Disclaimer: unless otherwise agreed by the Council of UPOV, only documents that have been adopted by the Council of UPOV and that have not been superseded can represent UPOV policies or guidance.

This document has been scanned from a paper copy and may have some discrepancies from the original document.

_____

Avertissement: sauf si le Conseil de l'UPOV en décide autrement, seuls les documents adoptés par le Conseil de l'UPOV n'ayant pas été remplacés peuvent représenter les principes ou les orientations de l'UPOV.

Ce document a été numérisé à partir d'une copie papier et peut contenir des différences avec le document original.

_____

Allgemeiner Haftungsausschluß: Sofern nicht anders vom Rat der UPOV vereinbart, geben nur Dokumente, die vom Rat der UPOV angenommen und nicht ersetzt wurden, Grundsätze oder eine Anleitung der UPOV wieder.

Dieses Dokument wurde von einer Papierkopie gescannt und könnte Abweichungen vom Originaldokument aufweisen.

_____

Descargo de responsabilidad: salvo que el Consejo de la UPOV decida de otro modo, solo se considerarán documentos de políticas u orientaciones de la UPOV los que hayan sido aprobados por el Consejo de la UPOV y no hayan sido reemplazados.

Este documento ha sido escaneado a partir de una copia en papel y puede que existan divergencias en relación con el documento original.

UPOV

BMT/4/8
ORIGINAL: English
DATE: January 28, 1997

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

WORKING GROUP ON BIOCHEMICAL AND MOLECULAR TECHNIQUES AND DNA-PROFILING IN PARTICULAR

Fourth Session
Cambridge, United Kingdom, March 11 to 13, 1997

SIMILARITY, CLUSTERING AND DENDROGRAMS

Document prepared by experts from the United Kingdom

1. Distance, Association, Similarity and Dissimilarity Measures

1.1 Introduction

The success of agriculture has relied on exploitation of diversity in crop plants. This diversity is used not only to try to achieve maximum benefit, but also to allow different crop types to be distinguished. As agriculture has developed, considerable effort has been made in order to produce crops to suit current requirements. To permit recompense or even profit from such work, legislation governing plant breeders' rights is in place in most countries, providing a system similar in many ways to patenting. Part of the basis of this system is the ability to identify differences between varieties so that the varieties can be distinguished one from another. The distinctness, uniformity and stability (DUS) tests carried out in assessing varieties rely on achieving a balance between declaring every plant distinct (and therefore declaring a variety non-uniform) and being able to distinguish between varieties. As the number of varieties increases greater resolution between varieties is required without making uniformity unnecessarily difficult to achieve, and without using characters that show undue variation in different environments. Statistical techniques have been used since the outset of this work, and are constantly under review. Appropriate combinations of statistical techniques and plant characters facilitate successful DUS testing (i.e. help meet the aims of DUS). This document discusses a variety of ways in which the differences or similarities between varieties may be quantified. Discussion of how these measures can be exploited is dealt with in other papers (see section 2 and paper by Piepho).

1.2 Varietal Variation

1.2.1 Source of Variation

To establish an appropriate method of quantifying differences or similarities between varieties we need to consider the biological basis of those differences. Primarily it is genetically based differences that are of interest for DUS work. Interactions between genotype and environment also merit consideration if the necessary conditions are easy to control (e.g. germination at two different temperatures). Within a crop different levels of genetic variation are present, variation within plants, variation between plants within a variety and variation between varieties. The variation present at each of these levels will depend on the breeding system of the crop in question, and on methods used by plant breeders. In a strongly outbreeding crop any variation within a variety will be distributed between the two classes of homozygotes and the heterozygotes. In a strongly inbreeding crop varietal variation is maintained as homozygotes of different genotypes.

1.2.2 Measurement of Variation

Many different types of characters can be measured for DUS. These include morphological, agronomic and molecular data. In order to analyse these data to best advantage, as much use as possible of the known biology of the characters should be made. Measurements of any type may be genetic in nature if the observed phenotypes can be converted to genotypes. Translation of measurements to allele frequencies at different loci or to DNA sequence information allows best possible use of character information. Specific diversity and distance measures exist for genetic data (see below). Using data in this way may be likened to studying the factors underlying the observed variation.

1.3 Measures

1.3.1 Types of Measure

A wealth of distance, association, similarity and dissimilarity measures exist in the literature - far too many to deal with here. This apparent excess of measures arises because it is often appropriate for a "new" measure or modification of an existing one to be used for a specific application. Many of the more commonly used measures are given in the appendix (the lists of synonyms are not exhaustive).

The measures may be conveniently grouped by the type of data for which they suitable. Data may be dichotomous (binary, presence or absence), qualitative (multi-state) or quantitative in nature. Some measures have analogues for each of these classes of data. The concept of similarity between different organisms is easy to grasp, and intuitively linked to dissimilarity. Producing a number that adequately summarises the similarity or difference between two organisms is rather more difficult. Our concept of similarity is also affected by context. For example if we consider all plants, or even all crop plants we would think of wheat and barley as being similar. If we consider different varieties of wheat, a rogue barley variety would be considered very different. Distance and similarity are relative measures. We might consider the relative measurement absolute if we were able to compare all possible characters. A number of organisms with small genomes have had their DNA or RNA sequenced entirely (e.g. AIDS viruses). Individuals may be compared on the basis of their entire genome. This is not likely to be the case for crop plants for some time despite the projected increases in sequencing speed.

1.3.2 Distance, Similarity and Dissimilarity

The convenience of presentation of character score information in graphical form reinforces the idea of distance and similarity (or dissimilarity) being related to one another. The measurement of distance between individuals can be illustrated by consideration of Euclidean distance. If one considers two characters, character states or scores for each character may be plotted on a two dimensional graph. Each individual will be represented by a point. The distance between a pair of points represents the distance between a pair of individuals, and can thus be measured. When there are three characters to be considered, this can be extended to three dimensions and, although difficult to visualise, when there are *n* characters the distance between two individuals may be measured in *n*-dimensional space. In many cases it is possible to express a similarity in terms of a dissimilarity measure. For example, for a similarity coefficient where $0 \leq s \leq 1$, there is an associated dissimilarity such that $d = 1 - s$. Similarity and dissimilarity measures may be thought of as complementary to one another.

1.3.3 Metric Measures

If we consider a particular character for three individuals A, B and C we can calculate the pair-wise distances between them D_{AB}, D_{AC}, D_{BC}. A distance measure can be said to be metric if $D_{AB} + D_{AC} \geq D_{BC}$ for all three-way combinations of individuals. Initially this seems generally true, but if negative values are present or if a measure is asymmetric (such that the distance from A to B is not the same as that from B to A), the measure will be non-metric. It is not essential that the measures used are metric, but the property can be useful.

1.3.4 Correlation Coefficients and Angular Coefficients

Correlation coefficients range from -1 to +1 and are therefore non-metric. They may be regarded as special cases of angular coefficients. These measures do not imply that all values being compared are identical, and are therefore not strictly similarity measures. They are, however, sometimes useful in comparing shapes such as the profiles obtained from HPLC or densitometry. Other commonly used coefficients include Pearson's product-moment correlation, Spearman's and Kendall's rank correlations, the percent similarity coefficient and the cosine shape coefficient.

1.4 Type of Data and Coding

The data associated with different characters may be of different types. Sometimes the way in which data is collected will also affect how it is subsequently used. Table 1 shows different ways in which this might be done for the length of hair covering a particular organism.

In any particular instance there may be an obvious choice. This will depend on the distribution of the character for the individuals being considered, the biology of the character, and for any applied use, the practicality of obtaining the data. The first three rows of Table 1 show how hair length might be measured. If individuals have hair of different lengths then it might be best to measure the length of the hairs. This may be impracticable, so treating the data as categorical data may be necessary (although some information may be lost). In the second case in Table 1 it is easy to see that the categories can be arranged in a meaningful order. This is not always the case.

Description	Type	Scores
Individuals either have or lack hair	Dichotomous or Binary	0/1 or +/- or present/absent
Individuals have long short or no hair	Qualitative or Multi-state	long/short/none
Individuals have different hair lengths	Quantitative or Continuous	measure of length of hairs on individual

Table 1 Hair length measurement

The three different examples of qualitative measurements of hair colour in Table 2 illustrate this. In the first case (black/grey/white) the colours may be arranged in a logical order. In the second case (red/yellow/orange/white) no single "correct" arrangement of these colours exits. If the character is divided into two - the presence or absence of red and of yellow pigments in the hair - two dichotomous characters are created which may be a better reflection of the observations. In the third case (red/yellow/blue) it is not possible to arrange the different classes into a sensible order. The characters might be reclassified as three separate dichotomous variables, and individuals scored for the presence of blue, red and yellow pigments. In this and many other cases, further investigation of the nature of the character under investigation gives valuable insight into appropriate representation or coding. Where characters correspond to different genotypes, it may be more appropriate to determine the genotype and use that to represent the character states. Further consideration of this is given below in the section on genetic measures.

Description	Type	Scores
Individuals have black or white hair	Dichotomous or Binary	0/1 or black/white
Individuals have one of several hair colours (black/grey/white)	Qualitative or Multi-state	black/grey/white
Individuals have one of several hair colours(red/yellow/orange/white)	Qualitative or Multi-state	red/yellow/orange/white
Individuals have one of several hair colours (red/yellow/blue)	Qualitative or Multi-state	red/yellow/blue
Individuals have hair varying between black and white	Quantitative or Continuous	measure on scale of different grey levels

Table 2 Hair colour measurement

1.4.1 Recoding Quantitative Measures

Quantitative measures can be expressed as single numerical values. Such data can clearly be ordered. It is possible to recode this information, although some of the information present in the original data may be lost. By subdividing a continuous range into a series of groups the data can be treated as multistate, or by reduction to two groups, as dichotomous. The data may also be reduced to multistate and then recoded to produce a series of dichotomous characters. Coding can be carried out to reflect information about the character in question e.g. additive, multiplicative etc. Choice of an appropriate number of groups in reduction to multistate data can result in retention of the bulk of the original information. It is worthy of note that recoding quantitative characters as dichotomous may distort distance measures.

1.4.2 Choice of Data

The use of rare and of common characters have both been advocated by different workers. Comparison of these two different approaches suggests that they make little difference to the final result.

The inclusion of non-variable characters in a data set does not increase the resolution of the measures obtained but merely rescales them. It may be argued that the resulting measures reflect the underlying absolute differences more accurately. The approach may be attractive when we are able to measure the genetic differences between varieties more completely.

1.5 Standardisation, Scaling, Transformation and Weighting

1.5.1 Standardisation

The magnitude of many of the measures mentioned in the appendix increase as the number of characters measured increases. Standardisation provides a method for overcoming this, and also for converting these measures to metrics. The following examples concentrate on the Euclidean distance measure $\left[\sum_i (x_i - y_i)^2\right]^{1/2}$

Standardised by range (Taxonomic distance) $\left[\sum_i \left((x_i - y_i) / \text{range}\right)^2\right]^{1/2}$

Standardised by standard deviation $\left[\sum_i \left((x_i - y_i) / \hat{\sigma}\right)^2\right]^{1/2}$

Average or Average Taxonomic Distance $\left[\sum_i (x_i - y_i)^2 / n\right]^{1/2}$ (Heincke 1898)

Mean Character Difference $\left[\sum_i (x_i - y_i)^2\right]^{1/2} / n$

Cosine θ or normalised Euclidean $\left[\sum\left(\frac{x_i}{\left(\sum_i x_i^2\right)^{1/2}} - \frac{y_i}{\left(\sum_i y_i^2\right)^{1/2}}\right)^2\right]^{1/2}$

1.5.2 Scaling

If several variables are measured so that they have different units they must either be scaled or weighted before proceeding with further calculations. Scaling may be by equalisation of the gross size of each character, by equalisation of the variability of each character or a combination of the two. A similar effect to scaling can sometimes be achieved through appropriate weighting.

1.5.3 Transformation

In common with many other statistical methods, maximum use can be gained from data if it is transformed correctly. Transformation may be valuable in providing a biologically and mathematically sensible form of data representation. The size of an organism may be best represented by its volume. Length may be better represented on a logarithmic scale. Many other forms of data transformation are available to aid in the presentation of data e.g. Fourier series and non-linear fitting.

1.5.4 Weighting

Except as an alternative method of scaling, weighting should be unnecessary. If characters are represented as closely as possible by their genotypes, appropriate weighting should result by virtue of the numbers of genes involved in complex phenotypic features. Some form of weighting may occasionally be appropriate for characters that show environmental interactions.

1.5.5 Missing Values

Missing data should be coded differently from the absence of a character. Providing an individual or a particular character does not have too many missing values calculation of the coefficient is carried out with appropriate adjustments. It is worthy of note that missing data may change the properties of the measure or metric.

1.6 Generalised Distances

In applied biology it is rare to find that all of the variables or characters being studied are of the same type. There is usually a range of different types of variables for which different measures of dissimilarity are appropriate. One way around this problem is to consider each class of variable separately, and then to take an average (with or without weighting) of the resulting coefficients. Another option is to use a generalised measure. Gower has defined a general coefficient of similarity which is suitable for all data types. This measure may be weighted, although determining appropriate weights may not be straight forward.

1.7 Other Techniques

Formation of a distance or similarity measure from character data is generally the first step of three. It is generally followed by a clustering technique and then by a form of graphical representation (see section 2). This is essentially an algorithmic approach. Once appropriate algorithms have been selected, computation proceeds to a final solution. Only one "answer" is produced. (For some clustering methods different solutions may sometimes be produced by changing the data order.) Another class of methods are those used in phylogenetic reconstruction. These methods are based on evolutionary assumptions which are used to search for an optimum tree. Every possible arrangement of tree can be considered and rated according to the evolutionary assumptions used. Thus each tree can be ranked to determine the "best" tree for a given set of rules. Maximum likelihood methods estimate the probability that a given tree gave rise to the current observed data and seem both reasonably robust and intuitive. Parsimonious methods seek to minimise the numbers of particular changes (the changes depend on the model involved). These techniques are more sensitive to violations of their assumptions. Investigation of the use of these techniques to study varietal information is clearly merited.

1.8 Conclusion

To get the best results, all information known should be utilised. This is important when choosing an appropriate scale, transformation and distance measure for an individual character, a group of similar characters or a range of diverse characters. Although quantitative characters can be simplified to one or more two state characters, information is lost, decreasing potential resolution. Where possible translation of phenotype scores to genotypes allows a better characterised analysis. Where this is not possible the exact choice of distance measure will depend on the nature of the data. Where data from different characters is of different types, generalised approaches like that of Gower are to be preferred, although appropriately weighted combinations of different measures may be made. Particular care should be exercised when computer packages are used as these may provide unsuitable default options, or lack suitable options entirely.

2. Clustering

2.1 Introduction

In general clustering tools exist to allocate a sample of N objects into G groups based on the measurements of P variables. Clustering tools can be considered as the filling of a sandwich - the outer layers being on the one hand the of choice of similarity or distance measures and on the other hand the dendogram or output from cluster tools.

Choice of similarity or distance measure
Choice of clustering tool
Interpretation of dendogram or output

The choice of the similarity or distance measure used can be driven by the type of data (binary, qualitative, quantitative, mixed), the quantity of the data and the scales of measurement.

Before employing a clustering tool it is essential to consider the objectives of the 'grouping'. Some primary questions are given below - not exhaustive.

- Objective to seek natural grouping only?
- Are groups/clusters of known shape sought?
- Are known 'controls' available to mark 'groups'?
- Are proposed clustering method appropriate to the (biological) mechanism that generated the data?
- Are hierarchical 'tree/branch' methods appropriate.?
- Are methods of density search, clumping and partitioning appropriate?
- How many groups are desired?
- Are overlapping groups allowable?
- Has the data been screened for outlier values?

2.2 *Agglomerative Methods*

This class of clustering tool are usually hierarchical in form and can be influenced by a poor initial starting 'group'. No reallocation of objects to groups is possible which may be important if the identification of natural groups is the primary aim. Chaining (see below) can occur in these methods. Some authors have questioned some mathematical properties of these methods under transformation.

Despite the above comments hierarchical clustering tools are extensively used - particularly as software is readily available to perform the computations. Listed below are a sample of hierarchical methods that differ in the decision rule applied when selecting existing groups to be fused together. Application of each of the methods outlined below are unlikely to generate exactly the same results. Differences between results derived from a range of methods can provide useful additional information about the distribution of objects within a group, the overall shape of clusters and the distribution of the clusters over the 'sample space'.

2.3 *Clustering Methods*

Many clustering methods have been developed over the past 30 years. Unless the 'clusters' are compact and regular in shape and well separated in space - it is unlikely that identical results will be obtained from all the methods listed below. This can be unsettling for users especially when the resulting clusters are very different. Some explanation of the decision rule used by the methods will reassure users that any differences are providing important information about the structure of the data - eg the distribution of objects within a group, the overall shape of clusters and the distribution of the clusters over the 'sample space'. Deviations from compactness or regularity or both can have marked effects on the resulting clusters. The detection of potentially aberrant observations is advised before clustering as such observations can greatly influence the resulting clusters. Application of clustering methods can be used to identify outliers. Removal of outliers should require that the selected distance measure is recalculated.

Using many clustering methods on the same data sets and picking the one that is 'best' (in some sense) is not recommended.

2.3.1 Nearest Neighbour (Single Linkage)

Decision Rule:- Minimise the inter-group distance defined as the distance from the closest member. This is illustrated diagramatically for three clusters in Fig 1. Nearest Neighbour method works well for regular and compact natural clusters. The method can be influenced by outliers and perform poorly in some cases of clearly distinguishable groupings with little separation in the data space (see Fig 7.)

2.3.2 Furthest Neighbour (Complete Linkage)

Decision Rule:- Minimise the inter-group distance defined as the distance from the remotest member. This is illustrated diagramatically for three clusters in Fig 2. Furthest Neighbour method works well for regular and compact natural clusters. The method can be influenced by outliers.

2.3.3 Centroid Cluster (UPGMC - unweighted pair group)

Decision Rule:- Minimise the inter-group distance defined as the distance from group Centroid and proportional to the size of the groups. This is illustrated diagramatically for three clusters in Fig 3. UPGMC method works well for regular and compact natural clusters. Although <u>all</u> the data is utilised in the computation of the centroid - similar centroids can stem from both compact and disparate within cluster data.

2.3.4 Median Cluster - Gower's Method (WPGMC - weighted pair group)

Decision Rule:- Minimise the inter-group distance defined as the distance from group Centroid independent of the relative sizes of the clusters. Weighted version of Centroid Cluster (3.3 above). WPGMC method works well for regular and compact natural clusters. This is illustrated diagramatically for three clusters in Fig 4. Can be applied to different types of data.

2.3.5 Group Average Cluster (UPGMA unweighted pair group average)

Decision Rule:- Minimise the inter-group distance defined as the average of all paired distances based on all individuals with the group. This is illustrated diagramatically for three clusters in Fig 5. UPGMA method works well for regular and compact natural clusters. Each observation, irrespective of location relative to the within group centroid, is given equal weighting when calculating the set of pairwise distances.

2.3.6 Ward's Method - Orloci (error sum of squares)

Decision Rule:- Minimises the increase in sum of squared deviations for each individual from the group Centroid This is illustrated diagramatically for three clusters in Fig 6. Ward's method works well for regular and compact natural clusters. The distribution of the within group observations relative to the group centroid is analogous to the familiar least squares approach used in analysis of variance/regression.

Various authors have recommended that Median, Wards and Centroid methods are not suitable for similarity coefficients. Not all the above methods will deal equally effectively with clusters of the distinctive shape as shown in the artificial examples in Fig 7.

While users may seek self contained and compact clusters the mathematical foundations of the applied clustering decision rules can yield chains of points where objects are added to clusters via close intermediate points. This is referred to as "Chaining". Some hierarchical methods have been shown to be prone to this effect.

As mentioned earlier hierarchical clustering may be appropriate when the data suggests a 'nested' structure for the relationships between the clusters.

2.4 Non-Hierarchical Methods

2.4.1 Divisive Methods.

Many such methods have been suggested in the literature including:-

- Splinter Group Methods.
- Association Analysis Method for binary data.
- Automatic Interaction Detector (AID).

This class of tool seek to divide the N objects into G groups but suffer from a similar fault as the hierarchical methods - that of failing to 'recover' from a poor initial grouping.

2.4.2 Partitioning Methods

If W is the pooled within group sum of squares and products matrix(SSP) then methods can be devised to partition the N objects by minimising the trace(W). Many such methods require *a priori* the number of clusters required. While theoretically a plot of trace(W) against the number of groups (or clusters) is expected to indicate the appropriate number of groups for a particular data set - the results of using such an approach in practice have been disappointing.

2.4.3 K-Cluster Means Methods.

This class of methods bases the clustering criterion on such statistics as:-

- trace(W)
- det|W|
- trace (BW^{-1})
- Average Entity Stability statistic
- Information Statistics.

2.4.4 Density Search Methods

This class of method seeks to identify dense and sparse portions of the total data space.
An initial radius R is chosen and a circle drawn round each of the N object data points. Circle around data points that contain K or more points are referred to as "dense" points. Radius R is increased gradually, which will increase the number of "dense" points until the various stopping rules are met.
This approach is illustrated in Fig 8.

3. Dendograms

Dendograms are ways of representing object/cluster relationships visually. They are usually produced following hierarchical clustering based on a suitable distance/similarity measure. This tool is not generally applicable to non-hierarchical clustering methods or for data that is not hierarchical in nature.

Dendograms are nodal systems and when viewed vertically can be considered akin to a child's mobile. A series of nodes suspended from "legs" which are in turn attached to "arms" passing through nodes. The lengths of the "arms" is such to achieve overall equilibrium and therefore is not important to the interpretation of the overall dendogram. Nodal position relative to the root and the length of the "legs" are important as they correspond to the degree of relatedness of the groups of objects/clusters.
Dendograms, unlike a child's mobile, are not restricted to being 'hung' vertically and can equally well be drawn horizontally. In a similar way that a mobile is free to rotate about any node, independently of other nodes but retaining overall equilibrium, Dendograms can also be rotated. Lower level nodes can also be rotated or remain fixed. Such a feature can offer a great deal of flexibility in the way any dendogram is viewed. The potential application of this feature is shown in the example Fig 9. With only six objects the initial dendogram shows A and F at opposite sides of the visualisation whereas after only three rotations A and F become adjacent. As can be seen dendograms do not produce unique visualisations of the observations/clusters.

A child's mobile is unlikely to remain in equilibrium if nodes are removed. Similarly nodes in dendograms should not be removed. The distances/similarities should first be recomputed and the appropriate clustering tool reapplied.

From a users perspective there is a tendency to accept a dendogram if it produces a visualisation that supports our *a priori* expectation. The focus of dendograms is also on the objects and less on the clustering of the objects which are fused into clusters at differing stages.

Several methods are proposed to give an objective assessment of the efficiency of a dendogram for 'known' solutions - e.g. using training data sets. Dendogram efficiency can be measured as the minimum number of node rotations which gives the optimal match with the 'known' solution. Alternatively the minimum number of rotations required to achieve complete randomness.

3.1 Alternatives to Dendograms

Given some of the limitations of dendograms mentioned above users are encouraged to consider a visual representation that focuses more on the relatedness of the formed clusters.

Sneath and Sokal (1973) give several possibilities that can be tailored to suit individual requirements. Some examples, based on the same data set, are given below.

- Contour intervals compared to the traditional dendogram. (Fig 10)
- Contours with minimum spanning 'tree' showing cluster grouping (Fig 11)
- First and second order distances (Fig 12)
- 'Ball and rod' method (Fig 13)

Other methods, such as dimension reduction techniques, can also be applied successfully particularly where much of the total information is contained in the first few 'dimensions'.

4. References

Sneath, H. A. And Sokal, R. R. (1973) Numerical Taxonomy. Freeman and Company. San Francisco

5. Appendix

5.1 Measures

The following three sections mention some of the more commonly used resemblance coefficients for quantitative and qualitative data. Qualitative data is considered as multistate data and as dichotomous or binary data. The coefficients mentioned are neither exhaustive nor are all synonyms listed. Some algorithms have anologues in each section. Despite this they may not bear the same name. Standardization and scaling have been incorported in some cases where a particular name is associated with the use of a particular type of standardization or scaling. For many of the measures standardization is necessary as the distances increase as more characters are added. Different forms of standardisation are considered below.

5.1.1 Measures for Quantitative Data

The following are measures that have been used to study quantitative data.

5.1.1.1 Power

The power distance can be regarded as general form of several popular distance measures. Use of different constants a and b allow modification of the coefficient for different purposes.

$\left[\sum_i |x_i - y_i|^a\right]^b$ where a and b are user-defined parameters

5.1.1.1.1 Minkowski

A specialized case of the above, where $a = 1/b$. As a increases, dissimilar units make a greater contribution.

$\left[\sum_i |x_i - y_i|^a\right]^{1/a}$ where a is a user-defined parameters. This measure is sometimes standardised by range.

5.1.1.1.2 Euclidean or Pythagorean or Taxonomic Distance

This is a specialized case of the Power and Minkowski metrics where $a = 1/b = 2$. It is perhaps the most widely used of the distance measures and is readily visualised in two dimensions where it corresponds to the distance between two points.

$$\left[\sum_i (x_i - y_i)^2\right]^{1/2}$$

5.1.1.1.3 Squared Euclidean $\sum_i (x_i - y_i)^2$

5.1.1.1.4 Manhattan or City Block

This is another specialized case of the Power and Minkowski metrics where $a = b = 1$. This is another very popular metric, and for a two dimensional case can be thought of as the distance between two points on the X axis plus the distance between the same two points on the Y axis.

$$\sum_i |x_i - y_i|$$

This measure is also frequently standardized by range by averaging, or by both. When averaged this measure is sometimes referred to as the Mean Character Difference (Cain & Harrison 1958, Czekanowski 1932, Haltenorth, 1937). The measure will always underestimate the true Euclidean difference; underestimation may be considerable wheresome differences are small and others large.

5.1.1.1.5 Ecological

This is like the City block measure, but has a weight of zero if $x_i = y_i$.

5.1.1.2 Bhattacharyya or Chord Distance

$$\left[\sum_i (x_i^{1/2} - y_i^{1/2})^2\right]^{1/2}$$

5.1.1.3 Canberra

$$\sum_i \frac{|x_i - y_i|}{(x_i + y_i)} \quad \sum_i \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

5.1.1.4 Unnamed

$$\sum_i \frac{|x_i - y_i|}{|x_i + y_i|}$$

5.1.1.5 Bray-Curtis

$$\frac{\sum_i |x_i - y_i|}{\sum_i (x_i + y_i)}$$

5.1.1.6 Czekanowski

$$1 - \frac{2\sum_i \min(x_i, y_i)}{\sum_i (x_i + y_i)}$$

5.1.1.7 Clark or Divergence

$$\left[\frac{1}{n}\sum_i \left(\frac{x_i - y_i}{x_i + y_i}\right)^2\right]^{1/2}$$

5.1.1.8 Chebychev

$$\mathrm{Max}\,|x_i - y_i|$$

5.1.1.9 Chi-square

$$\left(\sum_i \frac{(x_i - y_i)}{\sum_i x_i}\right)^{1/2}$$

5.1.1.10 Penrose Size

$$\left[\sum_i (x_i - y_i) / n\right]^2$$

5.1.1.11 Penrose Shape

$$\sum_i (x_i - y_i)^2 / n - \left[\sum_i (x_i - y_i) / n\right]^2$$

5.1.1.12 Soergel

$$1 - \frac{\sum_i |x_i - y_i|}{\sum_i \max(x_i, y_i)}$$

5.1.1.13 *Ware and Hedges*

$$\sum_i \left(1 - \frac{\min(x_i, y_i)}{\max(x_i y_i)}\right) / n$$

5.1.1.14 *Mahalanobis*

$\sqrt{\sum_i \sum_j V'_{ij}(\bar{x}_i - \bar{y}_i)(\bar{x}_j - \bar{y}_j)}$ where V'$_{ij}$ is the *i*the row of the *j*th column of the variance-covariance matrix between two populations.

5.1.2 Measures for Qualitative Data

Where variables are catagorical, but with more than two classes. In the explainations below *m* indicates a match, *u* unmatched (or mismatched), and *n* the total number of matches plus mismatches.

5.1.2.1 *Sneath or Simple Match Distance*

$$1 - (m / n)$$

5.1.2.2 *Rogers and Tanimoto Distance*

$$1 - (m / n + u)$$

5.1.2.3 *Harmann Distance*

$$1 - (m - u / n)$$

5.1.2.4 *Unnamed No1 Distance*

$$1 - (2m / n + m)$$

5.1.2.5 *Unnamed No3 Distance*

$$1 - (m - u)$$

5.1.2.6 *Pattern Difference*

The pattern difference can be extended to multistate characters (See below and Sackin M.J., J Gen Micro 122: 247)

5.1.3 Measures for Dichotomous Data

Considering the scores for two individials:

Individual A	Individual B	Frequency
+	+	a
+	_	b
_	+	c
_	_	d
Total		n

5.1.3.1 *Simple Match & Taxonomic Distance*

The proportion of variables that show disagreement between two individuals. Although not always ideal, this measure may be acceptable in many situations. The Taxonomic distance is the square root of the simple matching distance.

$$\frac{b+c}{n} \left(\text{or } 1 - \frac{a+d}{n}\right)$$

5.1.3.2 *Jaccard Distance*

Particularly suitable where the absence of a trait does not simply relate to the degree of similarity between individuals eg a lack of gills does not relevant when comparing make a tree to a mammal.

$$\frac{b}{a+b+c}\left(\text{or } 1-\frac{a}{a+b+c}\right)$$

5.1.3.3 Czekanowski or Sørensen or Dice Distance

Similar to the Jaccard algorithm, but giving extra weight to positive matches

$$\frac{b+c}{2a+b+c}\left(\text{or } 1-\frac{2a}{2a+b+c}\right)$$

5.1.3.4 Yule Distance

Yule's non-metric similarity measure may be converted into a 0 - 1 scale by the transformation (s + 1)/2, and into a distance by subtracting the result from 1.

$$\frac{bc}{ad+bc}\left(\text{or } 1-\frac{1}{2}\left(\frac{ad-bc}{ad+bc}+1\right)\right)$$

5.1.3.5 Kulczynski No 1

Values range from 0 to infinity. Sensible values are not obtained after a > 1/2.

$$\frac{a}{b+c}$$

5.1.3.6 Kulczynski No 2

$$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$$

5.1.3.7 Pearson's Phi

Pearson's Phi is a non-metric similarity measure. It may be converted into a 0 - 1 scale by the transformation (s + 1)/2, and into a distance by subtracting the result from 1.

$$\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

5.1.3.8 Russel and Rao Distance

$$\frac{b+c+d}{a+b+c+d}\left(\text{or } 1-\frac{a}{a+b+c+d}\right)$$

5.1.3.9 Anderberg 1 Distance

$$\frac{2(b+c)}{a+2(b+c)}\left(\text{or } 1 - \frac{a}{a+2(b+c)}\right)$$

5.1.3.10 Anderberg 2

$$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right)$$

5.1.3.11 Unnamed No5

$$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

5.1.3.12 Ochiai

$$\frac{a}{\sqrt{(a+b)(a+c)}}$$

5.1.3.13 Pattern Difference

A shape coeffieient known as the pattern coefficient has been developed to aid in the classification of bacterial identification. In this application a negative result may be due to sub-optimal conditions or be a true negative result. For a binary character the vigour is positive test results/total test results. The pattern difference between two individuals is calculated as:

$\sqrt{\left(\text{Difference in vigour}^2 - \text{Match Distance}^2\right)}$ where

$$\text{Match Distance} = \sqrt{\left|\left(\frac{a+b}{n}\right)^2 - \left(\frac{a+c}{n}\right)^2\right|}$$

5.1.3.14 Rogers and Tanimoto

See also below

$$\frac{b+c}{a+2(b+c)+d}\left(\text{or } 1-\frac{a+d}{a+2(b+c)+d}\right)$$

5.1.3.15 Sneath and Sokal

$$\frac{b+c}{2a+b+c+2d}\left(\text{or } 1-\frac{a+d}{a+(b+c)/2+d}\right)$$

5.1.3.16 Hamman

Hamman's non-metric similarity measure may be converted into a 0 - 1 scale by the transformation (s + 1)/2, and into a distance by subtracting the result from 1. When converted into a distance this measure is identical with the simple match distance.

$$1-\left(\frac{a-b-c+d}{a+b+c+d}\right)$$

5.1.4 General Measures

5.1.4.1 Gower

$1-\dfrac{\sum_i(w_i s_i)}{\sum_i w_i}$ where s_i is the score and w_i the weight. The score and weight for different types of variable are shown in Table 3: This can be expressed as a distance.

Data Type	s_i	w_i
Quantitative	$1-\frac{\lvert x_i-y_i\rvert}{\text{range}}$ 0 if $w_i=0$	0 for missing values 1 for all other situations
Dichotomous	1 for matches 0 for mis-matches 0 if $w_i=0$	0 for negative matches 1 for all other situations
Qualitative	1 for matches 0 for mis-matches 0 if $w_i=0$	0 for negative matches 1 for all other situations

Table 3 Calculations for Gowers Coefficient

5.1.5 Sequence and Genetic Data and Measures

Measurements of genetic data can be of two types. Those where the mesurements can be treated as genetic measurements and those where they must be treated essentially as phenotypic data. If measurements of a character can be related to different alleles at a locus assumptions consistent with this can be made. The nature of the data used to determine which allele(s) is (are) present does not affect this, although knowledge of the genetic architecture of the organism (ploidy, duplications, etc) is clearly important. The measurements may be of morphological characters, protein size, shape or charge, or information derived from the nucleic acid sequence directly. Where this is not possible care should be taken, particularly if the information is derived from nucleic acid data. Homologous or homeologous parts of the genome should be compared For example, bands of a particular size in a genomic digest should be treated as phenotypic data unless there is some specific reason for beleiving them to show homology. In this case this might be established by studies of inheritance or be hybridization with specific complementary probes.

The nucleic acid sequence itself might be regarded as the most basic form of data, but even here care must be exercised. The genetic arcitecture of the organism(s) under study may make it difficult to determine appropriate comparisons for some sequences. As mentioned in the introductory section, a modified or new measure is often the most appropriate for particular type of data. This is particularly true for genetic data, and for nucleic acid and protein sequences. There is still some debate about the most appropriate techniques for the study of some forms of molecular varaition.

5.1.5.1 Czekanowski's Mean Difference

Czekanowski's Mean Difference may be used for the study of genetic difference. The maximum value of this measure depends on the number of alleles. Czekanowski's mean difference can also be much less than one for two polymorphic populations which share no alleles.

5.1.5.2 Manhattan

The Manhattan metric is 2 when no shared alleles are present, and can thus be adjusted to account for this by division by 2. This is sometimes known as the Prevosti distance. Neither can be used to take in to account that gene frequency change is itself dependant on gene frequencies.

5.1.5.3 Rogers

Rogers (1972) genetic distance is effectively the euclidean distance corrected to allow for the fact that for gene frequency data this would allow the distance to be between 0 and $\sqrt{2}$. The measure is not proportional to evolutionary time or the number of gene substitutions. It also suffers from the property that the distance can be considerably less than one even when no alleles are in common between two polymorphic populations.

$$\left[\frac{1}{2}\sum_i (x_i - y_i)^2\right]^{1/2}$$

5.1.5.4 Sanghivi

Sanghivi's distance (1969) is essentially a modification of Mahalanobis' D^2 modified for use with gene frequency data. This distance is closely related to χ^2, such that for two populations of size n χ^2 = $2n$ x the square of

$$\left[\sum_i \frac{(x_i - y_i)^2}{(x_i + y_i)/2}\right]^{1/2}$$

5.1.5.5 Bhattacharyya's Angular Transformation

$$\left[\frac{1}{2}\sum_i \frac{(x_i - y_i)^2}{(x_i + y_i)}\right]^{1/2}$$ This takes a value of between 0 and 1 and is distributed as the root of 1/4 χ^2

5.1.5.6 Other Genetic Measures

Other important and frequently used measures include Cavalli-Sforza & Edwards (1967) Chord Distance, Nei (1978) unbiased genetic identity; Nei (1978) unbiased genetic distance; Nei (1972) genetic identity; Nei (1972) genetic distance; Nei (1978) unbiased minimum distance; Nei (1972) minimum distance;; Cavalli-Sforza & Edwards (1967) arc distance; Edwards (1971,1974) "E" distance.

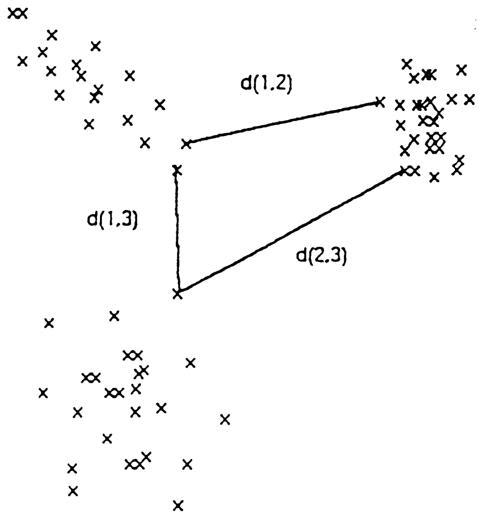

Fig 1. Nearest Neighbour (Single Linkage)

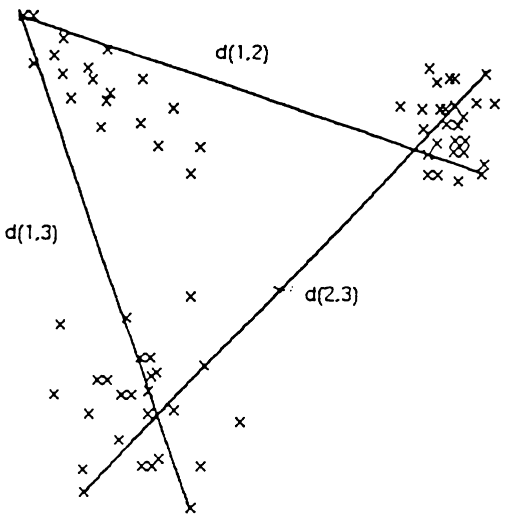

Fig 2. Furthest Neighbour (Complete Linkage)

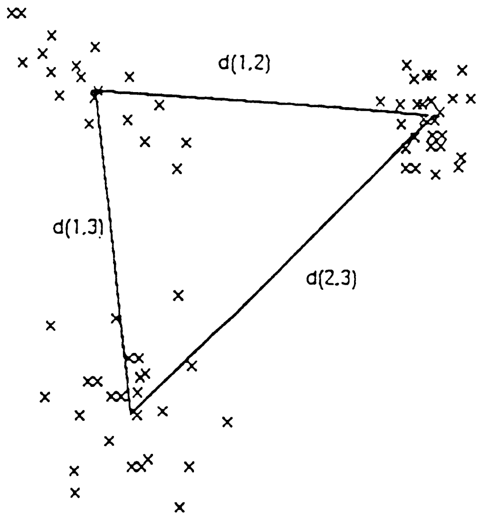

Fig 3. Centroid Cluster (UPGMC - unweighted pair group)

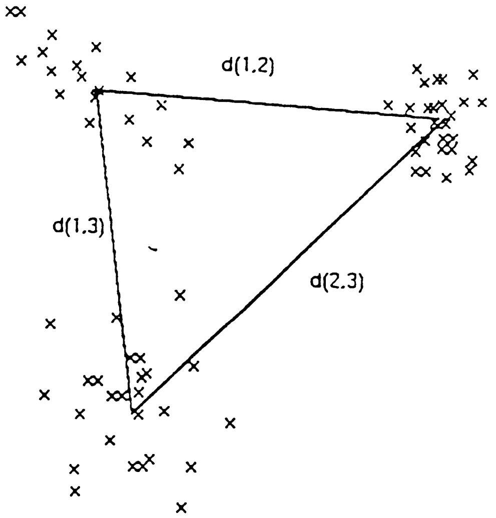

Fig 4. Median Cluster - Gower's Method (WPGMC - weighted pair group)

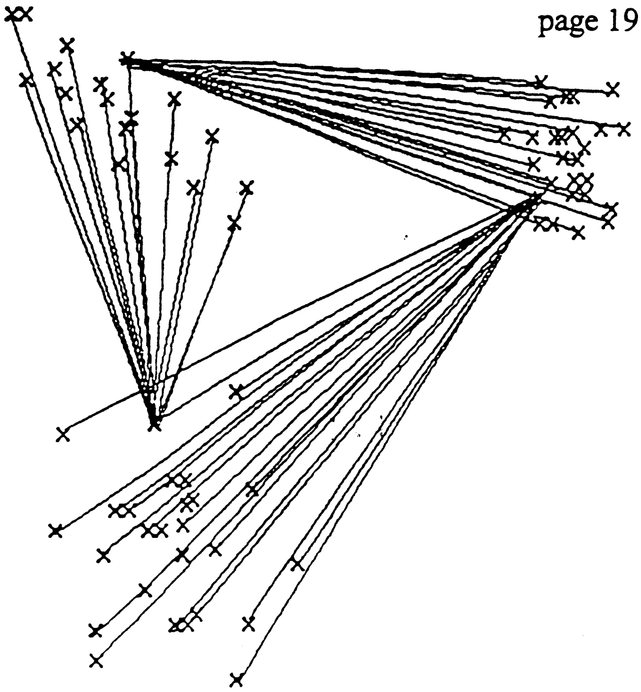

Fig 5. Group Average Cluster (PGGMA unweighted pair group average)
(only sample of lines shown)

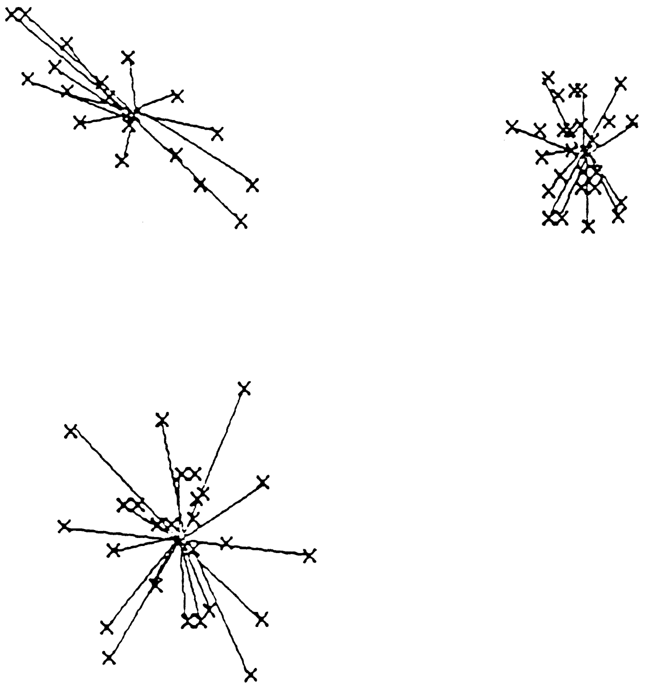

Fig 6. Wards Method (Orlocl - Error Sumof Squares)
(only sample of lines shown)

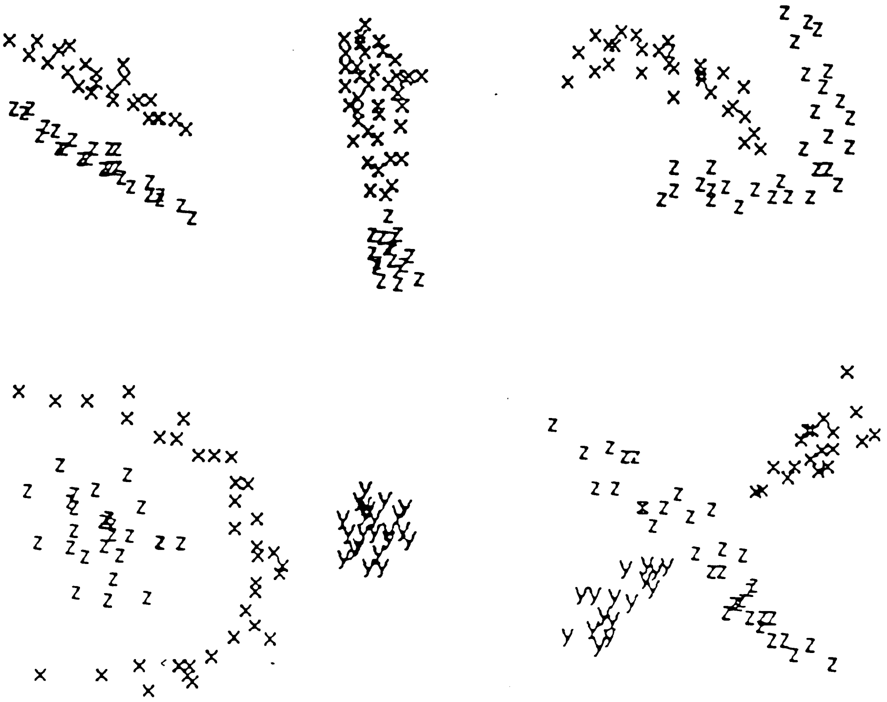

Fig 7. Clear clusters that can be difficult to detect.

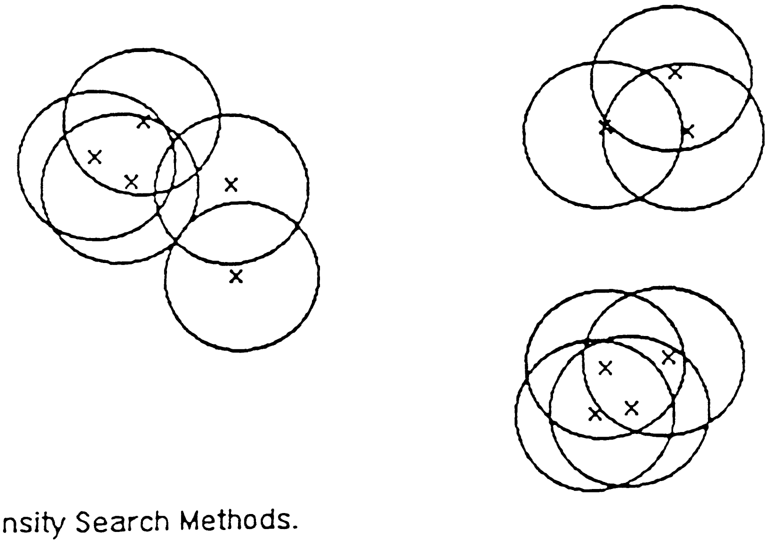

Fig 8. Density Search Methods.
(radius R)

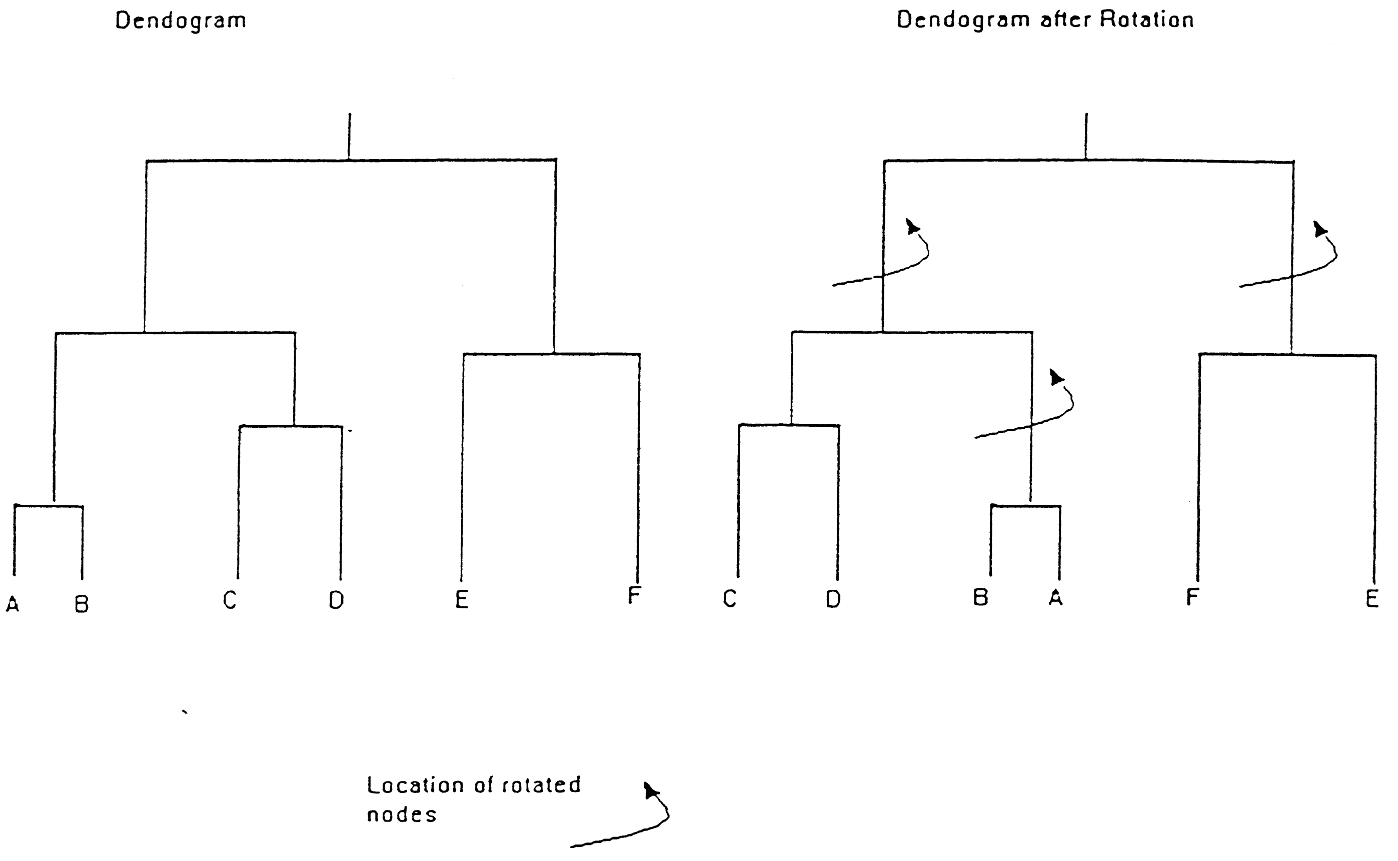

Fig 9. Example of Rotation in Dendograms

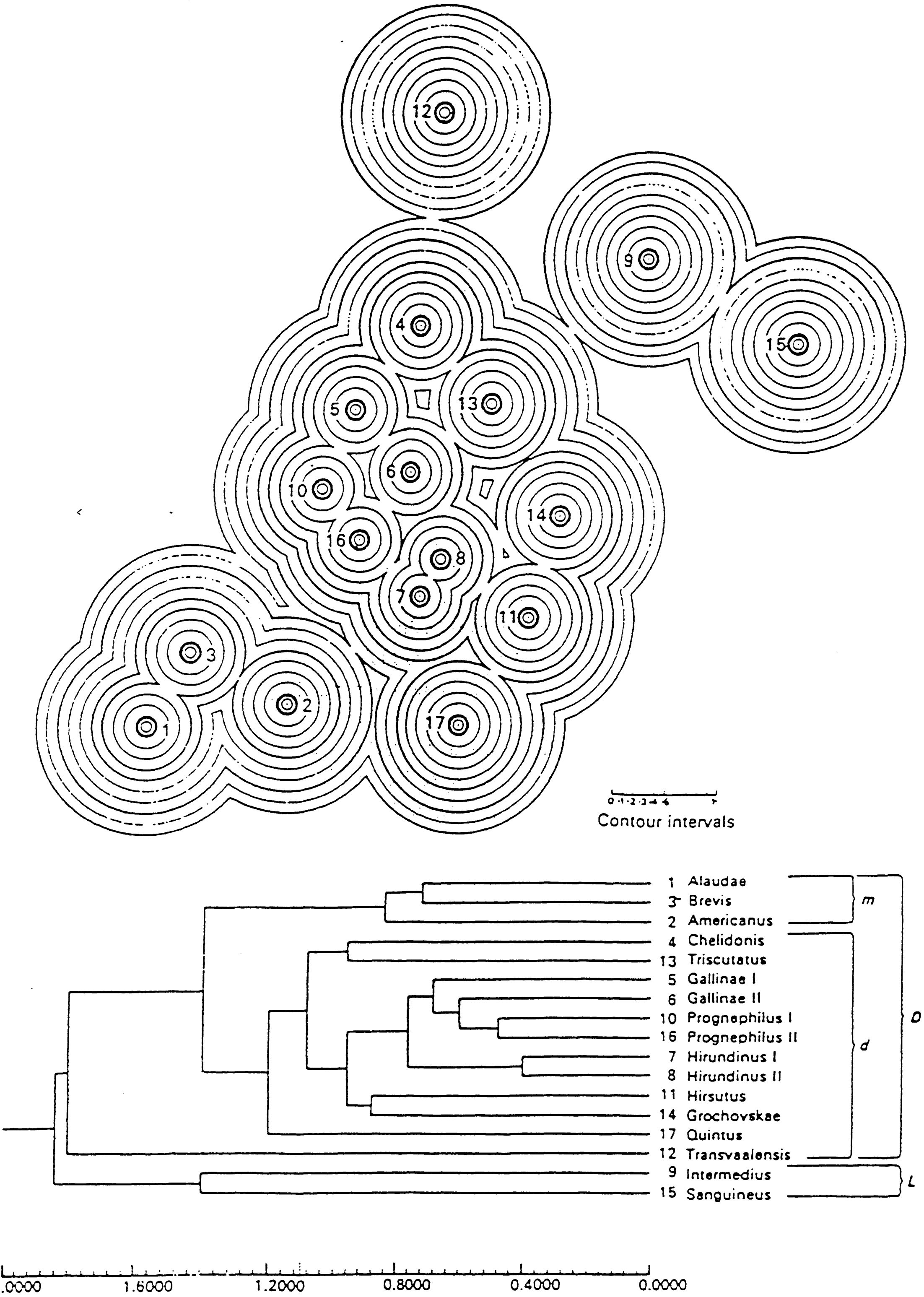

Fig 10. Contour intervals compared to the traditional dendogram.

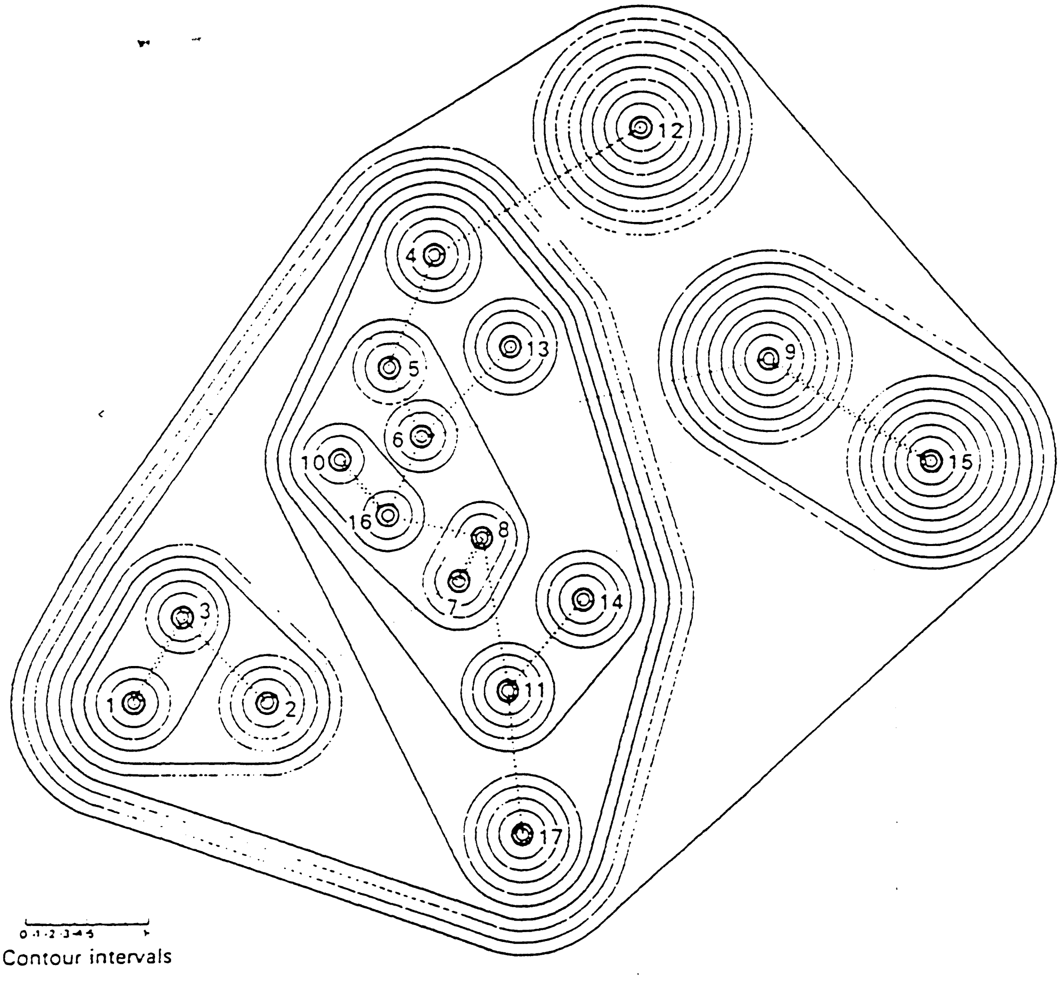

Fig 11. Contours with minimum spanning 'tree' showing cluster grouping

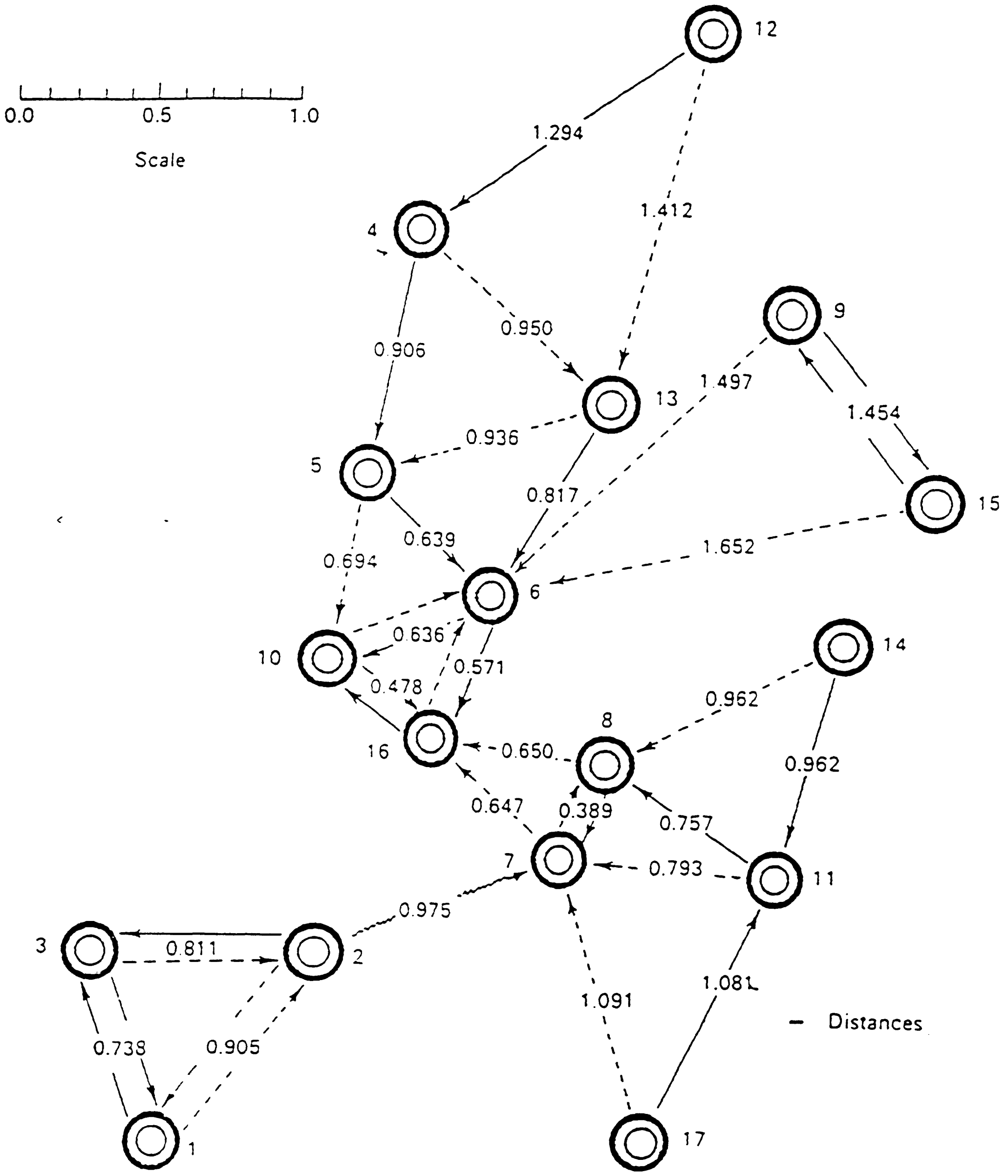

Fig 12. First and second order distances

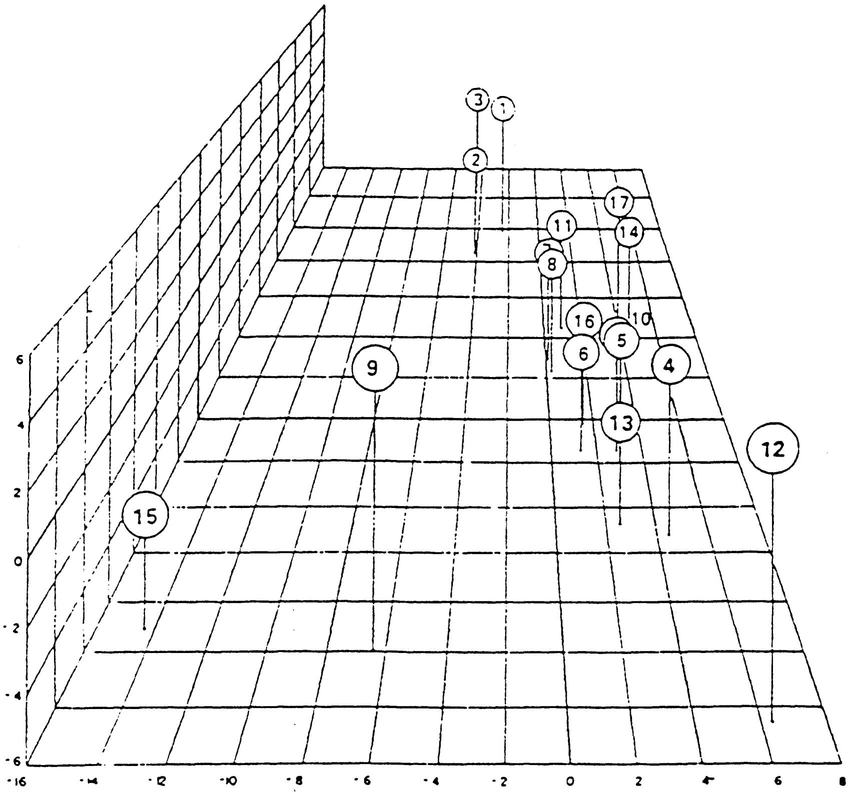

Fig 13. 'Ball and rod' method

[End of document]