 10 to 100.

No. of rows	Length of C_{middle}	Min. CFI	CFI < 0.8	CFI < 1
30	10	0.75	0.9%	19.4%
30	20	0.5357	1.8%	25.7%
30	30	0.6429	2.8%	31.3%
30	40	0.6429	3.0%	24.7%
30	50	0.4286	2.5%	28.1%
30	60	0.6786	3.0%	26.3%
30	70	0.6429	2.4%	26.0%
30	80	0.6071	3.2%	28.0%
30	90	0.5714	2.1%	24.8%
30	100	0.5714	2.0%	25.2%

Table 5: Summary of simulation parameters and CFI distribution for C_{middle} above 100.

No. of rows	Length of C_{middle}	Min. CFI	CFI < 0.8	CFI < 1	Mean CFI	Median CFI	SD
30	130	0.5714	2.2%	26.3%	0.9778	1	0.0516
30	150	0.4600	2.7%	24.4%	0.9761	1	0.0594
30	170	0.3929	3.0%	23.1%	0.9779	1	0.0589
30	200	0.5357	2.5%	21.5%	0.9801	1	0.0501
40	130	0.4737	3.0%	32.2%	0.9729	1	0.0576
40	150	0.5263	3.0%	31.7%	0.9739	1	0.0570
40	170	0.5000	2.0%	29.0%	0.9765	1	0.0542
40	200	0.4700	3.7%	32.2%	0.9721	1	0.0613

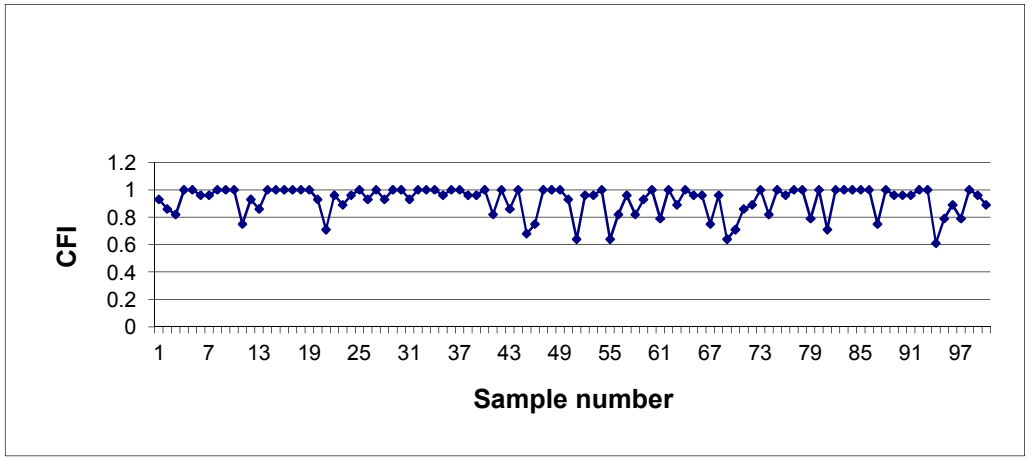


Figure 4: Consensus fork index for Dice and Jaccard.

The CFI results for 100 generated samples without in-built grouping (i.e. completely random), showed that about 16% of the samples had low CFI values ($0.6 \leq CFI < 0.8$). A plot of the CFI values for Dice and Jaccard is shown in Figure 4.

3.1.3 MDS and PCA Results for Trees with Low CFI (less than 0.8)

MDS was carried out for the samples with low CFI and mixing of objects from different groups in the corresponding UPGMA dendrograms. The objective of this analysis was to see whether the structure in the data will still be preserved in the MDS plots generated for these samples so as to confirm the results obtained from the dendrograms. The MDS plot and PCA plots of the first two axes showed that the structure in the data was preserved. However, a plot of the higher axes, showed some mingling (Figure 8).

PCA on the samples with low CFI values also revealed some mingling among the objects. A plot of the principal axis 1 against the principal axis 2 showed the perfect separation of the objects within each group. However, plots of higher principal axes that depict less variation in the data against each other revealed some more mingling among the objects of the two groups. It was observed that the MDS plot and a plot of principal axis 1 against axis 2 from the PCA produced similar results with respect to the

classification of the objects within the two groups. However, in the dendrograms, there was mingling (less than 10%) of objects from the two groups.

In Figure 5, the dendograms showed mingling for Jaccard coefficient (Figure 5A) and perfect separation for Dice (Figure 5B). The sample had $C_{middle} = 170$ columns, with 30 rows and the CFI was 0.393 while in Figures 5C and 5D, there was mingling among the objects of the two groups for Dice coefficient and perfect separation for Jaccard. This sample had $C_{middle} = 200$, with 30 rows and the CFI was 0.54. For the Dice dendrogram, B06 and B13 joined with the 'A' group while B05 and B10 formed a separate group. In Figure 6 however, there was mingling for both measures. $C_{middle} = 150$ columns, with 30 rows and CFI was 0.64. B14 join with the 'A' group in Jaccard (Figure 6A) while B05 and B08 formed another group while in Dice (Figure 6B), B14 formed a lone group, B05, B08 and B09 formed another group and A10 joined with B11 and the remaining 'B' group. Figures 7A and 7B also revealed mingling for both measures, with C_{middle} being 200 columns, 40 rows, and the CFI was 0.47. In Jaccard dendrogram, (Figure 7A), A16 joined with the 'B' group and all others. However in Dice, (Figure 7B), A19 mingled with the 'B' group while B14 joined the 'A' group.

The results from these dendrograms for both Dice and Jaccard measures showed that both measures would produce similar results in most situations. This suggests the result may not be unconnected with the fact that both measures do not give importance to the 0:0 factor (that is,'d') in their formulas (Table 1). It was also observed that mingling could occur irrespective of the size of the number of columns. However, for the experimental data, the Simple matching coefficient will be included along with the other two being analysed to see what effect this coefficient has in the different cases. It is to be recalled that the Simple matching coefficient does include the 0:0 factors ('d') in its formula as seen in Table 1. This could mean that the absence of a particular trait or character in the two individuals being compared is important to the researcher.

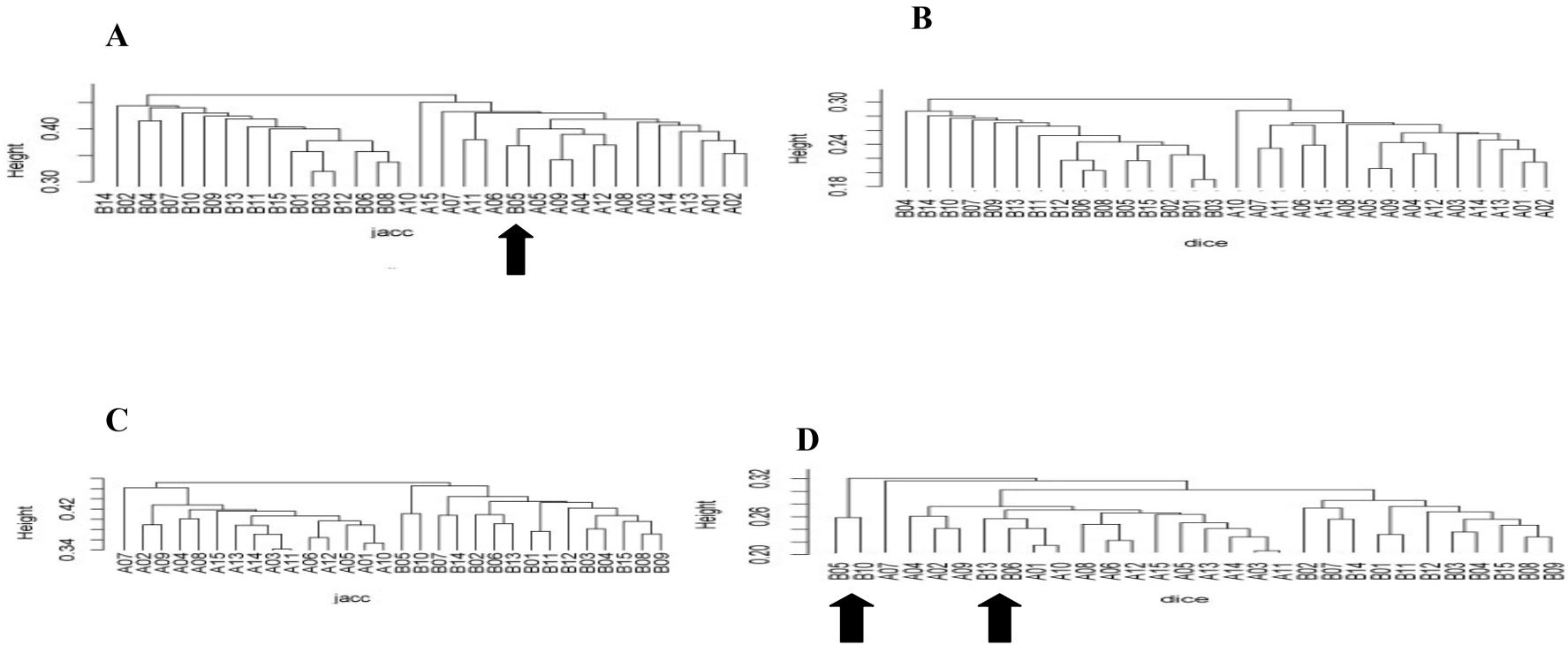


Figure 5: Dendrograms showing mingling and perfect separation for both Dice and Jaccard measures.

A and B - mingling for Jaccard (left) and perfect separation for Dice (right), C_{middle} = 170 columns, 30 rows and CFI = 0.393. C and D - mingling for Dice (right) and perfect separation for Jaccard (left). C_{middle} = 200, r = 30 and CFI = 0.54.

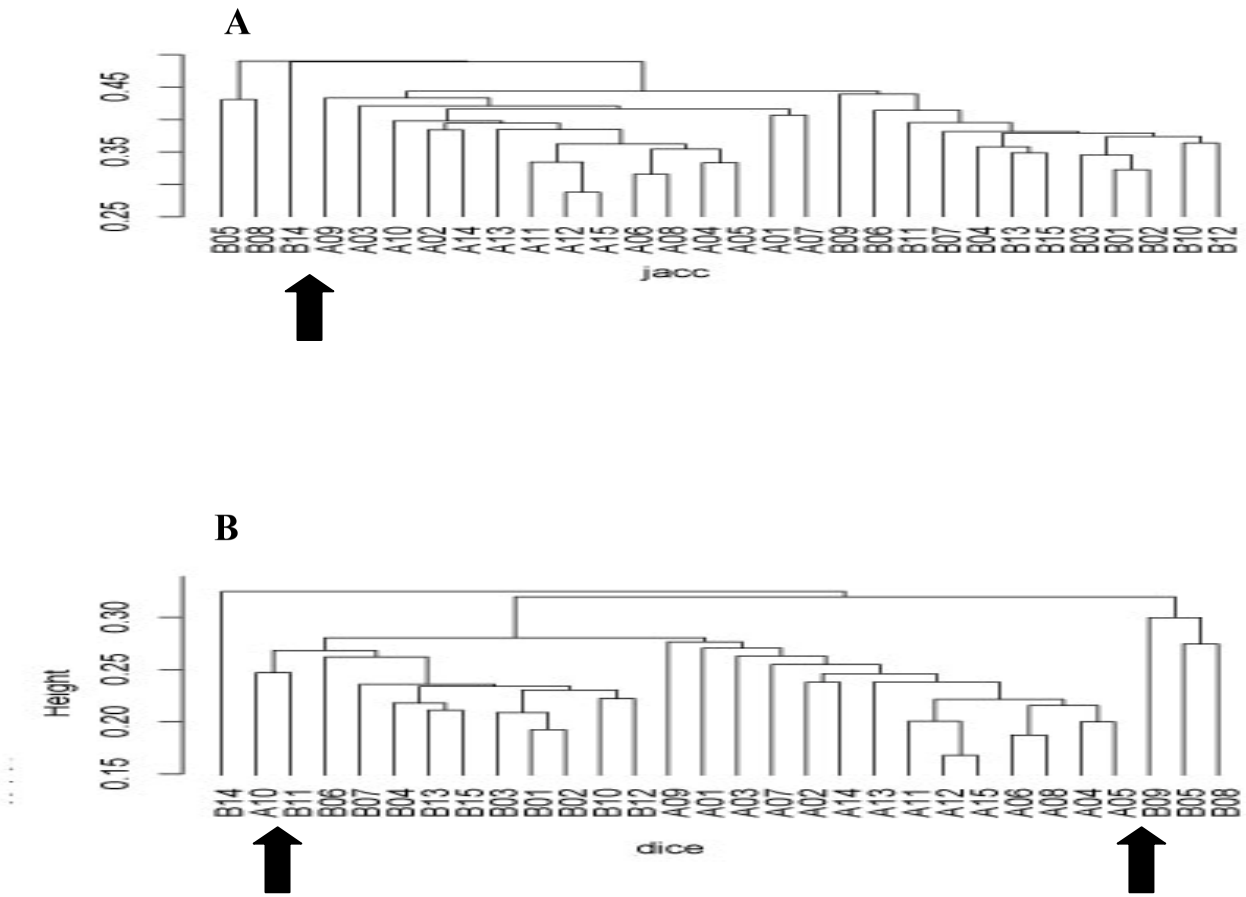


Figure 6: Dendrograms showing mingling for Dice and Jaccard measures with CFI = 0.64.

A and B - with C_{middle} = 150, r = 30

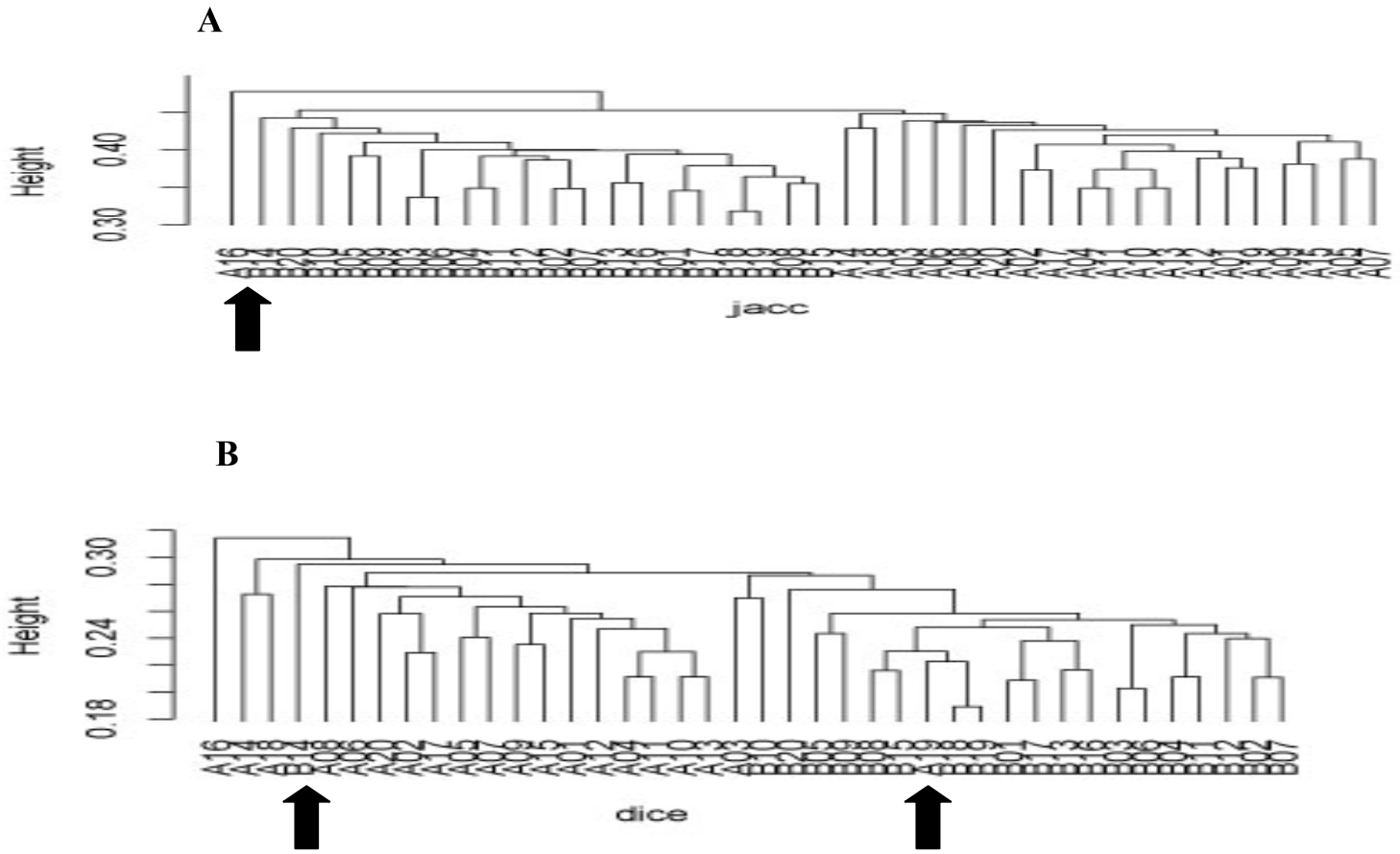


Figure 7: Dendrograms showing mingling for both Dice and Jaccard measures CFI = 0.47.

A and B with C_{middle} = 200 columns, 40 rows and CFI = 0.47.

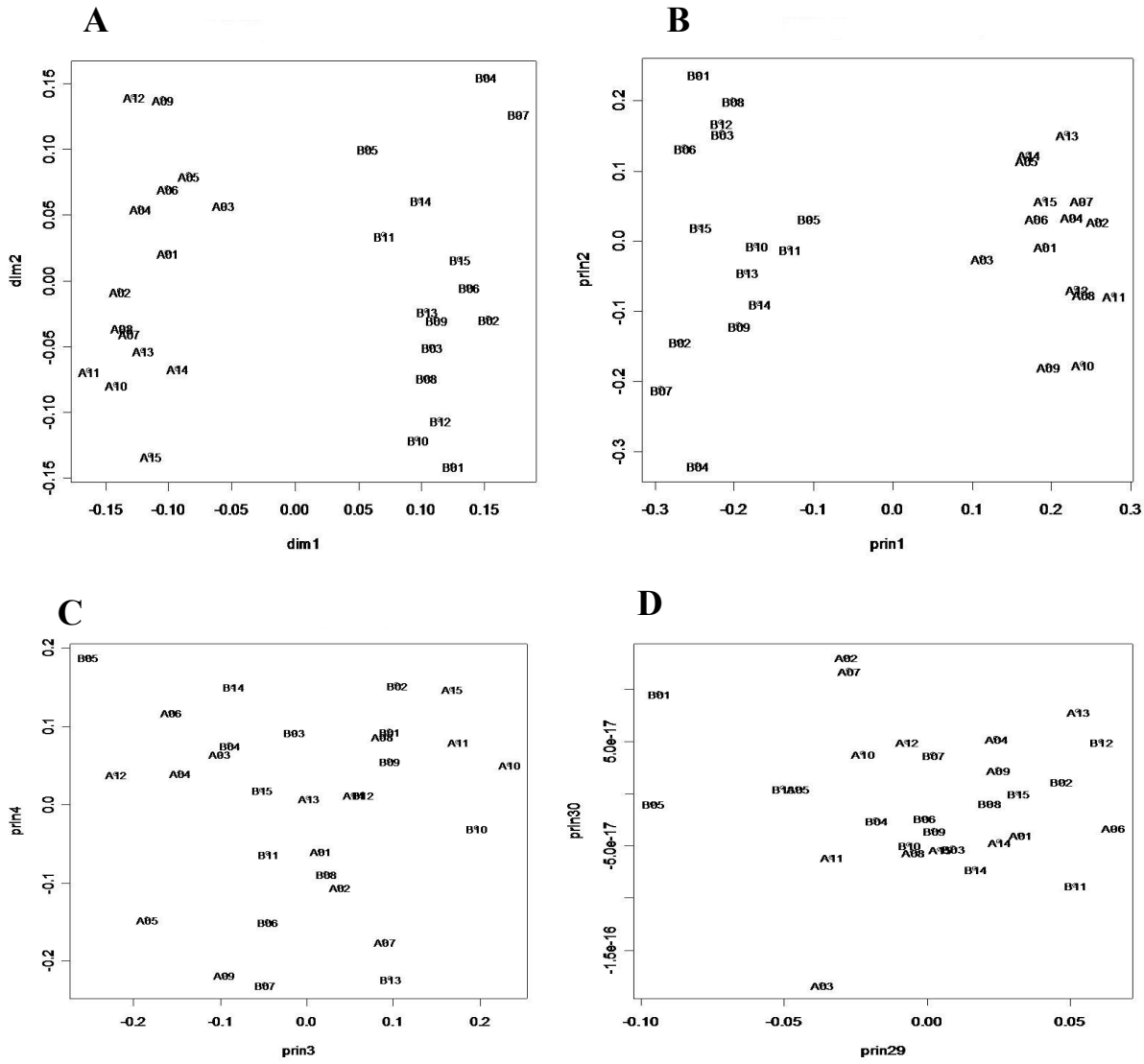


Figure 8: MDS and PCA plots for Jaccard measure.

Grouping in the data maintained in A and B. A- MDS plot, B- Axis 1 versus Axis 2 of PCA plot, C – Axis 3 versus Axis 4 and D – Axis 29 versus Axis 30 of PCA plots.

3.1.4 MDS and PCA Plots from Simulated Data Sets

The MDS result for both Dice and Jaccard coefficients revealed that there was total separation of the objects within the two groups, which confirms the structure in the data. There was similarity between the MDS plot and a plot of principal component 1 against principal component 2. This also confirmed the groupings within the data. However, a plot of principal component 3 against principal component 4 showed mingling between the objects of the two groups. Dividing the plot area into the positive side and the negative side from the X-axis, a mingling of about 43% was observed among the objects of the two groups on the positive side while a mingling of about 57% was observed on the negative side.

3.2 Results from Experimental Data Sets

Plantain and yam anthracnose data sets from Nigeria and powdery mildew data from Germany were analysed using different clustering methods for CA, PCA as well as MDS to compare the resulting classifications from the analyses. Three similarity measures (Dice, Jaccard and Simple matching) were used in the CA and five clustering methods. The objects in these data were the locations (states) where the Diagnostic Survey Samples on production constraints on plantain data were collected in Nigeria, powdery mildew isolates from Germany and yam pathogens from two agro-ecological zones in Nigeria. The effect of the similarity measures and CA methods on these locations and their distribution in space were investigated.

3.3 Diagnostic Survey Sample for Plantain (DSS Plantain) data

3.3.1 Dendrograms for 5 CA Methods and 3 Similarity Coefficients

In the DSS plantain data, the different locations (states) where the surveys were carried out were subjectively grouped into 12 clusters by the UPGMA method (Table 6), 11 clusters by the WPGMA, complete linkage and NJ methods while the single linkage method gave 8 clusters with more than 15 singletons. The UPGMA, complete linkage and single linkage produced identical classifications for both Dice and Jaccard coefficients while the WPGMA and NJ methods produced different classifications for the two coefficients. It was observed that there was a mixture of the states even in the regional groupings. Dendrograms showing the classification of the different locations for the Jaccard and Simple matching similarity coefficients and for the UPGMA clustering methods are

presented in Figures 9 and 10. It was also observed that some clusters were the same in the three dendrograms, however, the clusters in the Dice dendrogram are exactly the same as in the Jaccard dendrogram. Details of the classifications for other clustering methods apart from UPGMA as seen in the dendrograms are presented in Table A1-A4 (Appendix A).

3.3.2 MDS and PCA Results

The MDS and PCA results for the Simple matching coefficient are shown in Figure 11. In the DSS-Plantain data, three major groupings were found for all the three coefficients. The first three principal components for the Dice measure accounted for 65% of the total variation in the data while for Jaccard measure, they accounted for only 60% of the variation and in the Simple Matching, they accounted for 63% (Table 7). The MDS plot and the PCA plot of the first two principal axes gave the same grouping, although the PCA plot was slightly rotated in the resulting grouping. This comparison of the results provided by the bi-dimensional graphical dispersion of the different locations showed a lot of mixing among them. For the three coefficients, the plot of the principal component 1 against the principal component 2 revealed 3 major clusters. The plot for the Simple matching coefficient is as shown in Figure 11; it was observed that none of the states formed a unique group of its own. The PCA plot for the Dice measure showed that in the first group, the locations were a mixture of the eastern and mid-western region. Cross rivers state had about 56%, Rivers state had 54%, Imo, Akwa-Ibom and Anambra had 100% representation each in the group, Abia state had 80% while Delta and Edo states had 30% and 10% respectively. The second and third groups had a mixture of all the three regions. The second group consisted of Cross rivers state with 44% representation of the locations, Edo state with 70%, Delta state with 40% and Ondo state with 50%. In the third group, there were 33% representation from Ondo state, 20% from Delta state, 10% from Edo and about 18% from Rivers state. Overlapping in the plot was observed for some locations, which makes some farms in some states not to have a 100% report of the representation of the farms in these states.

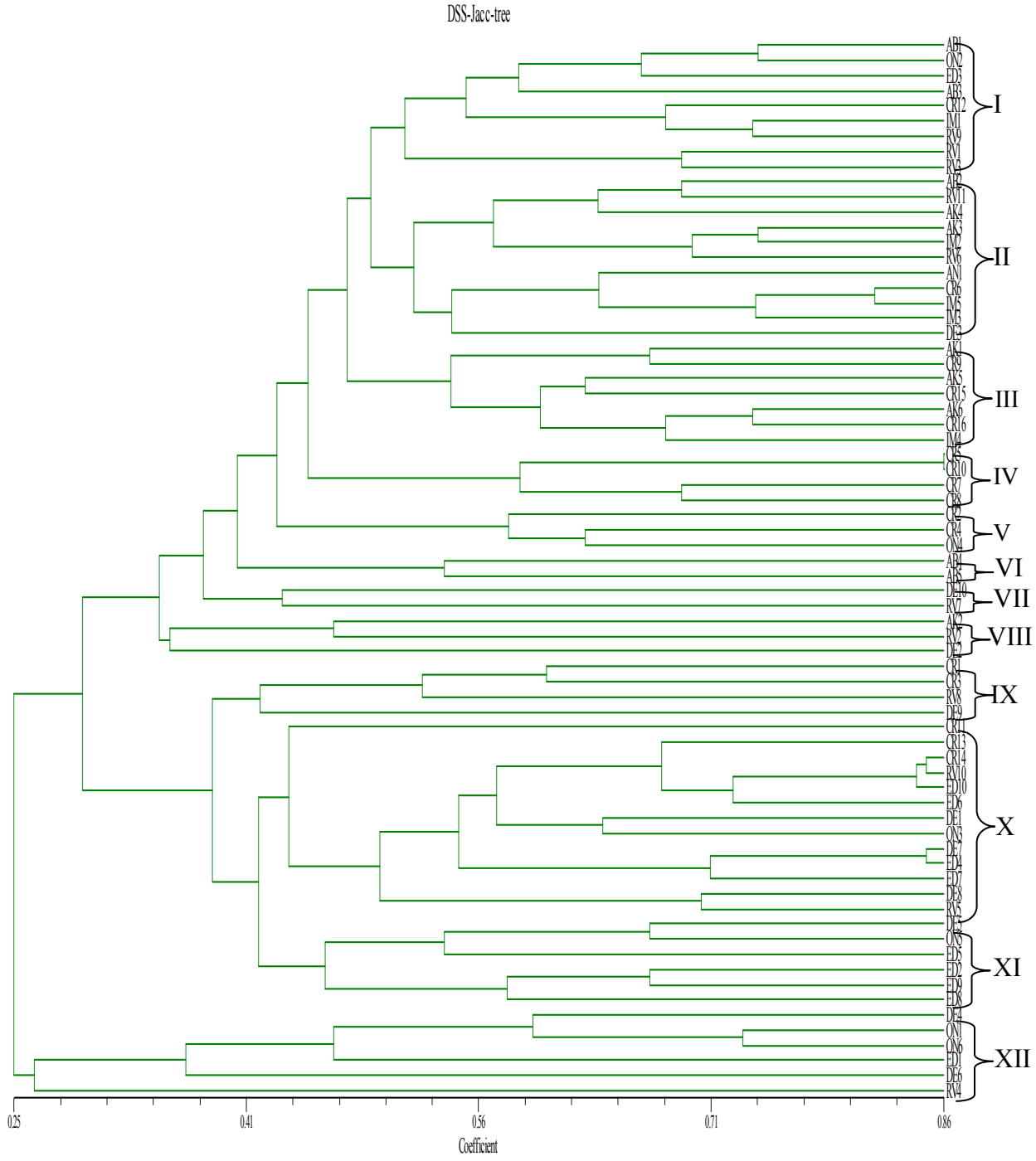


Figure 9: Jaccard based UPGMA dendrogram of the Plantain dataset.

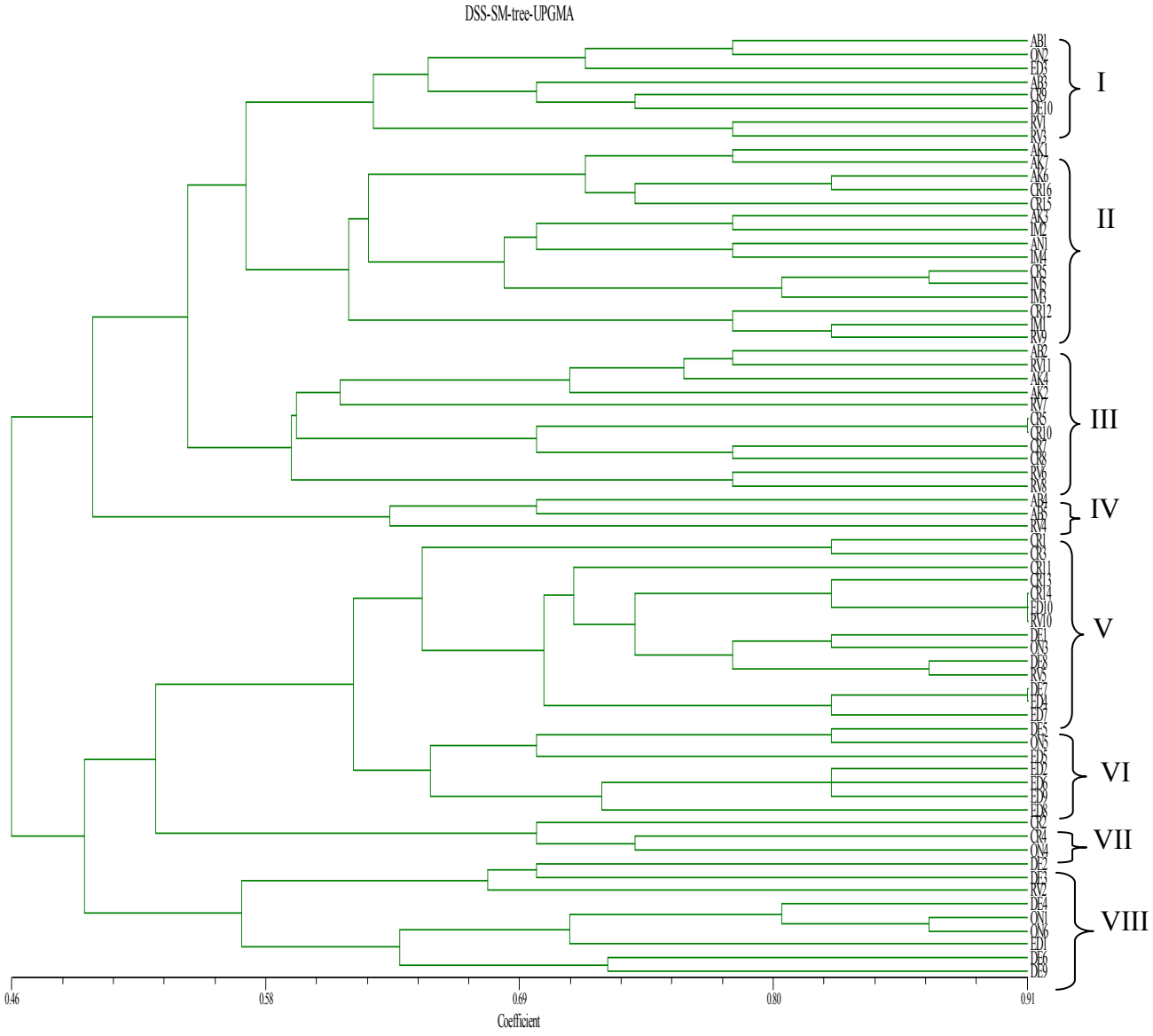


Figure 10: Simple Matching based UPGMA dendrogram of the Plantain.

Table 6: DSS-Plantain data clusters as seen in the dendrogram using UPGMA method.

Cluster	Dice and Jaccard	Simple Matching
I	AB1,ON2,ED3,AB3,CR12,IM1 RV9,RV1,RV3	AB1,ON2,ED3,AB3,CR9, DE10,RV1,RV3
II	AB2,RV11,AK4,AK3,IM2,RV6, AN1,CR6,IM5,DE3	AK1,AK7,AK6,CR16,CR15, AK3,IM2,AN1,IM4,RV9
III	AK1,CR9,AK5,CR15,AK6, CR16,IM4	AB2,RV11,AK4,AK2,RV7,CR5, CR10,CR7,CR8,RV6,RV8
IV	CR5,CR10,CR7,CR8	AB4,AB5RV4
V	CR2,CR4,ON4	CR1,CR3,CR11,CR13,CR14,ED10,RV10,DE1, ON3,DE8,RV5,DE7,ED4,ED7
VI	AB4,AB5	DE5,ON5,ED5,ED2,ED6,ED9,ED8
VII	DE10,RV7	CR2,CR4,ON4
VIII	AK2,RV2,DE2	DE2,DE3,RV2,DE4,ON1,ON6,ED1, DE6,DE9
IX	CR1,CR3,RV8,DE9	
X	CR11,CR13,CR14,RV10,ED10, ED6,DE1,ON3,DE7,ED4,ED7, DE8,RV5	
XI	DE5,ON5,ED5,ED2,ED9,ED8	
XII	DE4,ON1,ON6,ED1,DE6,RV4	

Table 7: Principal components proportion for plantain data.

Data	Principal component	% of each component	Accumulated percent
DSS- Plantain Dice	1	34.78	34.78
	2	19.05	53.83
	3	11.58	65.41
	4	7.45	72.86
	5	6.47	79.33
DSS- Plantain Jaccard	1	32.62	32.62
	2	16.48	49.10
	3	10.94	60.04
	4	7.26	67.30
	5	6.14	73.44
DSS- Plantain SM	1	37.06	37.06
	2	15.96	53.02
	3	10.28	63.30
	4	8.51	71.81
	5	6.19	78.00

SM – Simple matching

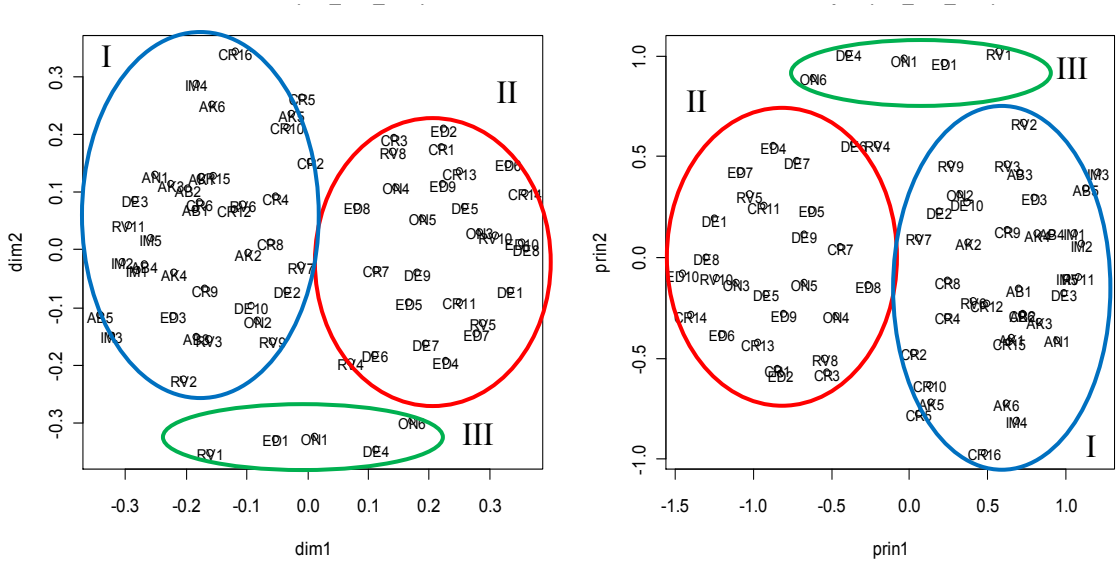


Figure 11: Simple matching MDS & PCA prin1 versus prin2 plot for plantain dataset.

3.4 Powdery Mildew Data

These data had two parts and two groups within each group. Mildewtrt13 consisted of data with treatments 1 and 3 (no fungicide) while Mildewtrt24 consisted of data with treatments 2 and 4 (with fungicide). The treatments were used as the groups in each data. The objects in the groups were samples of mildew isolates taken from mildew populations. Each sample consisted of 40 isolates and 22 differentials. Dendrograms were constructed using the Dice, Jaccard and Simple matching similarity coefficients for five CA methods. MDS and PCA were also carried out on the data. The results are discussed for each analysis.

3.4.1 Dendrogram Results for Isolates with Treatment 1 and Treatment 3

In the powdery mildew data Mildewtrt13, the UPGMA method produced eight clusters for the three measures; WPGMA method produced five clusters for Dice and Jaccard measures and six clusters for Simple matching. The complete linkage method also produced five clusters for Dice and Jaccard measures and seven clusters for the Simple matching. However, the single linkage method produced four clusters and lots of singletons for Dice and Jaccard, and one major cluster and a singleton for Simple matching. On the other hand, the NJ method produced five, seven and six clusters for the Dice,

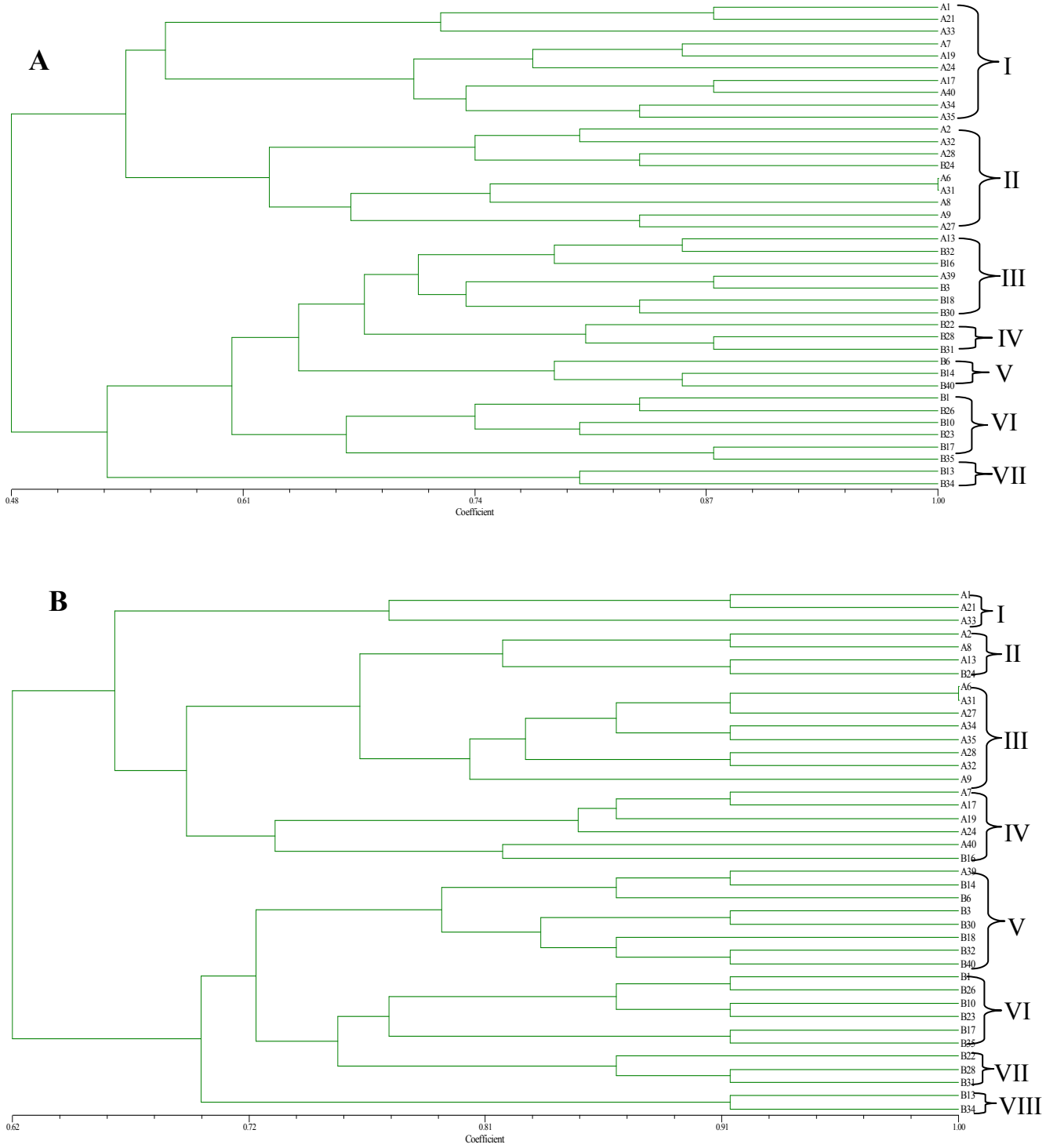


Figure 12: Jaccard and Simple matching based UPGMA dendrogram for Mildewtrt13 data set.

A – Jaccard; B - Simple matching

Table 8: Mildewtrt13 data clusters as seen in the dendrogram based on UPGMA method.

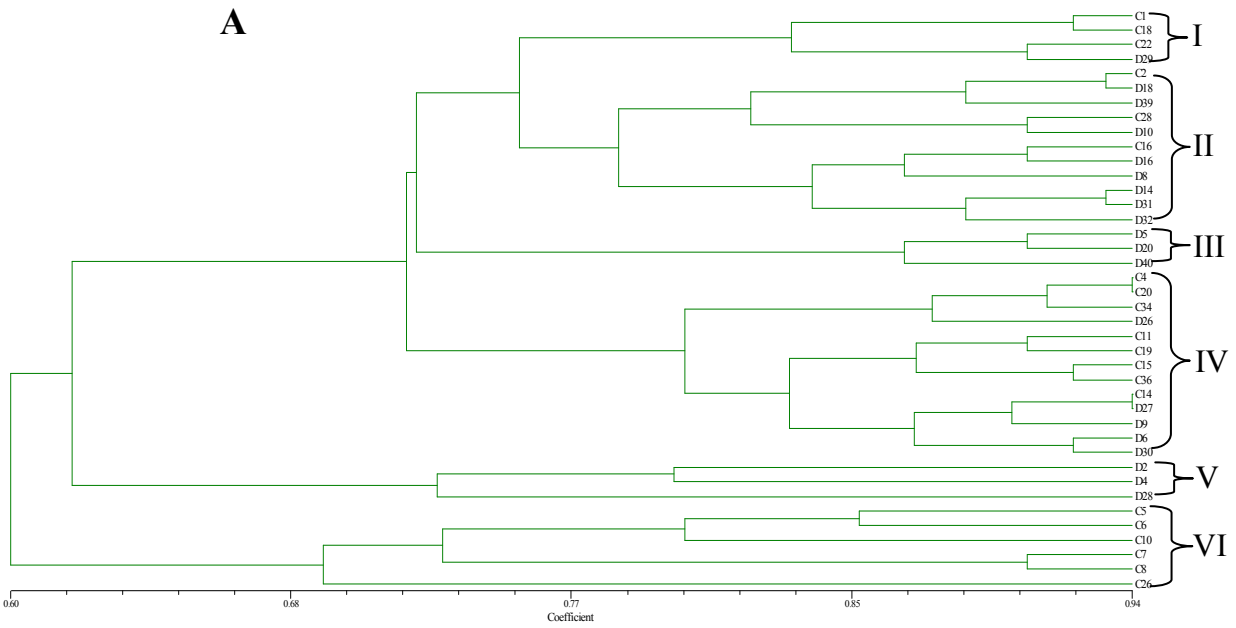
Cluster	Dice and Jaccard	Simple Matching
I	A1,A21,A33,A7,A19,A24,A17, A40,A34,A35	A1,A21,A33
II	A2,A32,A28,B24,A6,A31,A8, A9,A27	A2,A8,A13,B24
III	A13,B32,B16,A39,B3,B18,B30	A6,A31,A27,A34,A35,A28,A32,A9
IV	B22,B28,B31	A7,A17,A19,A24,A40,B16
V	B6,B14,B40	A39,B14,B6,B3,B30,B18,B32,B40
VI	B1,B26,B10,B23,B17,B35	B1,B26,B10,B23,B17,B35
VII	B13,B14	B22,B28,B31
VIII		B13,B14

Jaccard and Simple matching coefficients respectively. It was observed in all methods that the isolate B24 was always grouped with some other isolates from the A group while the isolates A13 and A39 were always grouped with some other isolates from the B group. The results of the UPGMA method for 40 isolates with 22 differentials with two groups, represented by A and B, for treatments 1 and 3 from the data are shown in Figure 12. The dendrograms for the Dice and Jaccard similarity coefficients were identical which was also reflected in the table of clusters as seen in the dendrogram (Table 8). The dendrograms showed that only the NJ method produced different classifications for Dice and Jaccard measures (Table B1-B4, Appendix B).

3.4.2 Results for Isolates with Treatment 2 and Treatment 4

The dendrogram results of the UPGMA method for 40 isolates with 22 differentials with two groups, represented by C and D, for treatments 2 and 4 from the data Mildewtrt24 are shown in Figure 13. It was observed that only the complete and single linkage methods produced identical classifications for Dice and Jaccard measures (Table B5-B8, Appendix B). The UPGMA and WPGMA produced six, five and six clusters for Dice, Jaccard and Simple matching measures respectively. However, the complete linkage method produced six clusters for both Dice and Jaccard measures and five clusters for Simple matching while the single linkage method produced two major clusters for Dice and Jaccard

measures and one cluster and three singletons for Simple matching. The NJ method produced six clusters each for both Dice and Simple matching and five clusters for Jaccard. More mingling was observed among the members of the two groups compared to the Mildewtrt13 data, the application of the fungicide could be responsible for this difference. The clusters as seen in the dendrograms for the three coefficients using the UPGMA method are shown in Table 9.



continued

Figure 13 continued

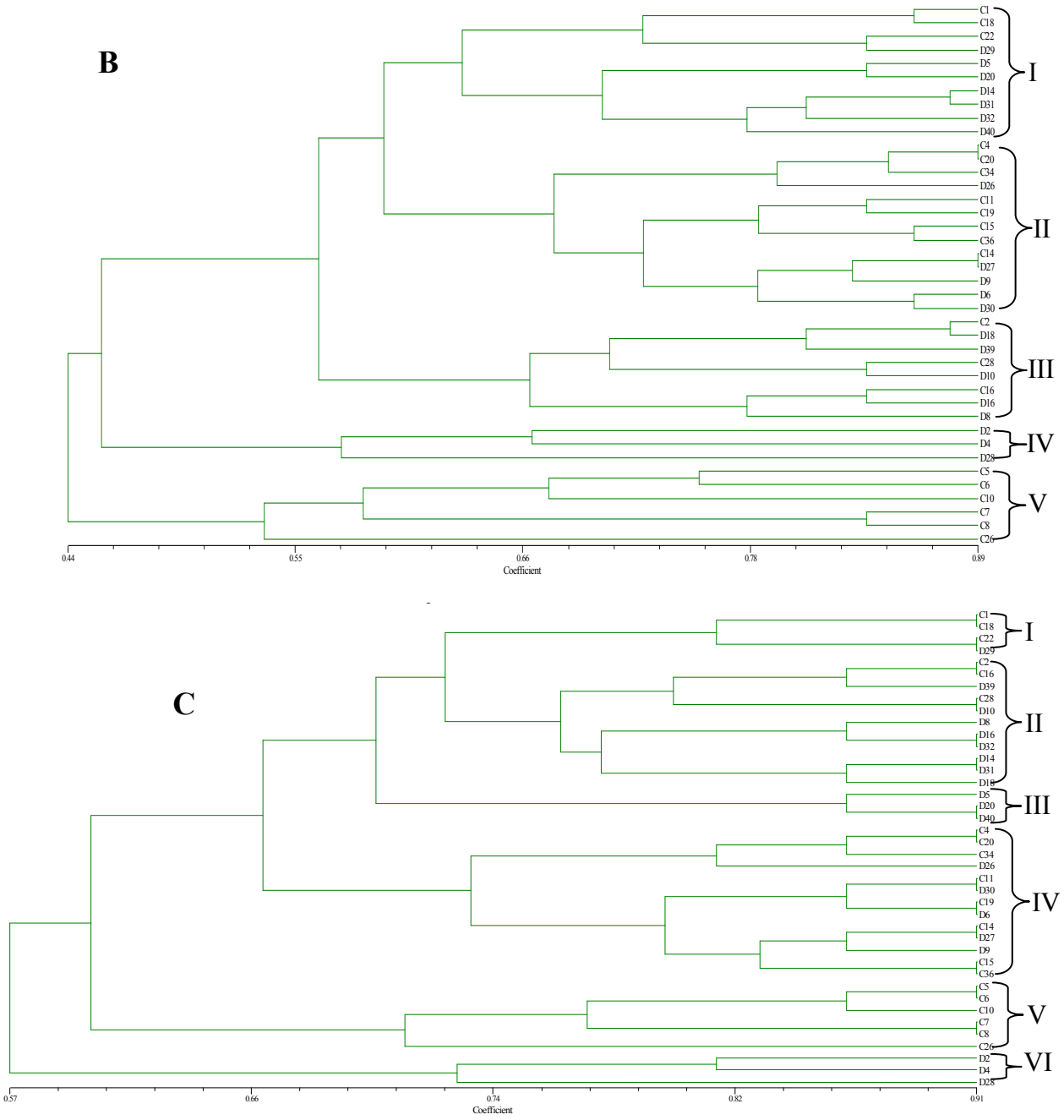


Figure 13: Dice, Jaccard and SM based UPGMA dendrogram for Mildewtrt24 data set

A - Dice, B -Jaccard and C - Simple matching

Table 9: Mildewtrt24 data clusters as seen in the dendrogram based on UPGMA method.

Cluster	Dice	Jaccard	Simple Matching
I	C1,C18,C22,D29	C1,C18,C22,D29,D5,D20, D14,D31,D32,D40	C1,C18,C22,D29
II	C2,D18,D39,C28,D10,C16, D16,D8,D14,D31,D32	C4,C20,C34,D26,C11,C19, C15,C36,C14,D27,D9,D6,D30	C2,C16,D39,C28,D10,D8, D16,D32,D14,D31,D18
III	D5,D20,D40	C2,D18,D39,C28,D10,C16, D16,D8	D5,D20,D40
IV	C4,C20,C34,D26,C11,C19, C15,C36,C14,D27,D9,D6,D30	D2,D4,D28	C4,C20,C34,D26,C11, D30,C19,D6,C14,D27,D9, C15,C36
V	D2,D4,D28	C5,C6,C10,C7,C8,C26	C5,C6,C10,C7,C8,C26
VI	C5,C6,C10,C7,C8,C26		D2,D4,D28

3.4.3 MDS and PCA Results from Powdery Mildew Data

The MDS and PCA results for Mildewtrt13 are shown in Figure 14 while those for Mildewtrt24 are shown in Figure 15. In the Mildewtrt13 data, the same grouping was presented in the MDS plot and the PCA plot of the principal axis 1 against the principal axis 2. However, the PCA plot of the first two axes looked like a transpose of the MDS plot (Figure 14). The two different groups were almost distinctively separated with a few isolates mixing with the other isolates from the other group. Group I consists of A isolates while Group II consists of mingling of B isolates and about 20% of A isolates. The same trend was observed for all the three measures for the PCA plot.

However, for the MDS plot, the percentage of mingling of the isolates from the A group differs for the Jaccard and Simple matching method. They had 15% of A isolates each in addition to the group B isolates while the Dice measure had 20% (Table 13). The first three principal axes accounted for about 80%, 77% and 78% of variation in the data for Dice, Jaccard and Simple matching coefficients respectively (Table 10). On the other hand, in the Mildewtrt24 data, the MDS and PCA plots formed three major groups that were not distinctively separated. For the Dice coefficient, in both MDS and PCA plots, group I consisted of isolates from the D group while group III consisted of isolates from the

C group. Group II had 75% of C isolates and 65% of D isolates in the MDS plot and 60% of C isolates and 85% of D isolates in the PCA plot. For the Jaccard coefficient, group I had 10% of C isolates and 15% of D isolates for both the MDS and PCA plots. Group II had 55% of C isolates and 75% of D isolates for MDS plot and 45% of C isolates and 65% of D isolates for the PCA plot while group III had all C isolates in the MDS plot and 35% of C isolates and 10% of D isolates in the PCA plot (Table 13). For the Simple matching measure, in the MDS plot, group I had 5% of C isolates and 15% of D isolates while in the PCA plot, it had 10% C isolates and 15% D isolates. Group II had 50% of C isolates and 80% of D isolates in the MDS plot but in the PCA plot had 60% of C isolates and 50% of D isolates. However, group III consisted of only C isolates in both the MDS and PCA plots (Figure 15C and D). High variation in the data was accounted for by the first 4 principal components. The Dice, Jaccard and Simple matching had 84%, 81% and 83% respectively (Table 10).

Table 10: Principal components proportion for Mildewtrt13 and Mildewtrt24 data.

Data	Principal component	% of each component	Accumulated percent
Mildewt13- Dice	1	44.79	44.79
	2	19.64	64.43
	3	15.54	79.97
	4	7.38	87.35
	5	5.08	92.43
Mildewt13- Jacc	1	42.06	42.06
	2	19.17	61.23
	3	15.35	76.58
	4	7.48	84.06
	5	5.04	89.10
Mildewt13- SM	1	42.62	42.62
	2	21.24	63.86
	3	14.68	78.54
	4	7.36	85.90
	5	6.49	92.39
Mildewt24- Dice	1	35.55	35.55
	2	19.87	55.42
	3	18.52	73.94
	4	10.51	84.45
	5	7.50	91.95
Mildewt24- Jacc	1	32.90	32.90
	2	19.91	52.81
	3	18.17	70.98
	4	10.41	81.39
	5	8.14	89.53
Mildewt24- SM	1	30.01	30.01
	2	20.12	50.13
	3	18.54	68.67
	4	13.97	82.64
	5	8.62	91.26

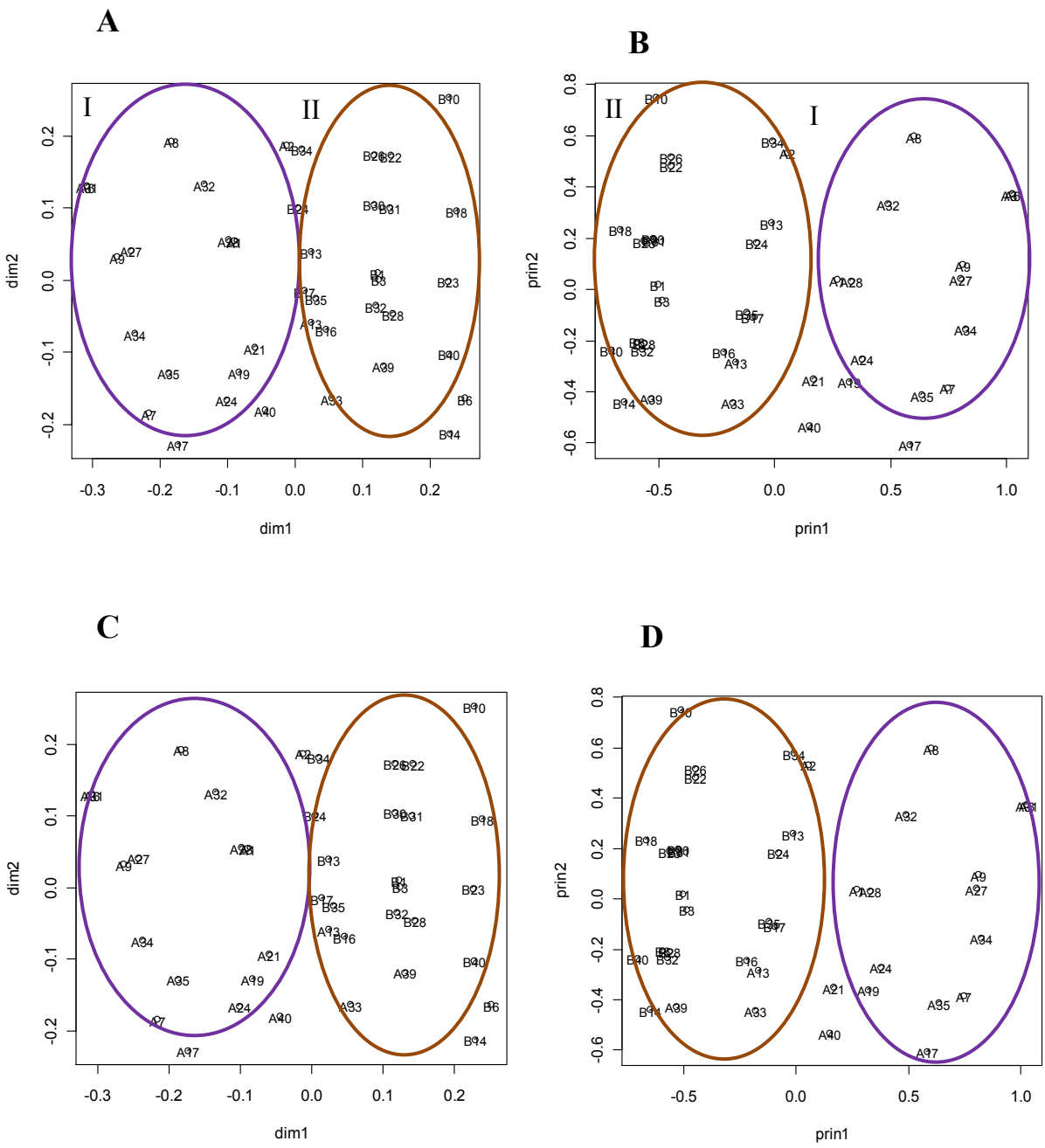


Figure 14: MDS and PCA plots for Jaccard and Simple matching for Mildewtrt13.

A – Jaccard MDS plot, B – Jaccard Prin1 vs Prin2 plot, C – Simple matching MDS plot and D – Simple matching Prin1 vs Prin2 plot.

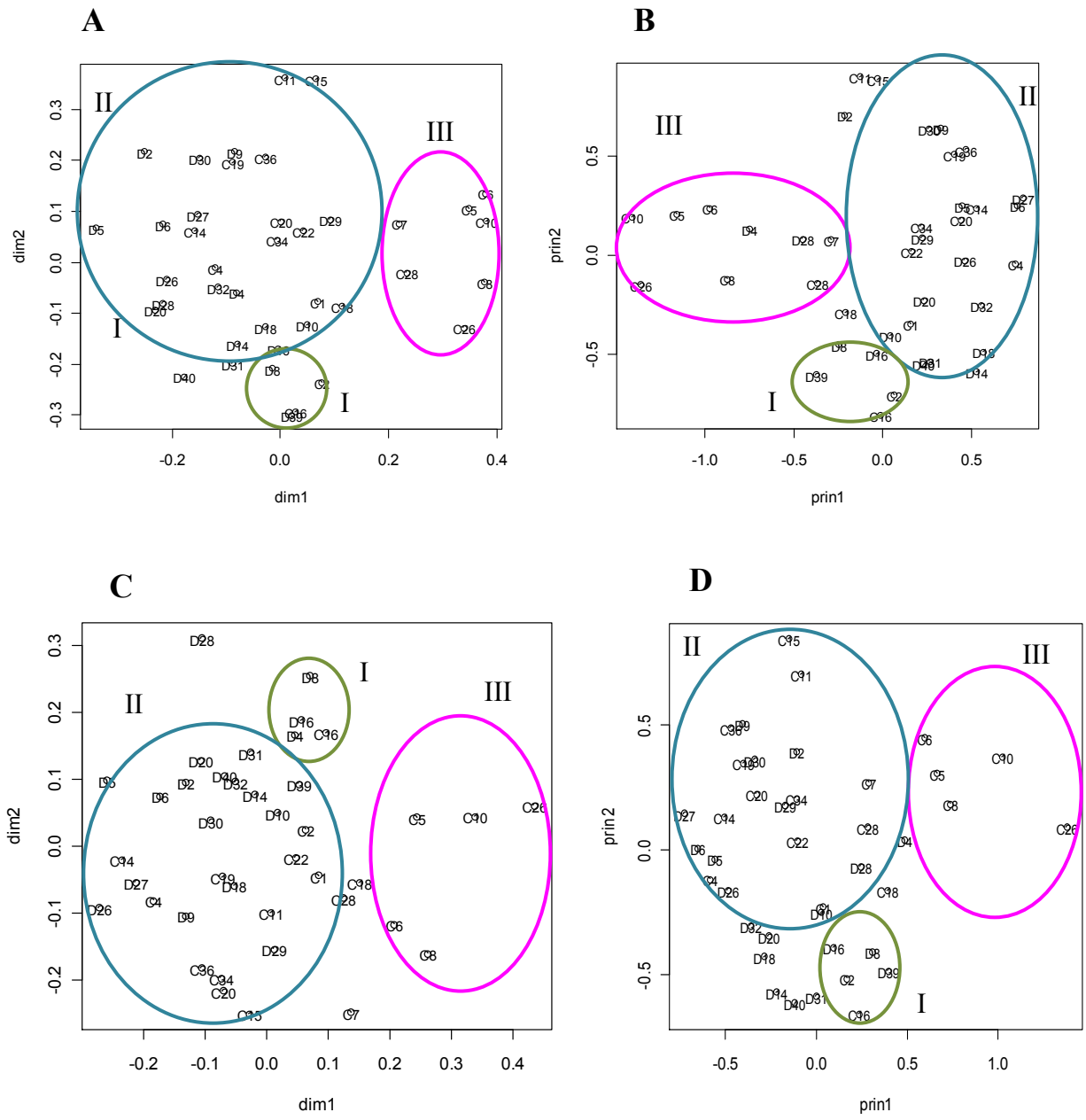


Figure 15: MDS and PCA plots for Jaccard and Simple matching for Mildewtrt24.

A – Jaccard MDS plot, B – Jaccard Prin1 versus Prin2 plot, C – Simple matching MDS plot and D – Simple matching Prin1 versus Prin2 plot.

3.5 AFLP Marker Data on Yam Anthracnose Disease

Ten primers were used to determine the presence of anthracnose disease on 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. Three similarity matrices were constructed from the resulting data; Dice, Jaccard and Simple matching. Dendrograms were constructed using the three similarity coefficients and five CA methods. 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. CFI was used to compare the similarity among the constructed dendrograms for the different similarity measures and CA methods. MDS and PCA were also carried out to compare the groupings.

3.5.1 Dendrogram Results for ACMA, AAMG and AAMO Primers

The dendrograms for the Jaccard and Simple matching similarity coefficients for the UPGMA clustering method are presented in Figures 16 to 18. The Dice dendrogram is also similar to the Jaccard dendrogram in all cases. 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 (Table 11, Table C1-C4, appendix C). The UPGMA produced five clusters and a singleton each for the three coefficients. The WPGMA produced five clusters and a singleton for Dice, five clusters and three singletons for Jaccard and four clusters for Simple matching. The single linkage gave seven clusters and ten singletons for Dice and Jaccard measures and four clusters with twelve singletons for Simple matching. The NJ method resulted in six clusters and three singletons for Dice, six clusters for Jaccard and only two clusters for Simple matching. However, in the AAMG primer data (Table 11, Table C5-C8, appendix C), only the NJ method did not result in identical classifications for Dice and Jaccard while in the AAMO primer data (Table 11, Table C9-C12, appendix C), NJ and WPGMA methods did not give identical classifications for the coefficients. In the AAMG data, UPGMA gave four clusters and two singletons for Dice and Jaccard measures and three clusters for Simple matching. WPGMA also gave three clusters for Simple matching and five clusters for Dice and Jaccard. The complete linkage produced four clusters for all three measures but the clusters from the Simple matching differs from that of Dice and Jaccard. The single linkage

produced two main clusters with eleven singletons for Dice and Jaccard while the Simple matching had five clusters with seven singletons. The NJ method produced three clusters each for Jaccard and Simple matching and three clusters with two singletons for Dice measure. 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. The dendrograms for Jaccard and Simple matching coefficients for UPGMA method and ACMA data are presented in Figure 16, AAMG data are presented in Figure 17 while those for the AAMO data are presented in Figure 18.

In the ACMA and AAMO 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. However, in the AAMG data, only the NJ method gave a different classification for both Dice and Jaccard measures. In all the data, the classification for the Simple matching coefficient was different from that of Dice and Jaccard for all methods.

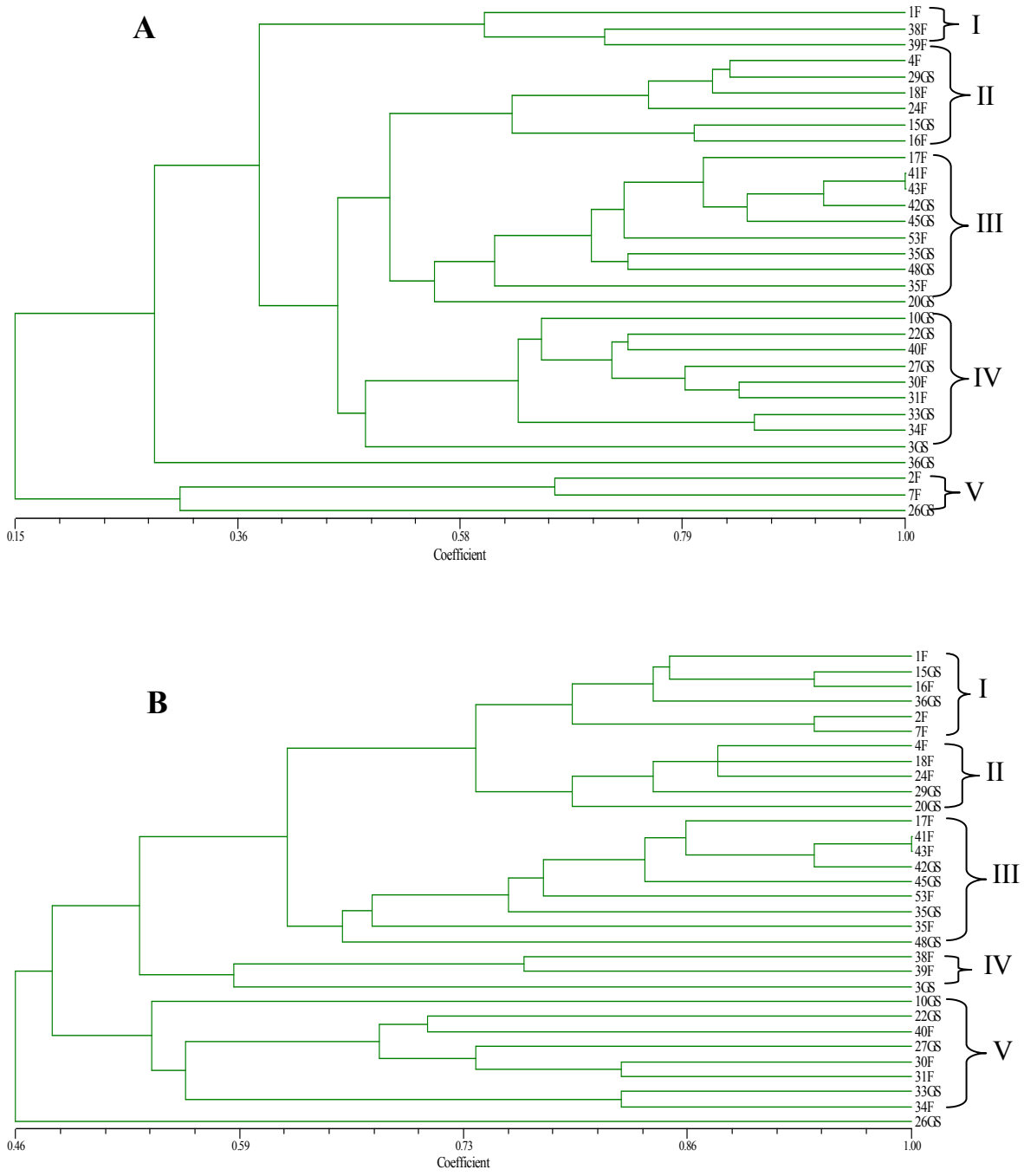


Figure 16: ACMA dendrogram for Jaccard and Simple matching coefficients (UPGMA).

A – Jaccard and B – Simple matching

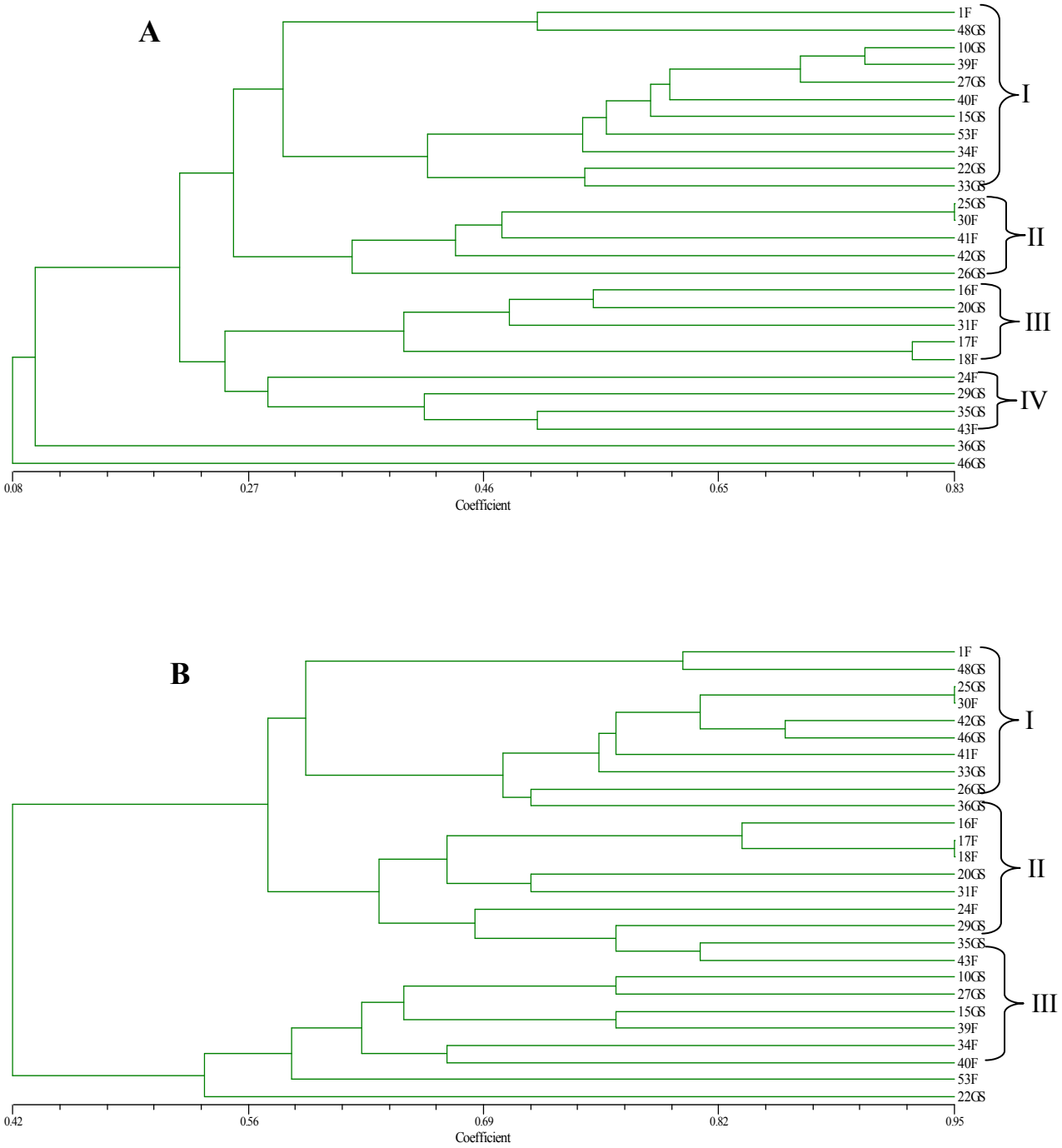


Figure 17: AAMG dendrogram for Jaccard and Simple matching coefficients (UPGMA).

A – Jaccard and B – Simple matching

Table 11: ACMA, AAMG and AAMO data clusters from dendrogram based on the UPGMA method.

Marker	Cluster	Dice and Jaccard	Simple Matching
ACMA	I	1F,38F,39F	1F,15GS,16F,36GS,2F,7F
	II	4F,29GS,18F,24F,15GS,16F	4F,18F,24F,29GS,20GS
	III	17F,41F,43F,42GS,45GS,53F 35GS,48GS,35F,20GS	17F,41F,43F,42GS,45GS,53F,35GS,35F,48GS
	IV	10GS,22GS,40F,27GS,30F, 31F, 33GS,34F,3GS	38F,39F,3GS
	V	2F,7F,26GS	10GS,22GS,40F,27GS,30F,31F,33GS,34F
	Singleton	36GS	26GS
AAMG	I	1F,48GS,10GS,39F,27GS,40F, 15GS,53F,34F,22GS,33GS	1F,48GS,25GS,30F,42GS,46GS,41F,a33GS,26GS,36GS
	II	25GS,30F,41F,42GS,26GS	16F,17F,18F,20GS,31F,24F,29GS, 35GS,43F
	III	16F,20GS,31F,17F,18F	10GS,27GS,15GS,39F,34F,40F,53F,22GS
	IV	24F,29GS,35GS,43F	
	Singleton	36GS,46GS	
AAMO	I	1F,27GS	1F,27GS
	II	10GS,43F,53F,15GS,16F,17F,18F,48GS,20GS,40F,29GS,3GS	10GS,43F,53F,15GS,16F,17F,18F, 48GS,29GS,20GS,40F
	III	26GS,35GS	9GS,39F,52GS,33GS,22GS,36GS,38F
	IV	36GS,38F	24F,41F,42GS,45GS,3GS,8GS
	V	41F,42GS,45GS	26GS,31F,35GS
	VI	9GS,39F,52GS,33GS	
	Singleton	24F,8GS,34F,31F,22GS	34F

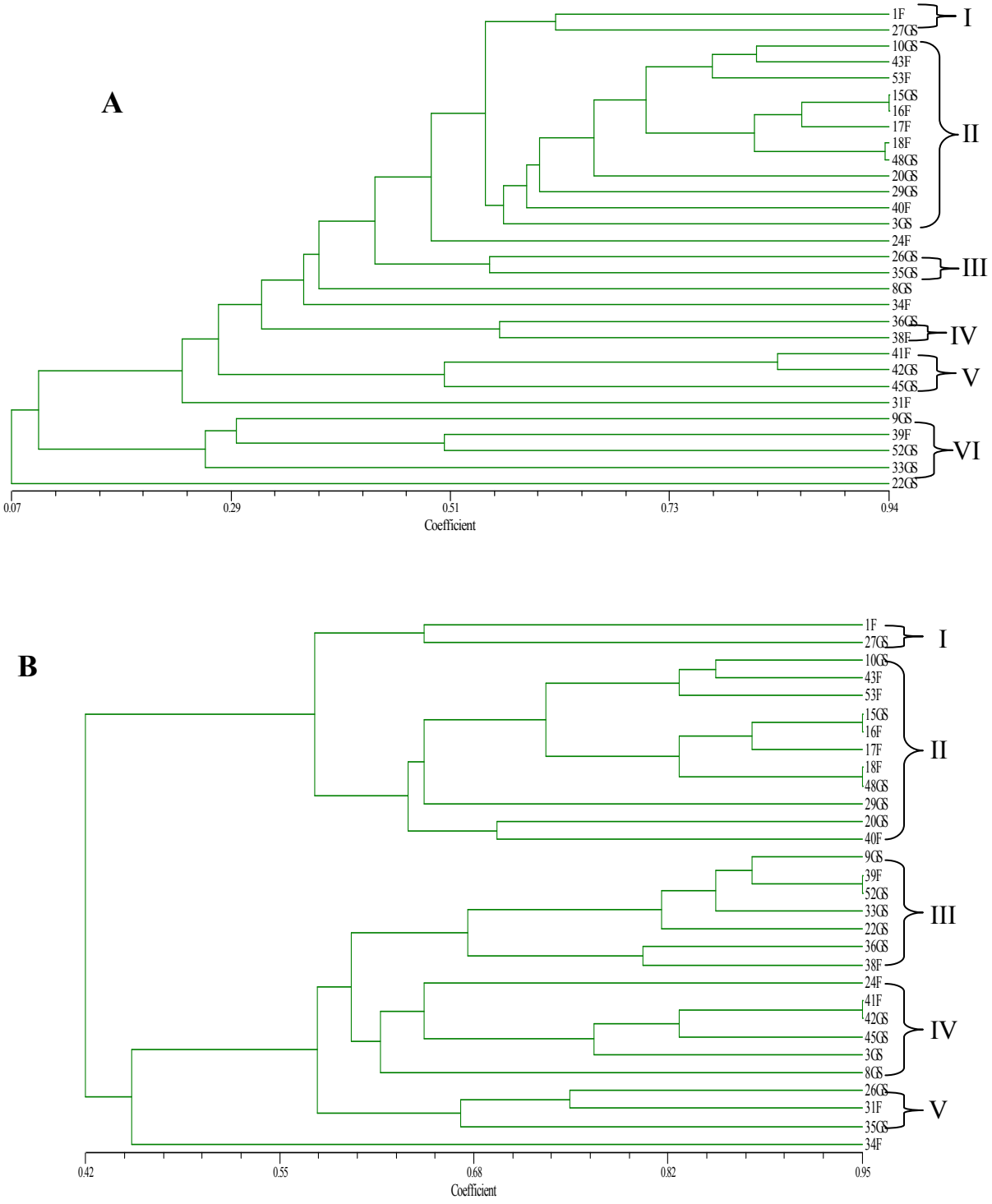


Figure 18: AAMO dendrogram for Jaccard and Simple matching coefficients (UPGMA).

A – Jaccard and B – Simple matching.

3.5.2 MDS and PCA Results

Comparative results of the MDS and PCA for ACMA primer for Jaccard measure are shown in Figure 19A and B, AAMG primer are shown in Figure 20 and AAMO primer are displayed in Figure 21. Four main clusters were observed in the MDS and PCA plots for the ACMA and AAMO data for all similarity measures. From the ACMA data, for the Dice measure, in the MDS plot, group I consisted of 11% pathogens that are from the Forest (F) region and 8% from the Guinea Savannah (GS); group II was made up of 15% GS pathogens and none from the F region; group III had 50% pathogens from F and 38% from GS while group III had 33% from F, 54% from GS. In the PCA plot, members in group I were the same as in the MDS plot, however, group II had 17% of F and 23% of GS; group III had 28% of F and 46% of GS and the group IV was made up of 33% F and 46% GS (Table 14).

For the Jaccard coefficient, in the MDS plot, group I was made up of 17% of pathogens from the F region and none from the GS; group II also had 17% of the F pathogens and 15% from GS; in group III there were 33% from F and 38% from GS and group IV consists of 28% from F and 46% from GS. On the other hand, in the PCA plot, group I had 11% F and 8% GS; group II had 17% F and 23% GS, group III is the same as in the MDS plot and group IV had 33% F, 38% GS. For the Simple matching, in the MDS plot, group I had 17% of F and 8% of GS; group II had 28% of F and 23% of GS, group III had 22% of F and 31% of GS while group IV had 33% of F, 38% of GS. In the PCA plot however, group I had 17% of F and none from GS; group II had 46% of F and none from GS, group III is the same as in the MDS plot and group IV had 33% of F, 46% of GS. In the MDS plot for AAMO data and for the Dice coefficient, group I consists of 8% pathogens from Forest and 13% from GS; group II had 62% of F and 50% of GS; group III had 8% of F and 19% of GS; Group IV had 15% of F, 25% of GS. In the PCA plot however, group I had 15% of F and 6% of GS; groups II and IV are the same as in the MDS plot and group III had 8% of F and 25% of GS.

For the Jaccard coefficient, in the MDS plot, group I was made up of 54% F and 50% of GS; group II had 15% of F and 6% of GS while groups II and IV had the same members as in the MDS plot. For the Simple matching, in the MDS plot, group I had 38% of F and 44% of GS; group II had 15% of F and 6% of GS; group II had 15% of F and 25% of GS and group IV had 23% of F, 25% of GS. On the

other hand, in the PCA plot, group I was made up of 46% of F and 44% of GS; group II had 15% of F and 13% of GS; group III had 8% of F and 25% of GS and group IV was the same as in the MDS plot.

However, in the AAMG, the three measures had different groupings. The Dice had four groupings in the MDS plot and five in the PCA plot. In the MDS plot, group I had 31% of F and 36% of GS; group II had 38% of F and 14% of GS; group III had 8% of F and 15% of GS; group IV had 15% of F and 36% of GS. However, in the PCA plot, group I had 31% of F and 43% of GS; group II had 46% of F and none from GS; group III is the same as in the MDS plot; group IV had 15% of F and 29% of GS and group V had 14% of GS and none from the F region.

The Jaccard measure had four groupings for both MDS and PCA plots while Simple matching had three groupings for both the MDS and PCA plot too. For the Jaccard measure, in the MDS plot, group I had 38% of F and 14% of GS; group II had 31% of F and 29% of GS; group II had 31% of F and 29% of GS; group III had 15% of F and 43% of GS; group IV had 8% of F and 14% of GS. In the PCA plot however, group I had 46% of F and 14% of GS while groups II to IV were the same as in the MDS plot. For the Simple matching, in the MDS plot, group I had 46% of F and 21% of GS, group II had 23% of F and 21% of GS and group III had 23% of F and 50% of GS. On the other hand, in the PCA plot, group I had 54% of F and 21% of GS; group II was the same as in the MDS plot and group III had 15% of F and 50% of GS. There was no clear separation among the pathogens with respect to their geographical locations. This also confirms the groupings from the dendrograms constructed. In the ACMA data, the first three principal axes accounted for 80%, 72% and 80% for the Dice, Jaccard and Simple matching measures respectively (Table 12); in the AAMG data, they accounted for 77%, 66% and 82% in a similar order (Table 12) and in the AAMO data, they accounted for 85%, 77% and 88% (Table 12).

Table 12: Principal component proportion for ACMA, AAMG and AAMO marker data.

Data	Principal component	% of each component	Accumulated %
ACMA- Dice	1	43.26	43.26
	2	21.68	64.94
	3	14.74	79.68
	4	6.47	86.15
	5	4.29	90.44
ACMA- Jaccard	1	35.55	35.55
	2	21.03	56.58
	3	15.13	71.71
	4	6.22	77.93
	5	4.64	82.57
ACMA- SM	1	43.90	43.90
	2	22.37	66.27
	3	13.99	80.26
	4	6.43	86.69
	5	5.45	92.14
AAMG- Dice	1	30.67	30.67
	2	29.90	60.57
	3	16.54	77.11
	4	6.75	83.86
	5	4.17	88.03
AAMG – Jaccard	1	27.90	27.90
	2	24.63	52.53
	3	13.33	65.86
	4	7.62	73.48
	5	4.97	78.45
AAMG – SM	1	52.85	52.85
	2	20.70	73.55
	3	8.42	81.97
	4	4.42	86.40
	5	4.06	90.47
AAMO – Dice	1	68.18	68.18
	2	10.12	78.30
	3	6.57	84.87
	4	4.41	89.29
	5	2.48	91.77
AAMO – Jaccard	1	60.90	60.90
	2	10.61	71.51
	3	5.84	77.35
	4	4.09	81.44
	5	3.39	84.83
AAMO – SM	1	72.68	72.68
	2	10.79	83.47
	3	4.86	88.33
	4	3.13	91.46
	5	2.28	93.74

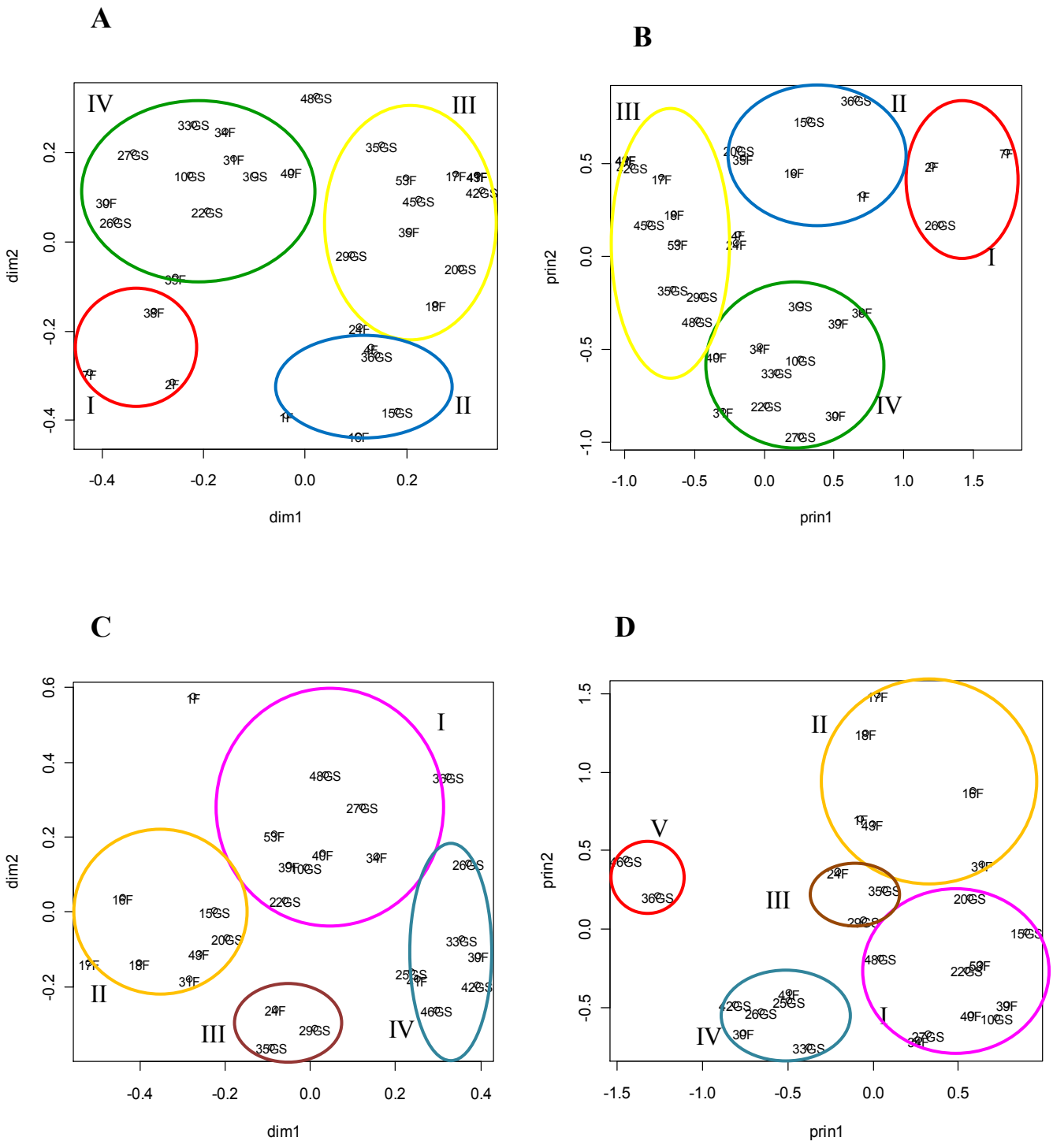


Figure 19: ACMA and AAMG MDS & PCA prin1 versus prin2 plot.

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

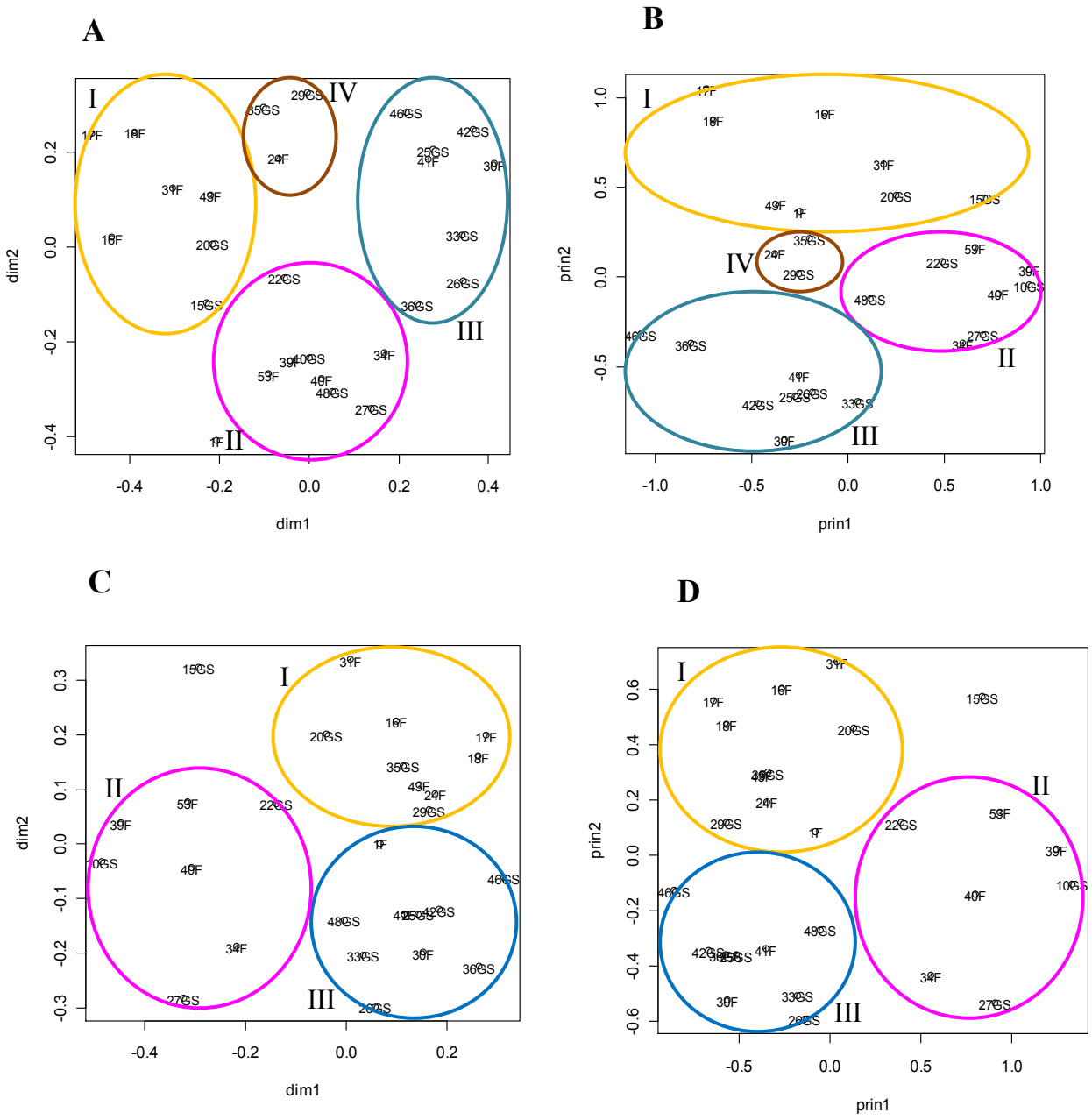


Figure 20: MDS & PCA plots for Jaccard and Simple matching coefficients (AAMG).

A - Jaccard MDS plot, B –Jaccard PCA plot, C – Simple matching MDS plot and D – Simple matching PCA plot.

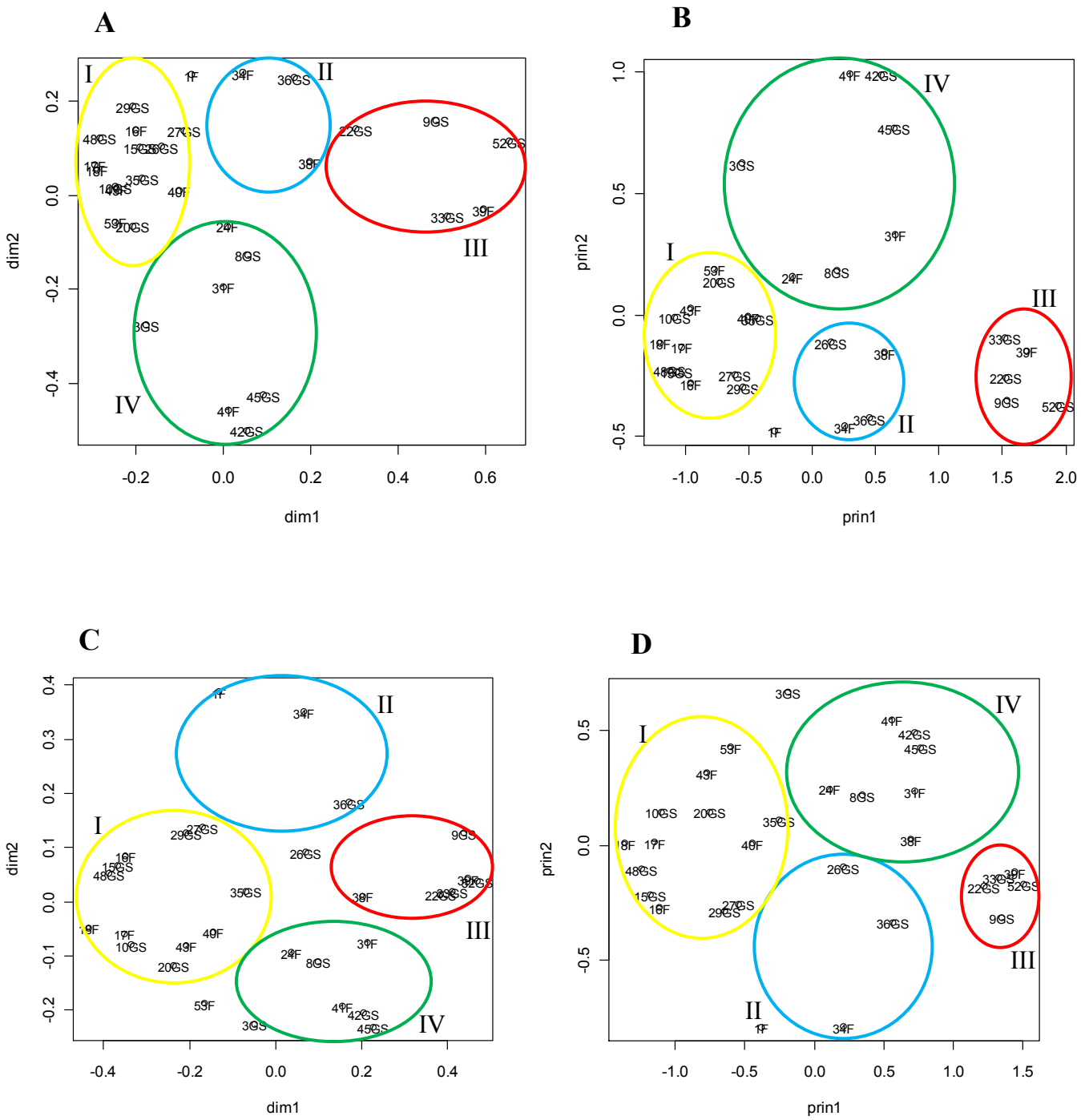


Figure 21: MDS and PCA plots for Jaccard and Simple matching (AAMO).

A - Jaccard MDS plot, B – Jaccard PCA plot, C – Simple matching MDS plot and D – Simple matching PCA plot.

3.5.3 Mingling of Objects between Different Groups

For all the experimental data sets, different degrees of mingling of the objects from the different groups were observed. There was no situation where all the members of a particular group were completely separated from the other. This could be due to some hidden relationship between the objects. It was observed from the dendrograms for the DSS-Plantain data (Figures 9 -10) that there was mingling of all the locations for all the clustering methods. That is, there was no complete separation of the different locations. There was no difference also with the grouping of the different locations into regions (i.e. East, West and Mid-West). A mixture was observed in all the CA methods, although there were cases where locations from the Eastern region formed their own clusters. A mixture of locations from the three regions was also observed (Figure 11).

For the virulence data on powdery mildew from barley with isolates that were susceptible (A) and those that were resistant (B) in the Mildewtrt13 data, the isolates B24, B16 were observed as forming clusters with isolates from group A and isolates A13, A39 formed clusters with isolates from Group B for most of the CA methods. In the other part of the virulence data, with isolates that were susceptible and treated with fungicide (C) and those that were resistant and treated with fungicide (D) in the Mildewtrt24 data, it was observed that there was no clear separation of the isolates from the two groups. They were most of the time, for all CA methods and similarity measures, mixed together in the clusters formed (Figure 15).

In the yam anthracnose AFLP markers data, it was observed for all the three primers data that the pathogens from the Forest and the Guinea Savannah were all mingled together in the clusters formed. For all measures and all CA methods, most of the groups formed were a mixture of pathogens from the two agro-ecological zones. Percentage mingling of the different groups for all experimental data, for both MDS and PCA are shown in Table 13 and Table 14.

Table 13: Percentage mingling of objects in the different groups for plantain and powdery mildew data.

Data	Coefficient	MDS (%)			PCA (%)		
		Group I	Group II	Group III	Group I	Group II	Group III
DSS - Plantain	Dice	CR=56, IM=100,AK=100,DE=30,ED=10,RV=54,AB=80,AN=100	CR=44, ED=70, DE=40, ON=50	ON=33, DE=20, ED=10, RV=18	CR=63,IM=80,AK=100, DE=30,ED=10,RV=64, AB=80, AN=100	CR=44, ED=70, DE=40, ON=50	ON=33, DE=20, ED=10, RV=9
	Jacc	CR=56, IM=100,AK=100,DE=30,ED=20,RV=36,AB=80,AN=100	CR=38, ED=70, DE=40, ON=50	ON=33, DE=20, RV=9	CR=56,IM=100,AK=100, DE=30,ED=10,RV=54, AB=80,AN=100	CR=38, ED=70, DE=40, ON=50	ON=33, DE=20, ED=10, RV=9
	SM	CR=56, IM=80,AK=100,DE=30,ED=10, RV=54,AB=80, AN=100	CR=44, ED=70, DE=40, ON=50	ON=33, DE=10, ED=10, RV=18	CR=63,IM=100,AK=100, DE=30,ED=10,RV=45, AB=60,AN=100	CR=44, ED=70, DE=40, ON=50	ON=33, DE=10, ED=10, RV=9
Mildewtrt 13	Dice	All A isolates	A=20	None	All A isolates	A=20	None
	Jacc	All A isolates	A=15	None	All A isolates	A=20	None
	SM	All A isolates	A=15	None	All A isolates	A=20	None
Mildewtrt 24	Dice	All D isolates	C=75, D=65	All C isolates	All D isolates	C=60, D=85	All C isolates
	Jacc	C=10, D=15	C=55, D=75	All C isolates	C=10, D=15	C=45, D=65	C=35, D=10
	SM	C=5, D=15	C=50, D=80	All C isolates	C=10, D=15	C=60, D=50	All C isolates

Legend

AB=Abia, AK=Akwa-Ibom, AN=Anambra, CR=Cross-River,
DE=Delta, ED=Edo, IM=Imo, ON=Ondo, RV= Rivers,
Jacc=Jaccard, SM=Simple matching
A=Isolates from group A, B=Isolates from group B,
C=Isolates from group C, D=Isolates from group D.

Table 14: Percentage mingling of pathogens in the different agro-ecological zones in the yam data.

Data	Coeffi-cient	MDS (%)					PCA (%)				
		Group I	Group II	Group III	Group IV	Group V	Group I	Group II	Group III	Group IV	Group V
ACMA	Dice	F=11,GS=8	F=0, GS=15	F=50, GS=38	F=33, GS=54	None	F=11, GS=8	F=17, GS=23	F=28, GS=46	F=33, GS=46	None
	Jacc	F=17,GS=0	F=17, GS=15	F=33,GS=38	F=28, GS=46	None	F=11, GS=8	F=17, GS=23	F=33, GS=38	F=33, GS=38	None
	SM	F=17,GS=8	F=28, GS=23	F=22,GS=31	F=33, GS=38	None	F=17, GS=0	F=22, GS=23	F=22, GS=31	F=33, GS=46	None
AAMG	Dice	F=31,GS=36	F=38, GS=14	F=8,GS=15	F=15, GS=36	None	F=31, GS=43	F=46, GS=0	F=8, GS=15	F=15, GS=29	F=0, GS=14
	Jacc	F=38,GS=14	F=31, GS=29	F=15,GS=43	F=8, GS=14	None	F=46, GS=14	F=31, GS=29	F=15, GS=43	F=8, GS=14	None
	SM	F=46,GS=21	F=23, GS=21	F=23,GS=50	None	None	F=54, GS=21	F=23, GS=21	F=15, GS=50	None	None
AAMO	Dice	F=8,GS=13	F=62, GS=50	F=8,GS=19	F=15, GS=25	None	F=15, GS=6	F=62, GS=50	F=8, GS=25	F=15, GS=25	None
	Jacc	F=54,GS=50	F=15, GS=6	F=8,GS=25	F=23, GS=25	None	F=54, GS=44	F=15, GS=13	F=8, GS=25	F=23, GS=25	None
	SM	F=38,GS=44	F=15, GS=6	F=15,GS=25	F=23, GS=25	None	F=46, GS=44	F=15, GS=13	F=8, GS=25	F=31, GS=25	None

Legend

F=Forest, GS=Guinea Savannah, Jacc=Jaccard,

SM=Simple matching

3.5.4 Consensus Fork Index values

The CFI values for all experimental data are presented in Table 15. 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 gave a range of 0.93 and 1 while for the NJ method 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. The single linkage method is well known for producing a long chain dendrogram with lots of singletons, this was 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, out of the six experimental data sets analysed, the UPGMA method produced the highest number of identical trees with 4 occurrences of CFI value of 1, followed by the complete linkage method with three occurrences and the WPGMA with two occurrences (Table 15). The single linkage and the NJ methods had no occurrences of CFI value of 1. The mean CFI plot for all experimental data is shown in Appendix D.

The CFI values comparing the Dice and Simple matching dendrograms (Table 15B) however, were very low. For the UPGMA method, the range was between 0.32 and 0.66; for WPGMA, 0.16 and 0.55; complete linkage 0.24 and 0.64; single linkage, 0.03 and 0.24 and NJ, 0.41 and 0.63. The same trend was observed in the CFI values for all the clustering methods, with UPGMA and complete linkage methods having the highest values. Also, the range for the CFI values for comparing the Jaccard and Simple matching dendrograms (Table 15C) for UPGMA was between 0.32 and 0.55. The ranges for WPGMA, single and complete linkage methods were the same with Dice and Simple matching while for NJ, the range was between 0.43 and 0.65. These CFI values for dendrograms between Dice and Simple matching as well as Jaccard and Simple matching also confirms the 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.

Table 15: CFI summary for different methods of CA for all experimental data.

A

Source	No of OTUs/Iso	No of differentials	UPGMA	WPGMA	SINGLE	COMPLETE	NJ
DSS-Plantain	70	23	1.00	0.66	0.63	1.00	0.54
Mildewtrt13	40	22	1.00	1.00	0.29	1.00	0.39
Mildewtrt24	40	22	0.89	0.89	0.21	1.00	0.55
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							
DSS-Plantain	70	23	0.40	0.40	0.24	0.41	0.41
Mildewtrt13	40	22	0.32	0.29	0.03	0.26	0.34
Mildewtrt24	40	22	0.66	0.55	0.03	0.45	0.63
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							
DSS-Plantain	70	23	0.40	0.44	0.24	0.41	0.65
Mildewtrt13	40	22	0.32	0.29	0.03	0.26	0.47
Mildewtrt24	40	22	0.55	0.55	0.03	0.45	0.55
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.

3.5.6 Results from Other Methods of Comparing Topology

Results for the Pearson correlation coefficients between the cophenetic distances, the node counts from the dendrograms and a combination of cophenetic distances and node counts for Dice and Jaccard coefficients and the different methods of cluster analysis for all experimental data are presented in Table 16. The Spearman correlation coefficient equivalents are presented in Table 17. The Pearson correlation coefficient for the cophenetic distances for both Dice and Jaccard

measures, for all experimental data for the UPGMA, WPGMA, single linkage and complete linkage methods revealed values ranging between 0.5815 and 0.9996 (Table 16A).

The DSS-plantain data had a correlation coefficient of 0.9893 for the UPGMA method, 0.8231 for WPGMA method, 0.9967 for single linkage method and 0.9878 for complete linkage method. Though the single linkage had the highest correlation coefficient value, the CFI for this measure was 0.63 (Table 15A). In Mildewtrt13 data, UPGMA had a correlation coefficient value of 0.9791, WPGMA had 0.9974, single linkage had 0.9996 and complete had 0.9916. Here also, the single linkage had the highest value; however the CFI value was only 0.29. In Mildewtrt24 data, UPGMA had a correlation coefficient value of 0.9933, WPGMA had 0.9101, single linkage had 0.9995 and complete linkage had 0.9926. Here also, the single linkage had the highest value; however the CFI value was 0.21 (Table 15A).

In the yam anthracnose disease data, for AAMG primer data, single linkage had the highest correlation coefficient value (0.9586) while the WPGMA had the lowest value (0.5815). In contrast however, the WPGMA had a CFI value of 1 while the single linkage had a CFI value of 0.72. For the AAMO primer, the single linkage also had the highest correlation coefficient value of 0.9928 while the WPGMA had the lowest value (0.9857). However, the CFI value for the WPGMA was 0.90 and that of single linkage was 0.77. These results confirm that high correlation coefficient does not necessarily imply similarity with respect to topology of trees. For the ACMA primer, the highest correlation coefficient value was for single linkage (0.9964) and the lowest was for WPGMA (0.9737). The CFI values for WPGMA and single linkage methods were 0.64 and 0.93 respectively. These results corroborate one another (Table 15A). However, the Spearman correlation coefficients for all experimental data and for UPGMA, WPGMA, single linkage and complete linkage methods presented a different trend from the Pearson Correlation. The single and complete linkages gave a correlation coefficient of 1 for all the data while the UPGMA gave values ranging between 0.8749 and 1 and WPGMA gave values ranging between 0.7198 and 1 (Table 17).

Table 16: Pearson correlation coefficients for Dice and Jaccard for different CA methods.

A

Source	No of OTUs/Iso	No of differentials	UPGMA	WPGMA	Single linkage	Complete linkage
DSS-Plantain	70	23	0.9893	0.8231	0.9967	0.9878
Mildewtrt13	40	22	0.9791	0.9974	0.9996	0.9916
Mildewtrt24	40	22	0.9933	0.9101	0.9995	0.9926
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

B

DSS-Plantain	70	23	0.9973	0.8418	1.0	1.0
Mildewtrt13	40	22	0.9454	1.0	1.0	1.0
Mildewtrt24	40	22	0.9801	0.8681	1.0	1.0
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

C

DSS-Plantain	70	23	0.9992	0.9625	0.9999	0.9999
Mildewtrt13	40	22	0.9887	0.9999	0.9999	0.9999
Mildewtrt24	40	22	0.9963	0.9780	0.9999	0.9999
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 17: Spearman correlation coefficients for Dice and Jaccard for different CA methods.

A

Source	No of OTUs/Iso	No of differentials	UPGMA	WPGMA	Single linkage	Complete linkage
DSS-Plantain	70	23	0.9988	0.7198	1.0	1.0
Mildewtrt13	40	22	0.9920	1.0	1.0	1.0
Mildewtrt24	40	22	0.9533	0.8303	1.0	1.0
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
B						
DSS-Plantain	70	23	0.9973	0.8214	1.0	1.0
Mildewtrt13	40	22	0.9318	1.0	1.0	1.0
Mildewtrt24	40	22	0.9726	0.7790	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
C						
DSS-Plantain	70	23	0.9995	0.9448	1.0	1.0
Mildewtrt13	40	22	0.9906	1.0	1.0	1.0
Mildewtrt24	40	22	0.9910	0.9523	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.

3.5.7 Correlation Coefficients between Cophenetic Distances and Original Distances

The Pearson correlation coefficients calculated between cophenetic distances and original distances for the three similarity coefficients and five CA methods are presented in Table 18. The UPGMA consistently gave the highest value out of all the methods and for all measures and all experimental data. In the Diagnostic Survey Sample data on plantain, the Jaccard measure had the highest correlation coefficient of 0.70 with the UPGMA method while the Dice measure had the lowest correlation coefficient value of 0.33 with the NJ method. The UPGMA had the highest value out of all the methods and for all measures, followed by the WPGMA, complete linkage, single linkage and NJ. However, for the Simple matching method, the NJ correlation coefficient was higher than the single linkage.

The Jaccard measure had the highest correlation coefficient of 0.61 with UPGMA method and Simple matching had the smallest correlation coefficient of 0.10 with the single linkage method in the powdery mildew data (Mildewtrt13). Rank of the highest correlation coefficient goes in the order UPGMA, WPGMA and complete linkage. However, the order between the single linkage and the NJ methods were not the same for all three measures. For the Dice and Simple matching measures, the NJ values were higher than the single linkage while for the Jaccard measure the single linkage value was higher than the NJ value. A different scenario was observed in ranking order in the powdery mildew data (Mildewtrt24), the order of the rank was UPGMA, WPGMA, complete linkage, single linkage for both Dice and Jaccard measures while for the Simple matching, the order was UPGMA, WPGMA, complete linkage, NJ and single linkage.

A different ranking order which was not the same as what was observed in the other three data sets was observed in the yam anthracnose marker data. The ranking order was the same for all three similarity measures in AAMG and AAMO primers. The order was UPGMA, WPGMA, single linkage, complete linkage and NJ methods. However, in the ACMA primer data the ranking was the same for Dice and Simple matching measures as in the other two primer data, but the order was slightly different for Jaccard measure. It was UPGMA, WPGMA, NJ, single linkage, and complete linkage methods. It was observed that the correlation coefficients for the AAMO primer data were the highest, followed by ACMA data and lastly by AAMG. It was also observed that the correlation coefficient for the Jaccard measure for NJ method in the AAMO data was negative. All the observations about the NJ method suggest that it could be unstable.

Table 18: Correlation coefficients from cophenetic matrices and original distances for all experimental data.

Data	Method/Similarity	Dice	Jaccard	SM
DSS-Plantain	UPGMA	0.66	0.70	0.59
	WPGMA	0.64	0.66	0.59
	Single linkage	0.52	0.55	0.43
	Complete linkage	0.53	0.56	0.55
	NJ	0.33	0.35	0.56
Mildewtrt13	UPGMA	0.58	0.61	0.58
	WPGMA	0.58	0.61	0.57
	Single linkage	0.33	0.35	0.10
	Complete linkage	0.50	0.54	0.54
	NJ	0.46	0.30	0.38
Mildewtrt24	UPGMA	0.66	0.68	0.60
	WPGMA	0.64	0.68	0.57
	Single linkage	0.56	0.56	0.38
	Complete linkage	0.57	0.59	0.47
	NJ	0.34	0.43	0.43
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

4 Discussion

Prerequisite for carrying out CA is the choice of a similarity measure and clustering methods and any combination of these two is possible. However, this choice depends on the experimental situation or questions to be answered by the researcher. The objective nature of cluster analysis is compromised by the subjective choices of clustering method and similarity measures keeping in mind that both the method and the similarity measure affect the outcome of the analysis (Jackson et al., 1989; Legendre and Legendre, 1983; Orloci, 1978; Pielou, 1984). Different combinations of the measure and clustering method may lead to very different results. Therefore in this study, the choice of appropriate similarity measure and clustering method combination for specific situations using binary data was investigated. It's been widely circulated that Dice and Jaccard measures usually result in similar classification. This was the motivation for our study initially. However, a third measure was introduced in analysing the experimental data after the initial results from the simulated data were obtained. The experimental data were binary marker data from anthracnose disease in yam, powdery mildew isolates and plantain diagnostic survey samples on plantain production constraints. The results from the simulated and different experimental data sets showed that there are various levels of interaction between the different clustering methods and the similarity coefficients. The discussion is based on the results from the simulated data and the results for the different methods of comparison for the experimental data in relation to other relevant studies with an outlook towards further investigation.

4.1 Comparing the Dendrograms by Visual Inspection and CFI

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 corroborating the similarity differences between the three measures (Duarte et al., 1999; Jackson et al., 1989; Meyer et al., 2004). These differences are revealed in the alterations in the levels in which the individuals are clustered. 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; Duarte et al., 1999; Jackson et al., 1989; 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 all the experimental data sets used in this study. However, none of these studies was specific for powdery mildew or yam anthracnose isolates. For a given data set, the calculated values of the Jaccard similarity coefficient are always smaller than those calculated using the Dice similarity coefficient. In contrast however, the calculated values of the Dice similarity coefficient may be greater or smaller than the calculated values of the Simple matching coefficient based on whether the number of positions with shared bands or attributes "a" is less or greater than the number of positions with shared absence of band or attributes "d", respectively (Dalirsefat et al., 2009). This is also clearly reflected in the definition of the different similarity coefficients as seen in Table 1.

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. As observed in the simulated data, there were cases where the Dice and Jaccard dendrograms constructed using the UPGMA method were not similar (Figures 5, 6 and 7). 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). However, some of our results showed that there could be some exceptions, where we have perfect separation in trees constructed using the Jaccard coefficient and mixture of objects in those constructed using Dice coefficient (Figure 5). In some cases, complete separation of members of a group was observed in one of the coefficients and in the other coefficient, they were mingled together. On the other hand, there were cases where in both coefficients, there was mingling of objects within the groups. In the Jaccard dendrogram in Figure 5A, there was mingling among the members of the groups while in the Dice dendrogram (Figure 5B), there was perfect separation. In Figure 5D, there was mingling in the Dice dendrogram and perfect separation in the Jaccard dendrogram (Figure 5C). On the other hand, in Figures 6 and 7, there was mingling in both the Dice and Jaccard dendrograms suggesting that the two measures could be used interchangeably with none being superior to the other.

One of the criteria for choosing the most appropriate coefficient of similarity depend 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, we would also recommend that the Dice or Jaccard similarity coefficient 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).

Another important observation in the simulated data was with the number of columns of the data analysed that produced mingling. Data without in-built grouping that had less than 100 columns did not show any mingling among the members of the different groups. However, in the experimental data, all the samples had less than 100 columns and there were mingling among the members. This suggests that this observation is not always consistent and that the incidence of mingling does not depend on the dimension of the data, it could depend on some other factors which might warrant further investigation. In the simulated data with less than 100 columns, it was also observed that the percentage of the samples with CFI values less than one are lower compared to the samples with columns above 100 also suggesting that the longer the number of columns, the higher the possibility of less identical trees.

The different locations (states) where the surveys were carried out were grouped into different clusters by the different CA methods with the single linkage producing the smallest size of clusters with a lot of singletons in the DSS plantain data. This suggests that the grouping produced by the single linkage method is not the most appropriate because of the singletons as most of the objects could not be grouped together. There were situations where some CA methods produced identical classifications for both Dice and Jaccard measures while the WPGMA and NJ methods produced

different classifications for the measures. This could further strengthen the similarity that is usually assumed or expected between these two measures. It was observed that there was a mixture of the states even in the regional groupings which may be attributed to the closeness of some of the locations of the farms within a state. Some of the states are closely located (Figure 3) suggesting that the locations have no effect on factors affecting the root health assessment, plant growth as well as disease evaluation with respect to plantain production in Nigeria. However, in the powdery mildew data involving treatments 1 and treatment 3 (Mildewtrt13), only the NJ method produced different classifications for Dice and Jaccard measures while in the data for treatment 2 and treatment 4 (Mildewtrt24), the complete and single linkage methods produced identical classifications. Introduction of fungicides to the resistant and susceptible isolates that constituted the data in Mildewtrt24 may be an explanation for this observation. It was also noted that the Mildewtrt13 data produced two main clusters in the MDS and PCA plots for the three similarity measures. One of the clusters consisted of all 'A' isolates while the second cluster consisted of a mingling of the members of the two isolate groups. On the other hand, Mildewtrt24 produced three main clusters that mostly consisted of mingled isolates. The difference in the number of clusters formed for the two data sets may not be unconnected with the application of fungicide to the Mildewtrt24 data confirming that the fungicide did have an effect on the resulting classification.

In the anthracnose disease markers data, the three primers data also produced different classifications. There was a mixture of the pathogens from the different agro-ecological zones 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; Dalirsefat et al., 2009; Duarte

et al., 1999; 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 (Tables 15). The complete linkage and the UPGMA methods tend to produce trees that are somehow similar, which was also reflected in the CFI values (Tables 15). In general, out of the six experimental data sets analysed, the UPGMA method produce 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. A plot of the mean CFI for all experimental data is shown in App. D1; this plot revealed that the UPGMA and the complete linkage had the same value for the mean CFI across all data sets. This observation was also reflected in the classifications for these two measures, the UPGMA and the complete linkage results were quite close in some of the resulting classifications.

4.2 Correlation Coefficients for Other Methods of Comparing Topology

4.2.1 Correlation Coefficients for Cophenetic Distances

The Pearson correlation coefficient for the cophenetic distances for both Dice and Jaccard measures, for all experimental data for the UPGMA, WPGMA, single linkage and complete linkage methods revealed a reasonably high level of correlation between these two measures. These results support what was reported in earlier studies stating the high correlation between the Dice and Jaccard measures (Duarte et al., 1999; Meyer et al., 2004). In all the experimental data, the cophenetic correlation coefficient for the single linkage method was close to one, even though the single linkage had the highest correlation coefficient value, the CFI for this measure was quite low suggesting that high correlation does not imply similarity in terms of topology of trees. In the yam anthracnose disease data for AAMG primer data, single linkage had the highest value (0.9586) while the WPGMA had the lowest value (0.5815). In contrast however, the WPGMA had a CFI value of 1 while the single linkage had a CFI value of 0.72. For the AAMO primer, the single linkage also had the highest correlation coefficient value of 0.9928 while the WPGMA had the lowest value (0.9857). However, the CFI value for the WPGMA was 0.90 and that of single linkage was 0.77. These results confirm that high correlation coefficient does not necessarily imply similar topology of trees. For the ACMA primer, the highest correlation coefficient value was for single linkage (0.9964) and the lowest was for WPGMA (0.9737). The CFI values for WPGMA and single linkage methods were 0.64 and 0.93 respectively, these results corroborate one another. A plot of the mean Pearson correlation coefficients for all experimental data sets is shown in App. D2. This plot also revealed that the single linkage method had the highest mean correlation coefficient value across all the data sets. Even though the single linkage had the highest Pearson correlation coefficient value, it was the method that produced the least similar topology of the trees compared.

However, the Spearman correlation coefficients for all experimental data and for UPGMA, WPGMA, single linkage and complete linkage methods presented a different trend from the results obtained from the Pearson Correlation. The single and complete linkages gave a correlation coefficient of 1 for all the data while the UPGMA and WPGMA methods did not. This suggests that the Dice and Jaccard coefficients are monotonically related for the Complete and Single linkage methods while this is not true for the UPGMA and WPGMA methods. A plot of the Spearman correlation coefficients for all experimental data is shown in App.D3; this plot revealed that the single and complete linkages always had a value of 1 for all experimental data, followed by the UPGMA and the WPGMA with the lowest correlation value.

4.2.2 Correlation Coefficients for Node Counts for Dice and Jaccard Measures

In order to develop another index that could be used to compare topology of trees, node counts were generated from node to node for all dendrograms and correlation coefficients were calculated. The Pearson and Spearman correlation coefficients for node counts generated from the individual dendrograms for Dice and Jaccard measures for all experimental data and for the complete linkage and single linkage revealed that the values for both methods were 1 while those for UPGMA and WPGMA were not. This could also suggest that the node counts for the two measures are monotonically related. However, it could also be that because the node counts data were roughly elliptically distributed and there were no prominent outliers, therefore the two correlation coefficients gave similar values.

4.2.3 Correlation Coefficients for Combination of Cophenetic Distances and Node Counts for Dice and Jaccard Measures

In order to look for another index that could properly explain the topology of the trees, correlation coefficients for the combination of node counts and cophenetic distances were calculated. For all the experimental data, the values for the Pearson correlation coefficient were close to one suggesting that the combination of the node counts and cophenetic distances revealed a high level of relationship between the two measures. However the Spearman correlation coefficients for single and complete linkage methods were seen to be consistent with the value 1 for all data sets and for all comparisons. This again suggests that the Dice and Jaccard values for these functions are perfectly monotonically related for these two clustering methods.

4.3 Correlation Coefficient between Cophenetic Distances and Original Distances.

To check the goodness of fit of a cluster analysis with the associated similarity/distance matrix, cophenetic correlation coefficients (Sokal and Rohlf, 1962) were calculated for all data sets. The Pearson correlation coefficients calculated between cophenetic distances and original distances for the three similarity coefficients and five CA methods suggests UPGMA gives consistent results. The UPGMA had the highest value out of all the methods and for all measures and all experimental data. This result is similar to what was previously reported (Koopman et al., 2001; Sesli and Yegenoglu, 2010). The authors while comparing the results of clustering method/similarity coefficient reported that the Jaccard coefficient with the UPGMA and Dice coefficient with the UPGMA method respectively had the highest correlation coefficient and are therefore considered as a convenient combination for detecting the genetic relationship for AFLP marker data set from *Lactuca, S L* Species and between cultivated olives respectively. However, from all the experimental data considered in this study ranging from data on AFLP markers to isolates from powdery mildew and diagnostic survey samples on plantain based on locations, the Jaccard coefficient with the UPGMA method gave the highest correlation coefficient. This could in part explain why many researchers use this combination of similarity coefficient and clustering method, albeit many of them do not give reasons for their choices. Ogunjobi et al. (2007, 2011) in studies on Cassava bacteria blight in Nigeria, Dalirsefat et al. (2009) in a study on AFLP markers in silkworm in Iran and Kumar et al. (2010) in a study on red rot in Indian sugarcane all used the Jaccard coefficient with the UPGMA method without any reason to justify their choice. Beharav et al (2010) reported a study on 36 randomly screened studies (1998-2008) that gave percentages of cases where the Dice, Jaccard and Simple matching coefficients were used. It was discovered also in their study that the Jaccard coefficient was more frequently used. Out of the 44 similarity coefficients used, the breakdown was given as follows: 25 (56.8%) used the Jaccard coefficient, 13 (29.6%) the Dice coefficient and 6 (13.6%) the Simple matching coefficient, with the Jaccard coefficient taking the lead.

For the Diagnostic Survey Sample data on Plantain, the Jaccard measure had the highest correlation coefficient of 0.70 with the UPGMA method while the Dice measure had the lowest correlation coefficient value of 0.33 with the NJ method. The rank of the clustering methods with respect to the correlation coefficients for this data was UPGMA, WPGMA complete linkage, single linkage and NJ. Therefore for data of this structure, one would suggest the combination of similarity and

clustering method of the Jaccard coefficient with the UPGMA method while it will not be recommended to use the combination of Dice with the NJ method.

In the powdery mildew data (Mildewtrt13) also, the Jaccard with the UPGMA was observed to be the best combination while the Simple matching with the single linkage was observed to be the least. The rank of the correlation values for the different methods here was also in the order UPGMA, WPGMA, complete linkage, single linkage and NJ for the Jaccard measure and UPGMA, WPGMA, complete linkage, NJ and single linkage for Dice and Simple matching measures. However, in the Mildewtrt24 data, the combination with the lowest correlation coefficient was the Dice coefficient with the NJ method and that with the highest correlation coefficient was also Jaccard with the UPGMA. It is suspected that the introduction of the fungicide to the isolates could be responsible for this change. The rank of the clustering methods was UPGMA, WPGMA, complete linkage, single linkage and NJ for Dice and Jaccard measures and UPGMA, WPGMA complete linkage, NJ and single linkage for Simple matching measure.

In the yam anthracnose marker data, the three primer data had a different order of ranking which was not the same as what was observed in the other three data sets. The combination with the highest coefficient for all the three primers was the Jaccard with the UPGMA method while the combination with the lowest correlation value was Jaccard with the NJ method in both AAMG and AAMO primer data and Simple matching with NJ in ACMA data. The ranking for the correlation coefficients for the AAMG (for all three measures) and AAMO (Dice and Jaccard) was UPGMA, WPGMA, single linkage, complete linkage and NJ while for AAMO (Simple Matching) was UPGMA, complete linkage, WPGMA, NJ and single linkage. The ranking for the ACMA data on the other hand differs for each similarity measure. For Dice measure, it was UPGMA, WPGMA, single linkage, complete linkage and NJ; for Jaccard measure, it was UPGMA, WPGMA, NJ, single linkage, complete linkage and for the Simple matching measure, it was UPGMA, single linkage, complete linkage, WPGMA and NJ. It is interesting to note that the correlation coefficients in general for the AAMO primer data were the highest, followed by ACMA data and lastly by AAMG. This further confirms the fact that different results are obtained from the same data using different primer combinations. It was also observed in this study that the correlation coefficient for the Jaccard measure for NJ method in the AAMG data was negative. All these suggest that the NJ method could be unstable.

4.4 Classification Using MDS and PCA

In the plantain data, three major groupings were found for all the three coefficients. The MDS plot and the PCA plot of the first two axes gave the same grouping. This comparison of the results provided by the bi-dimensional graphical dispersion of the different locations showed a lot of mixing among them. This confirms the groupings produced in the dendrograms for the DSS Plantain data, suggesting that the locations do not have strong effect on factors affecting the production of plantain. The first three principal axes accounted for about 65%, 60% and 63% of variation in the data for the Dice, Jaccard and Simple matching coefficients respectively. A plot of higher PCA axes also revealed a higher level of mixing among the locations as expected because the higher the PCA axes, the lower the percentage of variation accounted for in the data.

The Mildewtrt13 data produced two main groupings while the Mildewtrt24 data produced three main groupings. The isolates in the former data were well separated while only two out of the three groupings in the later data were well separated. In the Mildewtrt13 data, the first three principal components accounted for approximately 80%, 77% and 78% of the variation in the Dice, Jaccard and Simple matching coefficients respectively. On the other hand in the Mildewtrt24 data, the first four principal components accounted for approximately 84%, 81% and 83% of the variation in the data for the Dice, Jaccard and Simple matching respectively. The introduction of the fungicides to the isolates in the Mildewtrt24 data could be responsible for this change in the percentage variation and groupings suggesting that the fungicide did have an effect on the powdery mildew isolates which was reflected in the classification. The mingling observed among the isolates in the plot of the higher PCA axes also gives the idea that the first two axes depict the true structure of the data while there is a mixture of the isolates for the higher axes.

In the yam anthracnose with AFLP marker data, in the ACMA data, the first three principal components accounted for 80%, 72% and 80% of the variation in the data for the Dice, Jaccard and Simple matching coefficients respectively. In the AAMG data however, the first principal components accounted for 77%, 66% and 82% for the Dice, Jaccard and Simple matching coefficients respectively while in the AAMO for the same coefficients and in the same order, the first three components accounted for 85%, 77% and 88% respectively. 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 (Figures 19 C and D); the Jaccard had four groupings for both MDS and PCA plots (Figures 20 A and B) while Simple matching had three groupings for both the MDS and PCA plots (Figures 20 C and D). 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. There was also no clear separation among the pathogens with respect to their agro-ecological zones confirming the findings from the dendrograms constructed. 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.

5 Summary

Considering the fact that the choice of the similarity coefficient used in clustering could have a great impact on the resulting classification, there is need to study and understand these coefficients better, so as to be able to make the right choice for specific situations. Many studies have been carried out without apparent reason for the choice of the similarity coefficient or clustering method, however, the use of a particular similarity coefficient combined with different clustering methods may give different results. The Dice and Jaccard similarity coefficients have been reported to give very similar results with respect to dendrogram structures, despite the fact that Jaccard is metric while Dice is believed to be non-metric. On the other hand, the Simple matching coefficient, which takes into consideration the negative co-occurrences of the individuals being compared, is known to give a different structure. In this study, these three coefficients were employed in carrying out cluster analysis (CA) using five (Unweighted Pair-Group Mean Arithmetic (UPGMA), Weighted Pair-Group Mean Arithmetic (WPGMA), complete linkage, single linkage and Neighbour-Joining (NJ)) clustering methods for simulated and experimental binary data sets. In the simulated data, the UPGMA method was used for the Dice and Jaccard coefficients while in the experimental data, all three coefficients were used with all the five clustering methods. The influence of the similarity coefficient and clustering methods in CA with respect to different populations was investigated.

The simulated data was done such that in-built structure was given to the data with two known groups in order to see whether the structure in the data will still be preserved after carrying out CA and whether the chosen coefficients will produce similar results. It was observed that the Dice and the Jaccard coefficients did give similar results, with a few exceptions, suggesting that the two coefficients could be used interchangeably with none having superiority over the other. In order to validate the observations in the simulated data, three experimental data sets from different populations were analysed. The data sets were (i) plantain production constraints data from different locations (states) in Nigeria (ii) data on evolution of powdery mildew populations from Germany and (iii) yam anthracnose amplified fragment length polymorphism (AFLP) marker data from two agro-ecological zones in Nigeria.

The dendrogram results of the plantain production constraints data from different locations (states) in Nigeria showed that locations were all mingled together suggesting that the locations did not have a strong effect on the production constraints. The mingling among the regional groupings also suggests that the production constraints are not specific to region. The Dice and Jaccard coefficients produced similar results for the UPGMA, complete linkage and single linkage clustering methods while the WPGMA and NJ methods produced different groupings for these two coefficients. The Simple matching coefficient produced a different grouping from these two coefficients for all methods. The principal component analysis (PCA) and the Multi dimensional scaling (MDS) plots revealed three main groupings for all three coefficients. These three groups were however, a mixture of some of the states in each group, confirming what was observed in the dendrogram groupings.

For the field experiment data on evolution of powdery mildew populations in different selection regimes, four treatment regimes (treatment1 – susceptible host, treatment2 – susceptible host + fungicide, treatment 3 – resistant host and treatment4 – resistant host + fungicide) were generated by the application of host resistant genes and fungicides used in the four treatments. Treatments 1 and 3 were combined together to form a data set with two known groups while treatments 2 and 4 were also combined together to form another data set with two known groups. Therefore the powdery mildew data was divided into two data sets consisting of susceptible host and resistant host without fungicide as well as susceptible host and resistant host with fungicide. The result showed that all clustering methods except the NJ produced different grouping for Dice and Jaccard coefficients in the dendrograms plotted for powdery mildew data without fungicide. The MDS and PCA plots also revealed two main groupings for all three coefficients. However, the powdery mildew data with fungicide produced different groupings for Dice and Jaccard coefficients in the UPGMA, WPGMA and NJ dendrograms and similar grouping in the complete linkage and single linkage dendrograms. The MDS and PCA plots revealed three main groupings which was different from what was observed in the data without fungicide, suggesting that the change could be as a result of the fungicide introduced.

The AFLP yam anthracnose marker data from two agro-ecological zones in Nigeria were based on three primers (ACMA, AAMG and AAMO) that gave polymorphic bands out of which binary value matrices were constructed and used for the CA. The UPGMA, complete linkage and single linkage produced similar dendrograms for Dice and Jaccard coefficients while the WPGMA and the NJ methods produced different dendrograms for the two coefficients in the ACMA and AAMO data. The MDS and PCA plots also produced four major grouping for all three coefficients for both data sets. However, in the AAMG data, all clustering methods except NJ produced the same dendrogram for Dice and Jaccard coefficients. The MDS and PCA plots produced different groupings for the three coefficients. For the Dice coefficient, the MDS and PCA plots produced different grouping each, which was different from what observed in the plantain and powdery mildew data sets, where the number of grouping produced in the two plots were the same for each coefficient. The PCA plot showed that two pathogens were revealed as outliers in its plot compared to the MDS plot. This may further support our observation that the Jaccard coefficient could be more stable than the Dice coefficient. In addition, the results from all three data sets suggest that the grouping of the pathogens by the markers is not related to their agro-ecological zones.

The consensus fork index (CFI) results used to compare the dendrograms showed varying level of similarity for all the CA methods. The NJ and single linkage methods seemed to give the lowest values. Therefore the single linkage method is not suggested as an appropriate method because of its tendency to produce lots of singletons in classifications.

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 Spearman correlation coefficients revealed that the Dice and Jaccard values for single linkage and complete linkage methods are perfectly monotonically related. The MDS and PCA analyses confirmed most of the groupings of the isolates as seen in the dendrograms. 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., Mignouna, H.D., Green, K.R. and Asiedu, R. (2003). Molecular taxonomic, epidemiological and population genetic approaches to understanding yam anthracnose disease. African Journal of Biotechnology 2 (12) 486-496.

Akinyemi, S.O.S., Aiyelaagbe, I.O.O. and Akyeampong, E. (2010). Plantain (*Musa spp.*) cultivation in Nigeria: A review of its production, marketing and research in the last two decades. Acta Horticulturae 879 211-218.

Akinyemi, S.O.S., Kintomo, A.A., Ojurongbe, T., Sallah, P.Y.K., Ndabamenye, T. and Nkezabahizi, D. (2009). Effects of fertilizer, organic mulch and sucker hot water treatment on nematode population and productivity of plantain. Journal of Applied Biosciences 16 887-893.

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

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 7(3) 695-705.

Bremer, K. (1990). Combinable component consensus. Cladistics 6 369-372.

Bryant, D. (2003). A Classification of Consensus Methods for Phylogenetics. In Bioconsensus: DIMACS Series in Discrete Mathematics and Theoretical Computer Science, M. Janowitz, F. Lapointe, F. McMorris, B. Mirkin, and F. Roberts, eds. (Providende, Rhode Island).

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

Conner, R.L., Kuzyk, A.D. and Su, H. (2003). Impact of powdery mildew on the yield of soft white spring wheat cultivars. Can J Plant Sci 83 725–728.

Craenen, K. (1998). Technical manual on black sigatoka disease of banana and plantain. Ibadan, Nigeria, International Institute of Tropical Agriculture. 23pp.

Curtis, B.C., Rajaram, S. and Gomez Macpherson, H. (2002). Bread wheat improvement and production.

Dalirsefat, S., Meyer, A. and 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 (71) 1-8.

DeCoster, J. (1998). Overview of factor analysis http://wwwstat-helpcom/noteshtml Retrieved (01/06/2011).

Dias, L.A., Picoli, E.A., Rocha, R.B. and Alfenas, A.C. (2004). A *priori* choice of hybrid parents in plants Genet Mol Res 3 356-368.

Duarte, M.C., Santos, J.B. and Melo, L.C. (1999). Comparison of similarity coefficients based on RAPD markers in the common bean. Genetics and Molecular Biology 22 427-432.

Erkki, O. (1989). Neural networks, principal components, and subspaces. International Journal of Neural Systems 1 (1) 61-68.

Everts, K., Leath, S. and Finney, P.L. (2001). Impact of powdery mildew and leaf rust on milling and baking quality of soft red winter wheat Plant Dis 85 423–429.

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.

FAO (2006). Production yearbook (Rome, FAO).

Finch, H. (2005). Comparison of distance measures in cluster analysis with dichotomous data. Journal of Data Science 3 85-100.

Flinn, J.C. and Hoyoux, J.H. (1976). Le bananier platain en Afrique. Fruits 31 520-530.

Fongeyn, A. (1976). The problem of plantain production in Cameroon. Fruits 31 692-694.

Gaugh, H.G. and Whittaker, R.H. (1981). Hierarchical classification of community data. Journal of Ecology 69 537-557.

Glawe, D.A. (2008). The powdery mildews: a review of the world's most familiar (yet poorly known) plant pathogens. Annu Rev Phytopathol 46 27-51.

Groenen, P.J.F. and Van de Velden, M. (2004). Multidimensional scaling. In Econometric Institute Report (Rotterdam, The Netherlands, Erasmus University Rotterdam).

Guillemot, J. (1976). Le bananier plantain en Cote d'Ivoire, sa production, ses possibilites. Fruits 31 684-687.

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. Theoretical and Applied Genetics 88 123-128.

Hartigan, J.A. (1975). Clustering Algorithms (Probability and mathematical Statistics) (John Wiley and Sons Inc).

Hatcher, L. and Stepanski, E. (1994). A step-by-step approach to using the SAS System for univariate and multivariate statistics (Cary, NC, Sas Institute Inc).

Hill, M.O., Bunce, R.G.H. and Shaw, M.W. (1975). Indicator species analysis, a divisive polythetic method of classification, and its application to a survey of native pinewoods in Scotland. Journal of Ecology 63 597-613.

Hill, T. and Lewicki, P. (2008). Statistics: Methods and applications (Statsoft Inc).

IITA (2000). Annual report of project 5: Improvement of yam-based systems (Ibadan, Nigeria, International Institute of Tropical Agriculture), pp. 70.

Inuma, T., Khodaparast, S.A. and Takamatsu, S. (2007). Multilocus phylogenetic analyses within Blumeria graminis, a powdery mildew fungus of cereals Mol Phylogenet Evol 44 741–751.

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.

Jongman, R.H.G., Ter Braak, C.J.F. and Van Tongeren, O.F.R. (1995). Data analysis in community and landscape ecology (Cambridge University Press, UK).

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

Koopman, W.J.M., Zevenbergen, M.J. and Van Den Berg Ronald, G. (2001). Species relationships in *Lactuca* S.L. (*Lactuceae, Asteraceae*) inferred from AFLP fingerprint. Amer J Bot 88 (10) 1881-1887.

Kosman, E. and Leonard, K.J. (2005). Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid and polyploidy species. Molecular Ecology 14 415 - 424.

Kuhner, M.K. and Felsenstein, J. (1994). A Simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11 (3) 459-468.

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.

Lambert, J.M., Meacock, S.E., Barrs, J. and Smartt, P.F.M. (1973). AXOR and MONIT: Two new polythetic-divisive strategies for hierarchical classification. Taxon 22 173-176.

Legendre, P. and Legendre, V. (1983). Numerical Ecology (Amsterdam, Elsevier).

Lescot, T. (1998). Banana: Little-known wealth of variety. Fruitrop 51 8-11.

McGarigal, K., Cushman, S. and Stafford, S. (2000). Multivariate statistics for wildlife and ecology research (Springer).

Melin, P. and Djomo, E. (1972). Importance economique d ela banane plantain au Cameroon. Fruits 27 251-254.

Meyer, A., Garcia, A.A.F., 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 (1) 83 - 91.

Mobambo, K.N., Gauhl, F., Vuylsteke, D., Ortiz, R., Pasberg-Gauhl, C. and Swennen, R. (1993). Yield loss in plantain from black sigatoka leaf spot and field performance of resistant hybrids. Field Crops Research 35, 35-42.

Naku, M. (1983). La situation du plantain au Zaire: Cas du Mayombe. Fruits 38 (4) 306-308.

Nosofsky, R.M. (1992). Similarity scaling and cognitive process models. Annual Review of Psychology 43 25-53.

Oberhaensli, S., Parlange, F., Buchmann, J.P., Jenny, F.H., Abott, J.C., Burgis, T.A., Spanu, P.D., Keller, B. and Wicker, T. (2011). Comparative sequence analysis of wheat and barley powdery mildew fungi reveals gene colinearity, dates divergence and indicates host-pathogen co-evolution. Fungal Genetics and Biology 48 (3) 327 - 334.

Orloci, L. (1978). Multivariate analysis in vegetation research (The Hague Dr. W. Junk B.V.).

Palmer, M.W. (2008). Ordination methods - an overview. http://ordinationokstateedu/overviewhtm Retrieved 9.12.2011.

Pielou, E.C. (1984). The interpretation of ecological data: a primer on classification and ordination (New York, Wiley).

R (2008). R: A language and environment for statistical computing, D.C. Team, ed. (Vienna, Austria, R Foundation for Statistical Computing).

Reif, J.C., Melchinger, A.E. and Frisch, M. (2005). Genetical and mathematical properties of similarity and dissimilarity coefficicents applied in plant breeding and seed bank management. Crop Science 45 1-7.

Restrepo, G. and Villaveces, J.L. (2005). From trees (dendrograms and consensus trees) to topology. Croatia Chemica Acta CCACAA 78 (2) 275-281.

Rohlf, F.J. (1982). Consensus Indices for Comparing Classifications. Mathematical Biosciences 59 (1) 131-144.

Rohlf, F.J. (1992). Program numerical taxonomy and multivariate analysis system (New York, Applied Statistics Inc).

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

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

SAS (2011). Principal component analysis. http://supportsascom/publishing/pubcat/chaps/55129pdf Retrieved 02.09.2011.

Schill, P., Gold, C.S. and Afreh-Nuamah, K. (1996). Assessment and characterization of constraints in plantain prodcution in Ghana as an example for West Africa. Paper presented at: Plantain and banana: Production and research in west and central africa Proceedings of a regional workshop (International Institute of Tropical Agriculture, Onne, Nigeria).

Sesli, M. and Yegenoglu, E.D. (2010). Compare various combinations of similarity coefficients and clustering methods for *Olea europaea sativa*. Scientific Research and Essays 5(16) 2318-2326.

Shao, K. and Sokal, R.R. (1986). Significance tests of consensus indices. Systematic Zoology 35 582-590.

Shepard, R.N. (1987). Towards a universal law of generalization for psychological. Science 237 1317-1323.

Smith, L.I. (2002). A tutorial on principal component analysis.

Sneath, P.H.A. and Sokal, R.R., eds. (1973). Numerical Taxonomy (San Francisco, WH Freeman).

Sokal, R.R. and Rohlf, F.J. (1962). The comparison of dendrograms by objective methods. Taxon 11 33-40.

Speijer, P.R. and De Waele, D. (1997). Screening of musa germplasm for resistance and tolerance to nematodes. IN-IBAP technical guidelines 1. International Network for the Improvement of Banana and Plantain 47pp.

Speijer, P.R. and Gold, C.S. (1995). Root health assessment in banana and plantain. IITA research guide, Ibadan, Nigeria, International Institute of Tropical Agriculture 39 pp.

Speijer, P.R., Rotimi, M.O. and De Waele, D. (2001). Plant parasitic nematodes associated with plantain (*Musa* spp., AAB-group) in southern Nigeria and their relative importance compared to other biotic constraints. Nematology Vol 3(5) 423-436.

Steyvers, M. (2002). Multidimensional Scaling
In Encyclopedia of Cognitive Science (London, UK, Nature Publishing Group).

Stover, R.H. (1972). Banana, plantain and abaca diseases (Kew, UK, Commonwealth Mycological Institute).

Stuetzle, W. and Nuggent, R. (2007). A generalized single linkage method for estimating the cluster tree of a density. In Technical Report 514 (University of Washington).

Swennen, R. and De Langhe, E. (1985). Growth parameters of yield of plantain (*Musa* cv. ABB). Annals of Botany 56 197-204.

Swennen, R. and Vuylsteke, D. (1988). Bananas in Africa: Diversity, uses and prospects for improvement. Paper presented at: Crop Genetic Resources of Africa (International Institute of Tropical Agriculture, Ibadan, Nigeria).

Swofford, D.L. (1991). When are phylogeny estimates from molecular and morphological data incongruent? Phylogenetic analysis of DNA sequences (Oxford Univeristy Press).

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

Vandev, D.L. and Tsvetanova, Y.G. (1995). Ordered Dendrogram.

Williams, W.T. and Lambert, J.M. (1959). Multivariate methods in plant ecology I. Association-analysis in plant communities. Journal of Ecology 47 83-101.

Wilson, G. (1987). Status of bananas and plantains in West Africa. Paper presented at: Banana and plantain breeding strategies (Cairns, Australia).

Wilson, G.F. (1983). Production de plantains: Perspective pour ameliorer la situation alimentaire sous les tropiques. Fruits 38 229-239.

Declaration

"I declare that I have completed this dissertation single-handedly without the unauthorized help of a second party and only with the assistance acknowledged therein. I have appropriately acknowledged and referenced all text passages that are derived literally from or are based on the content of published or unpublished work of others, and all information that relates to verbal communications. I have abided by the principles of good scientific conduct laid down in the charter of the Justus Liebig University Giessen in carrying out the investigations described in the dissertation."

_____ _____

Taiwo Adetola Ojurongbe Place, Date

List of Figures

List of Tables

List of Abbreviations

AB, Abia

AFLP, Amplified Fragment Length Polymorphism

AK, Akwa-Ibom

AN, Anambra

CA, Cluster Analysis

CFI, Consensus Fork Index

CR, Cross-Rivers

DE, Delta

DPC, Dichotomized Production Constraints

DSS, Diagnostic Survey Sample

ED, Edo

FA, Factor Analysis

F, Forest

GIS, Geographic Information Systems

GS, Guinea Savannah

IITA, International Institute of Tropical Agriculture

IM, Imo

MDS, Multi Dimensional Scaling

NJ, Neighbour-Joining

ON, Ondo

PAHC, Polythetic Agglomerative Hierarchical Clustering

PCA, Principal Component Analysis

RV, Rivers

SAHN, Sequential Agglomerative Hierarchical and Non-overlapping

SM, Simple Matching

UPGMA, Unweighted Pair-Group Mean Arithmetic method

WPGMA, Weighted Pair-Group Mean Arithmetic method

UPGMC, Unweighted Pair-Group Method using Centroid Average

WPGMC, Weighted Pair-Group Method using Centroid Average

Appendix

Appendix A - Tables from Plantain data

Table A1: DSS-Plantain data clusters as seen in the dendrogram using WPGMA method.

Cluster	Dice	Jaccard	Simple Matching
I	AB1,ON2,ED3,AB3,CR12, IM1, RV9,AK1,CR9	AB1,ON2,ED3,AB3, CR12,IM1,RV9,RV1, RV3,AB4,AB5	AB1,ON2,ED3,AB3, CR9,DE10
II	AK5,CR15,AK6,CR16,IM4, AN1,CR6,IM5,IM3	AK1,CR9,AK3,IM2, RV6	AK1,AK7,AK6,CR16, CR15,CR12,IM1,RV9
III	AB2,RV11,AK4,AK3,IM2, RV6,RV1,RV3	AK5,CR15,AK6,CR16, IM4,AN1,CR6,IM5,IM3	AK3,IM2,AN1,IM4, CR5,IM5,IM3
IV	CR2,CR4,ON4,CR5,CR10, CR7,CR8	CR2,CR4,ON4	AB2,RV11,AK4,AK2, DE2,RV2,DE3
V	DE2,DE3	CR5,CR10,CR7,CR8	CR5,CR10,CR7,CR8, RV7
VI	AK2,RV2,DE10,RV7	DE10,RV7	AB4,AB5,RV4,DE4, ON1,ON6,ED1,RV1, RV3
VII	CR1,CR13,DE1,ON3,DE7, ED4,ED7,CR11,DE8,RV5	AB2,RV11,AK4,AK2, RV2,DE2,DE3	CR1,CR3,DE6,DE9, RV6,RV8
VIII	CR14,RV10,ED10,ED6,ED2, ED9,ED8,DE5,ON5,ED5	DE4,ON1,ON6,ED1, DE6,DE9	CR11,CR13,CR14,ED 10,RV10,DE1,ON3, DE8,RV5,DE7,ED4, ED7

Table A2: DSS-Plantain data clusters as seen in the dendrogram using complete linkage method.

Cluster	Dice and Jaccard	Simple Matching
I	AB1,ON2,CR9,ED3,AB3,RV1,RV3	AB1,ON2,ED3,RV7,CR12,IM1,RV9
II	AB2,RV11,AK4,AK2	AB3,CR9,DE10,RV1,RV3
III	AB4,AB5,AN1,CR6,IM5,IM3,DE3	AK1,AK7,AK6,CR16,CR15,CR1, CR3,CR2,CR4,ON4
IV	AK1,RV6,AK3,IM2,CR12,IM1,RV9	AB2,RV11,AK4,AK2
V	CR5,CR10,CR7,CR8,DE10,RV7	CR5,CR10,CR7,CR8,RV6,RV8
VI	DE2,RV2,DE6,DE9	AB4,AB5,RV4, AK3,IM2,CR5,IM5,IM3, AN1,IM4,DE3
VII	AK5,CR15,AK6,CR16,IM4, CR2,CR4,ON4, R1,CR13,CR3,RV8	CR11,DE8,RV5,CR13,CR14,ED10,RV10, DE1,ON3,DE7,ED4,ED7
VIII	DE5,ON5,ED2,ED9,ED8	DE5,ON5,ED5,ED2,ED6,ED9,ED8
IX	CR11,DE1,ON3,DE8,RV5	DE2,RV2,DE6,DE9,DE4,ON1,ON6,ED1
X	CR14,RV10,ED10,ED10,ED6, ED5,DE7,ED4,ED7	
XI	DE4,ON1,ON6,ED1,RV4	

Table A3: DSS-Plantain data clusters as seen in the dendrogram using single linkage method.

Cluster	Dice and Jaccard	Simple Matching
I	AB1,ON2,ED3,CR7,CR8	AB1,ON2,ED3
II	CR5,CR10	CR7,CR8
III	AB2,RV11,AK1,AK5,AK6,CR16, AK3,IM2,CR6,IM5,IM3,RV6,IM1, RV9,AN1,CR12,IM4	AB2,RV11,AK4,DE3
IV	RV1,RV3	AK1,AK6,CR16,AK7,CR9,RV6,AK3, IM2,CR5,IM5,IM3,IM1,RV9,AN1,IM4, RV8,CR12
V	DE5,ON5	CR5,CR10
VI	CR1,CR13,CR14,RV10,ED10,ED6, ED2,ON3,DE7,ED4,ED7,ED9,DE8, RV5	CR1,CR3,CR13,CR14,ED10,RV10,DE1, DE7,ED4,ED7,ED6,DE8,RV5,ON3, ED2,ED9,ED8
VII	DE4,ON1,ON6,ED1	DE5,ON5,ON4
VIII		DE4,ON1,ON6,ED1
IX		RV1,RV3
Singletons	AK4,RV8,CR9,DE3,AB3,CR15,CR4, ON4,AK2,DE2,RV2,DE1,ED5,ED8, CR2,CR3,AB5,DE9,DE10,AB4,RV7, CR11,DE6,RV4,DE10	AB3,AK2,DE10,RV2,CR11,DE9,ED5, CR2,DE2,CR15,CR4,DE6,RV7,RV4, AB4, AB5

Table A4: DSS-Plantain data clusters as seen in the dendrogram using NJ method.

Cluster	Dice	Jaccard	Simple Matching
I	AB1,ED3,ON2, AB2,CR7,RV7, CR5 CR10,CR8,AB4, AB5,RV4	AB1,ED3,ON2,AK1,CR9, CR12,IM1,RV9	AB1,ED3, ON2,AK1,CR9, CR12,IM1,RV9,AB3,RV1,RV3
II	AB3,CR9,DE10, AK2,RV2,AK4, DE3,RV11	AB2,RV7,CR5,CR10,CR7, CR8	AB4,AB5
III	DE4,ON1,ON6, ED1,RV1,RV3	CR1,CR3,CR11,CR13,DE1 ,CR14,ED10,RV10,DE8, RV5,ON3,DE7,ED4,ED7, DE6,DE9	DE2,RV2
IV	AK1,RV6,AK3, IM2	DE5,ON5,ON4,ED5,ED2, ED6,ED9,ED8,RV6,RV8	DE4,ON1,ON6,DE6,ED1,RV4

Continued

Table A4 continued

V	AK5,CR16, AK6,CR15, AN1,IM4, CR2,CR4, CR6,IM5	AK3,IM2,AK5,CR16, AK6,CR15,AN1,IM4, CR2,CR4	AK3,IM2,CR2,CR4,AK6, CR16,CR15,AK7,IM4, AN1,CR5,IM5,IM3,DE3, RV11
VI	CR1,CR3,RV8	AB3,AB4,AB5,RV4,DE2,RV2, DE4,ON1,ON6,RV1,ED1	AK2,AK4,DE10
VII	CR11,CR13,DE 1,CR14,ED10, RV10 DE8,RV5,ON3, DE7,ED4,ED7	AK2,AK4,DE10,RV3	AB2,RV7,CR5,CR10,CR7, CR8
VIII	DE5,ON5,ON4, ED5	CR6,IM5,IM3	CR1,CR3,CR11,DE1,CR14, RV10,ED10,DE8,RV5,ON3, DE7,ED4,ED7,DE9
IX	ED2,ED6,ED9, ED8	DE3,RV11	DE5,ON5,ON4,ED5,ED2, ED6,ED9,ED8
X	DE2,DE6,DE9		RV6,RV8
XI	CR12,RV9,IM1		

Appendix B - Tables from Mildew data

Table B1: Mildewtrt13 data clusters as seen in the dendrogram using WPGMA method.

Cluster	Dice and Jaccard	Simple Matching
I	A1,A21,A33,A7,A19,A24,A17, A40,A34,A35	A1,A21,A28,A32,A2,A8,A13,B24
II	A2,A32,A28,B24,A6,A31,A8, A9,A27	A6,A31,A27,A9,A34,A35
III	A13,B32,B16,A39,B3,B18,B30, B6,B14,B40	A7,A17,A19,A24,A33
IV	B1,B26,B10,B23,B17,B35, B22,B28,B31	A39,B14,B6,B3,B30,B18,B32,B40
V	B13,B14	A40,B16,B13,B34
VI		B1,B26,B10,B23,B22,B28,B31,B17,B35

Table B2: Mildewtrt13 data clusters as seen in the dendrogram using complete linkage method.

Cluster	Dice and Jaccard	Simple Matching
I	A1,A21,A33,B22,B28,B31,B17, B35	A1,A21,A28,A32,A13,B24
II	A2,A32,A28,B24,A6,A31,A8, A9,A27	A6,A31,A27,A9
III	A7,A19,A24,A17,A40,A34,A35	A7,A17,A19,A24,A34,A35
IV	A13,B32,B16,B18,B30,A39,B3,B6, B14,B40	A2,A8,B13,B34,B1,B26,B10,B23
V	B1,B26,B10,B23,B13,B34	B17,B35,B22,B28,B31
VI		A33,A40,B16,A39,B14,B6
VII		B3,B30,B18,B32,B40

Table B3: Mildewtrt13 data clusters as seen in the dendrogram using single linkage method.

Cluster	Dice and Jaccard	Simple Matching
I	A1,A21,A28	A1,A21,A28,A27,A2,A8,A6,A31,B34,B26,B24, B22,B31,B28,A9,B13,B10,B23,B1,B17,A7,A17, A19,A24,A34,A35,A32,A39,B14,B32,B16,B40, A40,B3,B30, B18,A13,B6,B35
II	A39,B3,B32,B14,B28,B 31	A33
III	B17,B35	
IV	A7,A17,A40,A19,A35	
Singletons	B24,A13,B22,B40,B16, B1,B6,B30, B18,A32,B23,A24,A27, A9,A34,B26, A2,A6,A31,B10,B13,B3 4,A33,A8	

Table B4: Mildewtrt13 data clusters as seen in the dendrogram using NJ method.

Cluster	Dice	Jaccard	Simple Matching
I	A1,A21,A6,A31, A8,A9,A27,A32, A28,A3,B24,A13	A1,A21,A28,A32,A2,B24	A1,A21,A28,A32,A2, B24
II	A7,A24,A17,A34, A35,A40,A19, B17,B35	A6,A31,A34,A8,B13,B34,A7, A24,A19,A9,A27,A35,A17,A40	A7,A17,A40,A24,A19, A33,A13
III	A39,B6,B14,B18, B40,B3,B30,B16, B32	A13,A33	A6,A31,A9,A27,A34, A35,A8,B13,B34,B16
IV	B1,B10,B23,B13, B34,B26	B1,B10,B23,B26,B17,B35,B22,B 31,B28	B1,B17,B35,B10,B23, B26
V	B22,B31,B28,A33	B16,B32	A39,B3,B6,B14,B40, B18,B30,B32
VI		A39,B3,B6,B14,B18,B40	B22,B31,B28
VII		B30	

Table B5: Mildewtrt24 data clusters as seen in the dendrogram using WPGMA method.

Cluster	Dice	Jaccard	Simple Matching
I	C1,C18,C22, D29	C1,C18,C22,D29	C1,C18,C22,D29
II	C16,D16,D8, D5,D20,D40, D14,D31,D32	C4,C20,C34,D26,C11,C19,D6,D 30,C14,D27,D9,C15, C36	C2,C16,D39,D14,D31,D18, C28,D10,D8,D16,D32
III	C4,C20,C34, D26,C11,C19, D6,D30,C14, D27,D9,C15, C36	C2,D18,D39,C28,D10,C16,D16, D8,D5,D20,D40,D14,D31,D32	C4,C20,C34,D26,C11,C19, D6,D30,C14,D27,D9,C15, C36
IV	C2,D18,D39, C28,D10,C7, C8	D2,D4,D28	D5,D20,D40
V	C5,C6,C10, C26	C3,C6,C10,C26	D2,D4,D28
VI	D2,D4,D28		C5,C6,C10,C7,C8,C26

Table B6: Mildewtrt24 data clusters as seen in the dendrogram using complete linkage method.

Cluster	Dice and Jaccard	Simple Matching
I	C1,C18,D14,D31,D32,C22, D29,D20,D40,C16,D16,D8	C1,C18,C22,D29,C7,C8,C2,C16, D39,D18,C28,D10,D14,D31,D16, D32
II	C4,C20,C34,C15,C36,C11,C19, D6,D30	C5,C6,C10,C26
III	C14,D27,D9,D5,D26	C4,C20,C34,C15,C36,C11,D30,C19, D6,C14,D27,D9
IV	D2,D4,D28	D5,D26,D20,D40
V	C2,D18,D39,C28,D10,C7,C8	D2,D4,D8,D28
VI	C5,C6,C10,C26	

Table B7: Mildewtrt24 data clusters as seen in the dendrogram using single linkage method.

Cluster	Dice and Jaccard	Simple Matching
I	C1,C8,C2,C4,C20,C34,C14, D27,D18,D26,C36,D6,D9,D 14,D31,D39,D10,D30,C15, C16,D5,C19,D40,D29,D32	C1,C18,C22,C19,C11,C15,D29,C2,C16, D10,D39,C28,C6,C5,D16,C8,D6,C4,C20, D26,D18,D8,D2,D30,C34,D14,D31,D40, D32,C36,D9,C14,D27,C7,C10,D5,D20
II	C7,C8	
Singletons	D20,D8,D16,C28,C22,C11, D2,C6,C5,C10,D28,C26,D4	D4,D28,C26

Table B8: Mildewtrt24 data clusters as seen in the dendrogram using NJ method.

Cluster	Dice	Jaccard	Simple Matching
I	C1,C18,C5,C10,C26, C6, C8,C7,C22,D29	C1,C18,C22,C5,C10, C26,C6,C8,C7	C1,C18,C5,C10,C26, C6,C8,C7
II	C2,D39,C16,D8, D28,D16,C28,D10	C2,D39,C28,D10,C16, D16,D8,D28,D14, D31,D18	C2,D39,C16,D8,D28,D16, D8,D28, D16,C28,D10
III	C4,D26,C20,C34	C4,D26,C20,C34,C14, D27, C36,D29	D5,D20,D40,D6,D32,D14, D31,D18
IV	C11,C15,D9,D2,D4, D30,D6,C19, C36,C14,D27	C11,C15,D9,D2,D4, D30	C4,D26,C20,C34,C14,D27, C36
V	D5,D20,D40,D32, D14,D31,D18	C19,D6,D5,D20,D40, D32	C11,C15,D2,D4,D30,D9,C19
VI			C22,D29

Appendix C – Tables from Yam marker data

Table C1: ACMA data clusters as seen in the dendrogram using WPGMA method.

Cluster	Dice	Jaccard	Simple Matching
I	1F,38F,39F	1F,38F,39F	1F,15GS,16F,36GS, 20GS,2F,7F,26GS, 33GS,34F
II	10GS,22GS,40F,27GS, 30F,31F	4F,29GS,18F,24F,15GS, 16F,20GS	4F,18F,24F,29GS,40F, 22GS, 27GS,30F,31F,38F,39F
III	17F,41F,43F,42GS, 45GS,53F,35GS,48GS, 33GS,34F	10GS,22GS,40F,27GS, 30F,31F	17F,41F,43F,42GS, 45GS,53F,35GS,48GS, 3GS
IV	4F,29GS,18F,24F, 15GS,16F, 20GS,36GS,35F	17F,41F,43F,42GS, 45GS,53F,35GS,48GS, 33GS,34F	10GS,35F
V	2F,7F,26GS	2F,7F,26GS	
Singleton	3GS	3GS,35F,36GS	

Table C2: ACMA data Clusters as seen in the dendrogram using complete linkage method.

Cluster	Dice and Jaccard	Simple Matching
I	1F,38F,38F	1F,15GS,16F,36GS,20GS,2F,7F, 4F,18F,24F,38F,39F
II	10GS,22GS,40F,27GS,30F,31F,33GS, 34F,35GS,48GS,3GS	17F,41F,43F,42GS,53F,45GS,35GS, 48GS,3GS
III	2F,7F,26GS	10GS,35F,26GS,33GS,34F
IV	4F,29GS,18F,24F,17F,41F,43F,42GS, 45GS,53F,20GS	22GS,29GS,40F,27GS,30F,31F
V	15GS,16F,36GS,35F	

Table C3: ACMA data clusters as seen in the dendrogram using single linkage method.

Cluster	Dice and Jaccard	Simple Matching
I	4F,29GS,18F,40F,24F	1F,15GS,16F,2F,7F,4F,24F,18F,20GS,29GS, 36GS
II	17F,41F,43F,42GS,45GS	17F,41F,43F,42GS,45GS
III	27GS,30F,31F,35GS	30F,31F
IV	33GS,34F	33GS,34F
V	15GS,16F	
VI	38F,39F	
VII	2F,7F	
Singletons	1F,53F,48GS,22GS,20GS,10GS, 35F,3GS,26GS,36GS	40F,53F,35GS,35F,27GS,22GS,39F,38F,26GS, 48GS,10GS,3GS

Table C4: ACMA data clusters as seen in the dendrogram using NJ method.

Cluster	Dice	Jaccard	Simple Matching
I	1F,38F,39F,15GS, 16F,36GS	1F,28F,39F	1F,38F,39F,2F,7F,26GS, 36GS,15GS,16F,24F,4F, 29GS,18F,20GS
II	4F,24F,18F,20GS, 29GS	4F,29GS,18F, 20GS,15GS,16F, 24F	10GS,40F,22GS,27GS,30F, 31F,33GS,34F,48GS,3GS, 35GS,53F,17F,35F,41F,43F, 45GS,42GS
III	17F,35F,41F,43F, 42GS,45GS,53F, 35GS,48GS,3GS	10GS,40F,22GS, 30F,31F,33GS, 34F	
IV	27GS,30F,31F	17F,41F,43F,42 GS,45GS, 53F,35GS,48GS, 3GS	
V	2F,7F,26GS,	36GS,35F	
VI		2F,7F,26GS	
Singletons	22GS,40F,10GS		

Table C5: AAMG data clusters as seen in the dendrogram using WPGMA method.

Cluster	Dice and Jaccard	Simple Matching
I	1F,48GS,36GS	1F,48GS,25GS,30F,42GS,46GS, 41F,33GS,26GS,36GS
II	10GS,39F,27GS,40F,34F,53F, 22GS,33GS	16F,17F,18F,20GS,31F,24F,29GS, 35GS,43F
III	25GS,30F,41F,29GS,42GS,26GS	10GS,27GS,15GS,39F,34F,40F,53F,22GS
IV	15GS,31F,16F,20GS,17F,18F	
V	24F,35GS,43F,46GS	

Table C6: AAMG data clusters as seen in the dendrogram using the complete linkage Method.

Cluster	Dice and Jaccard	Simple Matching
I	1F,48GS,26GS,36GS	1F,48GS,25GS,30F,42GS,46GS,41F, 26GS,36GS
II	15GS,31F,53F,16F,20GS,17F,18F, 24F,35GS,43F	16F,17F,18F,24F,29GS,35GS,43F
III	10GS,39F,27GS,40F,34F,22GS,33GS	20GS,31F,53F
IV	25GS,30F,41F,29GS,42GS,46GS	10GS,27GS,34F,40F,15GS,39F,22GS, 33GS

Table C7: AAMG data clusters as seen in the dendrogram using single linkage method.

Cluster	Dice and Jaccard	Simple Matching
I	10GS,39F,15GS,27GS,53F, 40F,34F,31F,22GS,33GS, 16F,17F,18F,20GS	1F,48GS
II	35GS,43F	16F,17F,18F
III		25GS,30F,42GS,46GS,29GS,41F,33GS, 36GS
IV		35GS,43F
V		10GS,27GS,15GS,39F,53F
Singletons	1F,48GS,25GS,30F,42GS, 41F,26GS,29GS,24F,36GS, 46GS	20GS,26GS,31F,22GS,24F,34F,40F

Table C8: AAMG data clusters as seen in the dendrogram using NJ method.

Cluster	Dice	Jaccard	Simple Matching
I	1F,48GS,53F,40F, 10GS,27GS,39F	1F,48GS,25GS,30F, 41F,42GS,46GS, 26GS,36GS	1F,48GS,36GS,46GS, 25GS,30F,26GS,41F,42GS
II	15GS,16F,17F,18F, 31F,20GS,24F,29GS, 35GS,43F	16F,17F,18F,20GS, 31F,24F,29GS,35GS, 43F,22GS,33GS	16F,17F,18F,20GS,31F, 29GS,35GS,43F,24F
III	25GS,30F,41F,42GS, 46GS,26GS,36GS, 33GS	10GS,27GS,15GS, 39F,40F,53F,34F	10GS,39F,27GS,40F,53F, 15GS,34F,22GS,33GS
Singletons	34F,22GS		

Table C9: AAMO data clusters as seen in the dendrogram using WPGMA method.

Cluster	Dice	Jaccard	Simple Matching
I	1F,27GS	1F,27GS,29GS	1F,27GS,34F
II	10GS,43F,53F,15GS,16F, 17F,18F,48GS,29GS, 20GS,40F	10GS,43F,53F,15GS, 16F,17F,18F,48GS,3GS, 20GS,40F,24F	9GS,39F,52GS, 22GS,33GS,41F, 42GS,45GS,8GS, 36GS,38F
III	36GS,38F	26GS,35GS	26GS,31F,35GS
IV	26GS,35GS,34F	36GS,38F,8GS	10GS,43F,53F, 3GS,24F
V	41F,42GS,3GS,45GS,8GS	41F,42GS,45GS	15GS,16F,17F, 18F,48GS,29GS, 20GS,40F
VI	9GS,39F,52GS,22GS, 33GS	9GS,39F,52GS,22GS, 33GS	
Singleton	24F,31F	34F,31F	

Table C10: AAMO data clusters as seen in the dendrogram using complete linkage method.

Cluster	Dice and Jaccard	Simple Matching
I	1F,27GS,29GS	1F,27GS,29GS
II	10GS,43F,18F,48GS,15GS, 16F,17F,20GS,40F	10GS,43F,18F,48GS,20GS, 40F, 15GS,16F,17F
III	24F,53F,3GS	24F,53F,3GS,35GS,8GS, 36GS,38F
IV	36GS,38F	9GS,39F,52GS,22GS,33GS,41F, 42GS,45GS,26GS,31F,34F
V	26GS,35GS,34F,31F	
VI	41F,42GS,45GS,8GS	
VII	9GS,39F,52GS	
VIII	22GS,33GS	

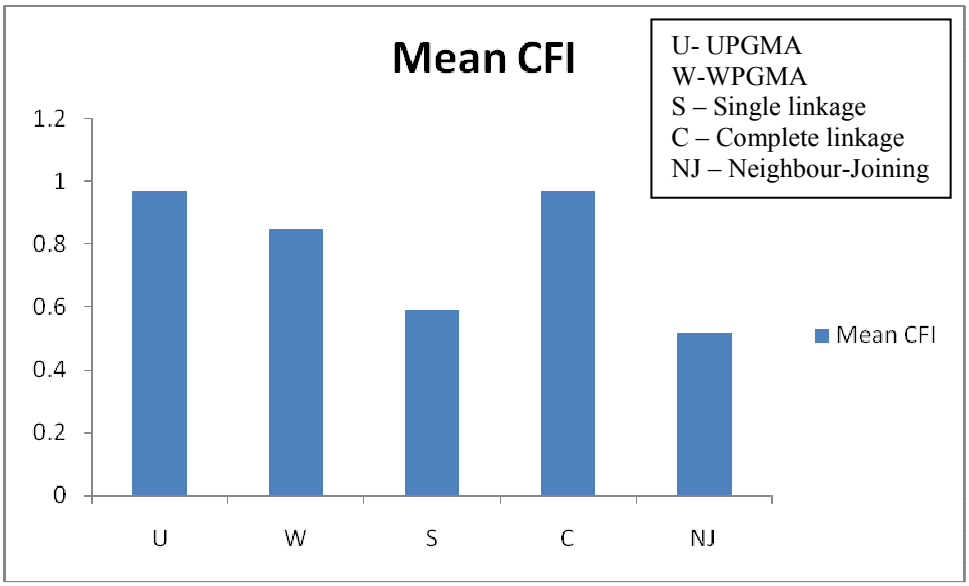
Table C11: AAMO data clusters as seen in the dendrogram using single linkage method.

Cluster	Dice and Jaccard	Simple Matching
I	1F,10GS,15GS,16F,17F,18F , 48GS,43F,53F,20GS,27GS, 29GS,3GS	9GS,39F,52GS,33GS,22GS,38F, 36GS
II	41F,42GS	10GS,15GS,16F,17F,18F,48GS,43F, 53F,3GS,20GS,41F,42GS,45GS
III	36GS,38F	
IV	39F,52GS	
Singletons	40F,35GS,24F,45GS,34F, 26GS,8GS,31F,9GS,33GS, 22GS	1F,27GS,26GS,35GS,31F,29GS,24F,40F, 8GS,34F

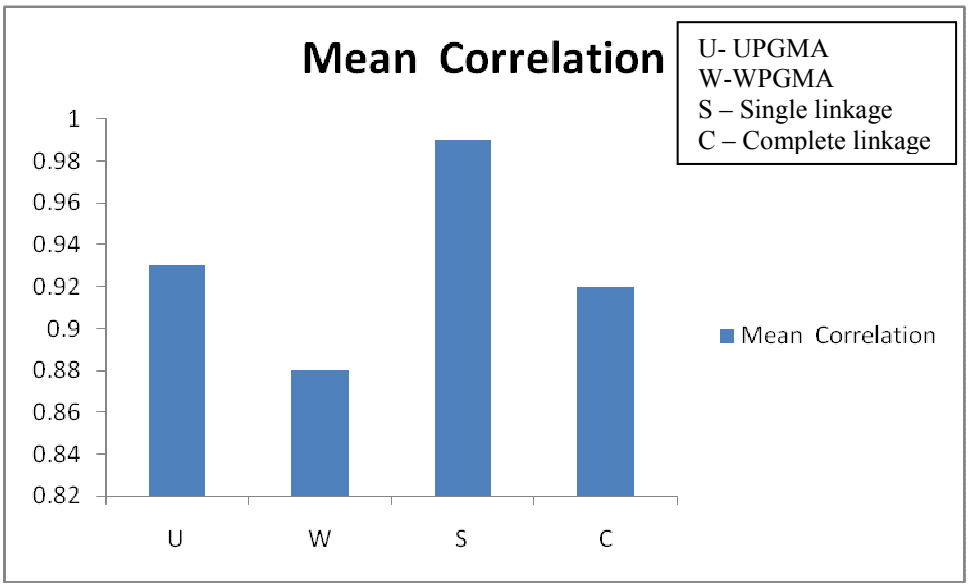
Table C12: AAMO data clusters as seen in the dendrogram using NJ method.

Cluster	Dice	Jaccard	Simple Matching
I	1F,34F,27GS	1F,34F,27GS	1F,34F,26GS,35GS
II	10GS,20GS,15GS, 16F,17F,18F,48GS, 29GS,26GS,31F, 35GS,43F,53F	9GS,39F,52GS,22GS, 33GS,36GS,38F,24F	10GS,20GS,15GS,16F, 17F,18F,48GS,43F,27GS, 29GS,40F
III	41F,42GS,45GS, 3GS,8GS	26GS,31F,35GS, 41F,42GS,45GS, 3GS,8GS	9GS,39F,52GS,22GS, 33GS,31F,41F,42GS,45GS ,8GS,36GS,38F,24F
IV	9GS,39F,52GS, 22GS,33GS,36GS, 38F	20GS,40F	53F,3GS
V		10GS,43F,53F,15GS, 16F,17F,18F,48GS	
Singletons	40F,24F	29GS	

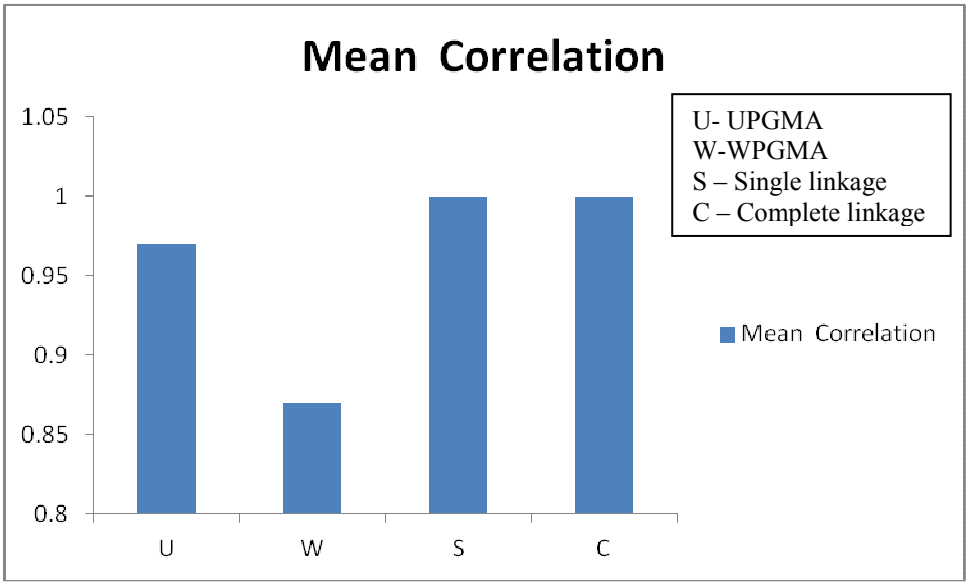
Appendix D – Histogram for Mean CFI and Mean Correlation for All CA methods and All Experimental Data.



App. D1: Mean CFI for Dice and Jaccard for all experimental data and different clustering methods.



App. D2: Mean Pearson Correlation Coefficient for Dice and Jaccard Cophenetic distances plot for all experimental data and different clustering methods.



App. D3: Mean Spearman Correlation Coefficient for Dice and Jaccard Cophenetic distances plot for all experimental data and different clustering methods.

Dedication

This work is dedicated to the Alpha and Omega, the beginning and the ending, the One who has made this work come through, The Almighty GOD.

Acknowledgement

I would like to express my profound gratitude to Prof. Dr. Matthias Frisch for allowing me to obtain my doctoral degree under his supervision. I would like to appreciate Dr Gabriel Schachtel for his valuable contributions, advice and guidance in the course of this project; I am indeed very grateful, vielen Dank! I am sincerely grateful to Prof. Dr. Dr. Wolfgang Friedt for accepting to be my co-supervisor. Many thanks for your guidance and wonderful advice. The financial support of the German-Israeli Foundation for Scientific Research and Development (GIF Project) from 2004 to 2006 is gratefully acknowledged.

I am sincerely grateful to all the staff of Biometry and Population Genetics both (old and new). Words cannot express my gratitude to Frau Renate Schmidt for her motherly advice and encouragement all the time, "Herzlichen Dank". My sincere appreciation goes to Prof. Dr. Wolfgang Koehler, Dr Joern Pons-Kuehnemann, Dr Gerrit Eichner, Dr. (Mrs) Foye Aduramigba-Modupe, Dr. (Mrs) Omolara Olaniyi for their valuable contributions towards the successful completion of this program, God bless you all. To all my friends both here in Germany, the United States and Nigeria, the IBC family in Giessen and FeG family in Tuebingen, Angelika & Glenn Carlson, Harry & Marianne Boettiger, Herr Hans Joachim Selzer for his financial support, Pastor Buddy and Kathy thank you all for your support and for being there all the time. I say a big thank you, for all your prayers, encouragement and support, God bless you all.

To the members of my family, my parents, my sisters and brothers, you are all wonderful!! Thanks so much for all the support in every way, spiritual, moral and physical. I appreciate you all. To my wonderful angels: IfeOluwa and AanuOluwa, thanks for your understanding and love, for your prayers and being there to support and encourage mummy in her work! The Lord bless you real good. To my husband, My Love and best friend, Dr Olusola Ojurongbe, I will forever be grateful to you for all your support, love, care, prayers and encouragement. You are ONE in a million, thank you so much, the Lord bless you real good. The Lord will perfect all that concerns us to His glory and we will continue to shine for HIM and HIM alone.

Now unto the King eternal, immortal, invisible, the only wise God, be all glory, honor, power and praise, now and forever more, amen. THANK YOU MY FATHER, YOU ARE THE BEST!!!