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agreed by the Council of UPOV, only documents that have been adopted by the Council of UPOV and that have not been superseded can represent UPOV policies or guidance.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

BMT/4/7 Rev. 2
ORIGINAL: English
DATE: April 17, 1997

UPOV

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

A REVIEW OF METHODS FOR CLUSTER ANALYSIS OF MARKER DATA

Document prepared by experts from Germany

A REVIEW OF METHODS FOR CLUSTER ANALYSIS OF MARKER DATA

H.P. Piepho[1], F. Laidig[2]

1. Introduction

Unambiguous identification of plant cultivars and of their genetic interrelation is of relevance in DUS testing and in the identification of essentially derived (ED) varieties. Various molecular techniques are now available for varietal identification, which are more powerful than traditional morphological comparisons and isozyme techniques (Cooke RJ, 1995). For a recent review comparing polymerase chain reaction (PCR) based DNA profiling methods with the well-established restriction fragment length polymorphism (RFLP) technique see Morell et al. (1995). Statistical analysis of DNA profile data usually consists of three steps: (i) Scoring the profile; (ii) Calculating genetic distances; (iii) Summarizing genetic relationships, e.g. as a dendrogram. Dendrograms are useful for studying the genetic relationships among crop cultivars or inbred lines. The purpose of this paper is to describe the computational steps for generating dendrograms from marker data.

The type of distance measure suitable for analysing a given data set depends on the data. In the following we will therefore describe the type of data arising from DNA profiles and how to score such profiles (Section 2). Then, a brief account is given of some distance and similarity measures in widespread use (Section 3). A short description of some common clustering algorithms is presented in section 4.

2. Type and scale of marker data

Various DNA profiling methods are available such as RFLPs and PCR based DNA profiling (RAPD = random amplification of polymorphic DNA, STS = sequence-tagged sites analysis, AFLPs). Any of these methods yields a DNA profile consisting of a specific banding pattern on an electrophoretic gel. Two basic types of pattern can be distinguished: banding data and allelic data. Banding data are easier to obtain in practice, while allelic data (as derived from banding data) are more informative.

2.1 Banding data

When the banding pattern is complex and the genotype cannot be determined directly, it is common to convert the banding pattern by assigning a series of 1s and 0s representing band presence or absence, respectively. The resulting array of 1s and 0s can be used to calculate (genetic) similarities or distances.

1 Institut für Nutzpflanzenkunde, Gh Kassel, Steinstrasse 19, 37213 Witzenhausen, Germany (piepho@wiz.uni-kassel.de)
2 Bundessortenamt, Hannover, Osterfelddamm 80, 30627 Hannover, Germany

Table 1: Example of scores for banding data of two genotypes *x* and *y* (9 band positions)

Band position		*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*
Genotype	*x*	0	0	1	1	0	1	1	0	1
	y	1	1	0	1	1	1	1	0	1

The cultivars or lines to be compared may be thought of as populations. The number of individuals scored per population depends on the degree of genetic homogeneity within populations. For self-fertilized varieties, which are homogeneous, for homozygous inbred lines, and for clonal varieties, one individual may be taken to represent the whole population. By contrast, a large number of plants (100 or more) may have to be sampled from heterogeneous populations of allogamous crops. In the latter case banding patterns are expected to vary among individuals of the population. Banding data of a sample of plants from heterogeneous populations may be summarized by computing the band frequency at each banding position, where the frequency may take any value between zero and unity (Table 2). For a sample from a completely homogeneous population, the frequencies would be either 0 or 1 at a position, leading to the same data as for each individual plant. In the sequel, banding data from a single plant (that represents a homogeneous population) will be referred to as binary banding data. Frequencies derived from banding data of a sample of plants from a heterogeneous population are referred to as band frequency data.

Table 2: Example for computing band frequency data from binary band data

	Band position			
Plant no.	*a*	*b*	*c*	*d*
1	0	0	1	0
2	0	1	1	0
3	1	1	0	1
4	1	1	0	0
	0.5	0.75	0.5	0.25
	band frequencies			

2.2 Allelic data

Allelic data may be obtained, when the marker genotype can be determined from the banding pattern (usually by RFLPs or isozymes, which are codominant). For example, with RFLP analysis, each probe-enzyme combination may be considered to be a RFLP locus, and each unique banding pattern a RFLP variant (Melchinger et al., 1991; Bernardo, 1993). Provided the mode of inheritance is known and each banding pattern can be assigned a marker genotype, allele frequencies for marker loci can be calculated (see Appendix) to compute genetic distances between populations of individuals based on differences in allele frequencies (Melchinger, 1993). When the populations are homozygous and homogeneous as for inbred lines, the frequencies for a particular allele are either zero or unity.

3. Genetic distance measures

Based on either banding data or allelic data, distances and similarity measures can be computed. Such measures may be viewed as convenient means of data reduction, and they need not involve any genetic concept (Weir, 1996: 190). Some of the measures for allelic data were designed based on genetic models specifying the processes underlying the divergence of populations. It should be checked, whether or not these assumptions are met in practice.

3.1 Binary banding data

Sneath and Sokal (1973) list four classes of similarity measures: (i) distance coefficients, (ii) association coefficients, (iii) correlation coefficients, and (iv) probabilistic similarity coefficients. Most measures relevant for the analysis of binary banding data fall within the class of association measures, which are based on qualitative data (multistate or two-state). Occasionally association measures turn out to be special cases of distance coefficients or correlation coefficients. For a comprehensive overview see Sneath and Sokal (1973), Clifford and Stephenson (1975), and Gower (1985). In the following, we will give a few measures which we found to be frequently used in genetical studies. The data of two genotypes can be arranged as 2 x 2 frequency table

		x	
		0	1
y	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

From this basic table the following frequencies can be computed (Armstrong et al., undated)

n_{00} = number of band positions scored 0 for x and 0 for y
n_{10} = number of band positions scored 1 for x and 0 for y
n_{01} = number of band positions scored 0 for x and 1 for y
$n_{11} = n_{xy}$ = number of band positions scored 1 for x and 1 for y
$n_x = n_{01} + n_{11}$ = number of bands present in x
$n_y = n_{10} + n_{11}$ = number of bands present in y
$m_{xy} = n_{00} + n_{11}$ = number of matches
$n = n_{00} + n_{10} + n_{01} + n_{11}$ = number of band positions

The most important distinction is between measures that ignore negative matches (0, 0 comparisons) and measures that do not. It is debatable whether or not exclusion of negative matches is useful in the context of DNA profiles. Take the following simple example with three genotypes x, y and z:

Table 3: Example of scores for banding data of three genotypes *x*, *y* and *z* (4 band positions)

Band position		*a*	*b*	*c*	*d*
Genotype	*x*	0	0	1	1
	y	1	1	1	1
	z	0	1	1	1

In a way, *z* is as similar to *x* as it is to *y*, because in both comparisons, three of four comparisons are concordant. The only difference is that in the *z*-*x* comparison, only two of the three concordant observations are positive matches, while in the *z*-*y* comparisons all are positive matches. On the other hand, for a negative match to be observable, the corresponding band must be observed for at least one of the other genotypes. Thus, any similarity measure which takes into account negative matches, will depend on the particular set of genotypes included in the study, which is a point in favor of measures ignoring negative matches. Moreover, there are several ways in which a genotype may loose a band/DNA fragment, so it may be argued that basing similarity on the mutual absence of a character is improper (Vierling and Nguyen, 1992).

The similarity measures (*s*) given here take values in the range from zero to unity. For identical genotypes $s = 1$, while for completely distinct measures $s = 0$. The distance measure corresponding to these similarity measures may be computed as $d = 1 - s$.

3.1.1 Measures that ignore negative matches

(1) Nei and Li (1979):

$$NL_{xy} = 2n_{xy}/(n_x + n_y) = 2n_{11}/(2n_{11} + n_{01} + n_{10})$$

This is probably the most popular similarity measure in genetic analyses. It is equivalent to the Dice coefficient (Sneath and Sokal, 1973: 131) and assesses the proportion of bands shared by two genotypes *x* and *y*. Under certain statistical assumptions, NL_{xy} may be employed to derive an estimate of the mean number of nucleotide substitutions per nucleotide site (Nei and Li, 1979), which is a useful parameter in evolutionary studies. The underlying assumptions may be realistic in natural populations, but they are probably doubtful in plant breeding. If the assumptions are violated, there is no longer a direct biological interpretation (Swofford and Olsen, 1990: 435). If the computations are exclusively based on single-banded RFLP patterns, then NL_{xy} is equal to Rogers distance (see below) (Melchinger, 1993).

(2) Jaccard (Sneath and Sokal, 1973: 131):

$$J_{xy} = n_{11}/(n - n_{00}) = n_{11}/(n_{11} + n_{01} + n_{10})$$

NL_{xy} is the same as J_{xy}, except that positive matches carry double weight. It has been suggested (Link et al., 1995) that NL_{xy} is more appropriate for RFLP data, while J_{xy} should

be used with RAPD data. The reasoning is as follows: RAPD markers either produce a band in a certain position or the band is absent. Thus, one band position usually corresponds to one marker locus. By contrast, RFLPs produce fragments of varying lengths for different alleles. For two cultivars differing at a marker locus, fragments are produced for both alleles, but they differ in their position on the gel. Hence, a locus is represented by two band positions. When the cultivars are identical, however, the locus is manifest in only one band position for the pairwise comparison. Thus, matches should receive double weight compared to mismatches, as in NL_{xy}. This reasoning implies that RAPDs show no length polymorphisms and that each RFLP allele produces only one band on the gel. Both of these assumptions are idealizations, but they may be reasonable approximations in practice.

It should be remarked that there is a monotonic relationship between the coefficients by Nei and Li and by Jaccard:

$$NL_{xy} = 2J_{xy}/(1 + J_{xy})$$

Both measures will give identical rank orders of similarities. With some clustering methods (single linkage, complete linkage, see below) the dendrograms will be essentially the same with both measures (Digby and Kempton, 1987: 129). NL_{xy} is non-metric [ie. they do not satisfy the triangular inequality $d_{xy} \leq d_{xz} + d_{yz}$, where d_{xy} is the distance between x and y], while J_{xy} is metric, which is a point in favor of the latter (Sneath and Sokal (1973:131)).

3.1.2 Measures, which treat positive and negative matches alike

These measures are symmetric in n_{00} and n_{11}, i.e. the formula stays the same when n_{00} and n_{11} are exchanged. Only the most popular measure is given here. For other measures see Sneath and Sokal (1973) and Clifford and Stephenson (1975).

(3) Simple Matching (Sneath and Sokal, 1973: 132)

$$SM_{xy} = m_{xy}/n = (n_{11} + n_{00})/n$$

The simple matching coefficient measures the proportion of positive and negative matches. In order to compare SM_{xy} with measures that ignore negative matches, we computed some similarities for the example in Table 3. SM_{xy} yields the same similarity/distance for the pairs x-z and y-z, while measures ignoring negative matches such as J_{xy} and NL_{xy} indicate a larger similarity between y and z.

$J_{xz} = 0.67$	$NL_{xz} = 0.80$	$SM_{xz} = 0.75$
$J_{yz} = 0.75$	$NL_{yz} = 0.86$	$SM_{yz} = 0.75$

3.2 Allelic frequency data and band frequency data

In the following x_i and y_i will denote the frequencies of allele i at a given locus for populations x and y, respectively. Alternatively, x_i and y_i may denote the band frequency at band position i, when banding data are used.

(1) Euclidean distance

The frequencies x_i and y_i can be viewed as coordinates of points in a multidimensional space. The geometric distance may be interpreted as distance between populations x and y.

$$E_{xy} = [\Sigma_i(x_i - y_i)^2]^{0.5}$$

When allelic data from several loci are available, the distances for individual loci may be averaged. E_{xy} takes a value between 0 and $\sqrt{2}$. A standardization to values between 0 and 1 leads to Rogers' distance (Nei, 1987: 211)

$$RD_{xy} = [0.5\ \Sigma_i(x_i - y_i)^2]^{0.5}$$

In the important case that x and y are inbred lines and allelic data are used, Rogers' distance (RD_{xy}) equals the percentage of loci which differ between lines x and y. It's expectation is related to the coefficient of coancestry (Melchinger et al., 1991). The Rogers distance has the following deficiency: When the two populations are both polymorphic but share no common alleles, this measure can become much smaller than unity even if the populations have entirely different sets of alleles (Nei, 1987: 209).

(2) Nei's standard genetic distance

Nei's measure is intended for allelic data. When no allelic information is available, it may be computed from band frequency data. If this is done, however, the measure does not have the genetic interpretation as if computed from allelic data. The *normalized identity of genes* or simply *genetic identity* is given by

$$I_{xy} = J_{xy}/(J_x J_y)^{1/2}$$

where $j_{xy} = \Sigma x_i y_i$, $j_x = \Sigma x_i^2$, $j_y = \Sigma y_i^2$ and J_x, J_y and J_{xy} are the averages of j_x, j_y and j_{xy} over all scored loci. I_{xy} is 1 when the two populations have identical gene frequencies over all loci and is 0 when they share no alleles. Because of this property, I_{xy} has been used for measuring the extent of genetic similarity between populations. The quantity $D_{xy} = -\ln(I_{xy})$ is the *standard genetic distance*. Under the assumption that the rate of gene substitution per locus is uniform across both loci and lineages and some other assumptions, it is an estimator for the number of codon differences per locus between two populations x and y (Nei, 1987: 219; Nei, 1972). While I_{xy} ranges from zero to unity, D_{xy} varies between zero and infinity.

Loarce et al. (1996) computed the genetic identity based on band frequencies of RAPD fragments from bulked DNA samples of two rye cultivars. O'Donoughue et al. (1994) computed D_{xy} for band frequencies from RFLPs in oats. When computed from band frequency data, D_{xy} probably does not generally allow the interpretation as a measure of the number of codon differences, though it is a valid descriptive distance measure.

Nei's distance has been critizised because it violates the triangular inequality. Both the distances by Nei and by Rogers (as well as the Euclidian distance) are heavily influenced by within-population heterozygosity. A measure, which does not show this undesirable property, is the arc distance of Cavalli-Sforza (Swofford and Olsen, 1990:434).

(3) Cavalli-Sforza

$$CS_{xy} = \sqrt{[(1/L)\ \Sigma_L(2\theta/\pi)^2]}$$

where L is the number of loci and $\theta = \cos^{-1}\Sigma_i\sqrt{(x_i y_i)}$

4. Clustering methods

Based on a matrix of pairwise distances, a cluster analysis may be performed, usually producing a tree or dendrogram, which represents the relationship among cultivars. For inbred or clonal varieties that exhibit little genetic variation within the taxon, production of a dendrogram often completes the analysis. For outcrossing varieties further analysis is appropriate and may be essential to identify varieties (Morell et al., 1995). There is a host of clustering techniques (not all require a distance matrix), and, unfortunately, different methods may yield different groupings (Blackith and Reyment, 1971: 278). Clustering methods, therefore, do not lead to purely objective and stable classifications. Rather they should be seen as tools for data exploration (Dunn and Everitt, 1982: 104). To begin with, we give the following citation, which highlightens the problem of choosing the 'right' clustering method: "... *the subjectiveness of the choice of clustering procedures stems essentially from the impossibility of defining a cluster on other than arbitrary terms* (C.R. Rao, 1952; cited in Blackith and Reyment, 1971: 284)".

Many clustering procedures were developed by and for taxonomists and evolutionists, who are often interested in recovering the true phylogenetic relationship among organisms. There are a number of methods, by which a phylogenetic tree is constructed by considering various possible ways of evolution, following certain rules and choosing the best possible tree (parsimony methods, maximum likelihood methods). An underlying assumption is that present-day diversity of organisms is the result of a branching evolutionary tree. A phylogenetic tree constructed by these methods is often called a *cladogram*. Two popular software packages implementing various phylogenetic methods are PAUP (by David Swofford, National Museum of National History, Washington D.C.) and PHYLIP (by Joe Felsenstein, University of Washington). For an introduction to phylogenetic methods see Felsenstein (1982), Swofford and Olsen (1990), Pankhurst (1991: 68), and Weir (1996). Often, however, classifications are intended mainly to describe the (genetic) resemblence of

present-day organisms. Such classifications are called *phenograms*. Nei (1987: 291-292) argues that the assumptions required for cladistic methods are not always satisfied with molecular data. Backeljau et al. (1995) discuss conceptual problems limiting the reliability of RADP data in parsimony analyses. Swofford and Olsen (1990) discourage the use restriction-fragment data for input to phylogenetic analysis, primarily because these data violate the crucial assumption of independence. Many of the cladistic methods require data on nucleic acid sequences or protein sequences. The phylogenetic approach is useful in the study of natural populations and of interspecific relationships (Song et al., 1990; van de Ven et al., 1993; Monte et al., 1993), where forces of evolution are a dominating factor, whereas intraspecific relationships among crop cultivars and lines seem better suited for phenetic methods. Relationship among crop cultivars are often quite complicated due to artificial crosses among all kinds of parents, so the 'phylogeny' in this context is difficult to visualize. It is probably not similar to a nice hierarchical tree diagram. All we can hope for is that a dendrogram gives a reasonable grouping based on today's genetic similarities. The temptation to read all dendrograms as a phylogeny is one of the potential abuses. Here, we will consider the phenetic approach only.

4.1 The UPGMA method

Various algorithms are available for phenetic classification but the UPGMA (Unweighted Pair-Group Method using Arithmetic averages) method is the most commonly used. This method is described first and then contrasted with two other common methods: single linkage and complete linkage. UPGMA is based on the matrix of pairwise genetic distances. Suppose there are four cultivars. Let the matrix of distances be given by

		cultivar			
		1	2	3	4
	1	0			
	2	d_{12}	0		
cultivar	3	d_{13}	d_{23}	0	
	4	d_{14}	d_{24}	d_{34}	0

where d_{xy} is the distance between x and y. Clustering starts from the cultivars with the smallest distance. More distant cultivars are then gradually added to the cluster. If d_{34} is the shortest distance, cultivars 3 and 4 are clustered with a branching point located at distance d_{34}. Cultivars 3 and 4 are then combined into a cluster (34). New distances between this cluster and the other clusters (single cultivars up to this point) are calculated:

		cluster		
		1	2	(34)
	1	0		
cluster	2	d_{12}	0	
	(34)	$d_{1(34)}$	$d_{2(34)}$	0

where $d_{1(34)} = (d_{13} + d_{14})/2$ and $d_{2(34)} = (d_{23} + d_{24})/2$. Next, the objects with the smallest distance are clustered. If, e.g., this is $d_{2(34)}$, we join 2 and (34) with a branching point at

distance $d_{2(34)}$. At the last step, cultivar 1 is joined with (234) at a branching point $d_{1(234)} = (d_{12} + d_{13} + d_{14})/3$. The following hypothetical data set is employed to demonstrate the method.

Table 4: Hypothetical banding data for genotypes *A* to *D* (10 band positions)

	Band position									
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*
Genotype										
A	1	1	1	0	1	1	1	1	0	1
B	0	1	0	1	1	1	1	1	0	0
C	0	1	1	1	1	0	1	0	1	0
D	1	1	1	0	1	1	0	0	1	1

From these data, matrices of Dice similarities (NL_{xy}) and distances (1 - NL_{xy}) are computed.

Dice similarity NL_{xy}:

	A	*B*	*C*	*D*
A	1.00			
B	0.71	1.00		
C	0.57	0.67	1.00	
D	0.80	0.46	0.62	1.00

Dice distance 1 - NL_{xy}:

	A	*B*	*C*	*D*
A	0			
B	0.29	0		
C	0.43	0.33	0	
D	0.20	0.54	0.38	0

The shortest distance is d_{AD}, so *A* and *D* are fused into one cluster. After the first clustering cycle there are three clusters: (*AD*), *B* and *C*. For these, a new distance matrix is computed. The distance d_{BC} (= 0.33) does not involve the new cluster (*AD*) and need not be re-computed. The new distance $d_{B(AD)}$ is computed as the average of the distances between *B* and members of the cluster (*AD*):

$d_{B(AD)} = (d_{AB} + d_{BD})/2 = (0.29 + 0.54)/2 \sim 0.42$ and likewise
$d_{C(AD)} = (d_{AC} + d_{CD})/2 = (0.43 + 0.38)/2 \sim 0.41$

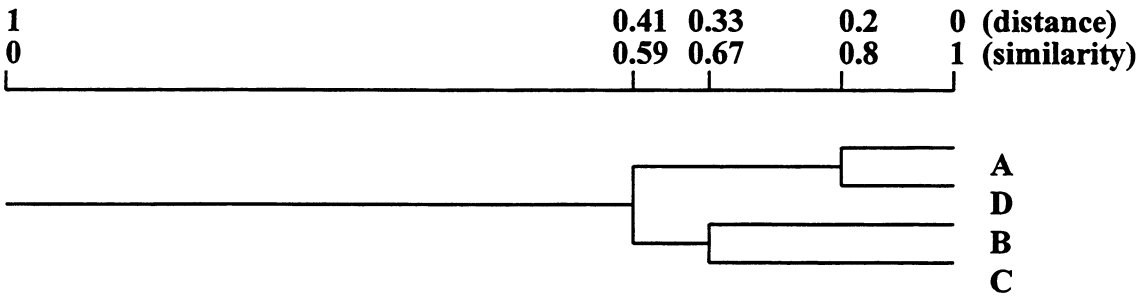
Dice distance 1 - NL_{xy} after 1st cycle:

	B	*C*	(*AD*)
B	0		
C	0.33	0	
(*AD*)	0.42	0.41	0

Next, *B* and *C* are grouped into one cluster, since $d_{BC} = 0.33$ is the shortest distance. The last step is to fuse clusters (*AD*) and (*BC*). The average distance between these two clusters is the average of all pairwise distances, where one individual is from (*AD*) and the other from (*BC*):

$$d_{(AD)(BC)} = (d_{AB} + d_{AC} + d_{BD} + d_{CD})/4 = (0.29 + 0.43 + 0.54 + 0.38)/4 = 0.41$$

The outcome of this clustering process is graphically displayed in a dendrogram. The scale above the dendrogram indicates at which distance (similarity) clusters were fused.



A few remarks clarifying the acronym UPGMA are in order. The method is an average linkage method, i.e. an average similarity or dissimilarity between candidate cultivars/clusters is computed at each cycle. UPGMA employs the arithmetic average. Alternatively, the socalled centroid can be computed (UPGMC). Average linkage may be contrasted to single linkage/nearest neighbor methods (distance computed for the two closest members of candidate clusters) and complete linkage/furthest neighbor methods (distance computed for the two furthest members of candidate clusters).

Another aspect is that of weighted vs. unweighted clustering. In UPGMA, being an unweighted method, all cultivars in a cluster are weighted equally when computing a distance to a new candidate. By contrast, the Weighted Pair-Group Method using Arithmetic averages (WPGMA) weights distances in favor of the most recent arrival within a cluster. Finally, UPGMA is a pair-group method, i.e. the dendrogram is restricted to bifurcations, while variable-group methods allow multiple furcations. The former are easier to program, while the latter take into account that differences among potential clustering candidates may be too small as to warrant separate furcations.

4.2 Single linkage (nearest neighbor)

As for average linkage, *A* and *D* are fused first, but the new distance matrix is computed differently. The single linkage distance $d_{B(AD)}$ is the shortest distance between *B* and members of (*AD*), i.e.

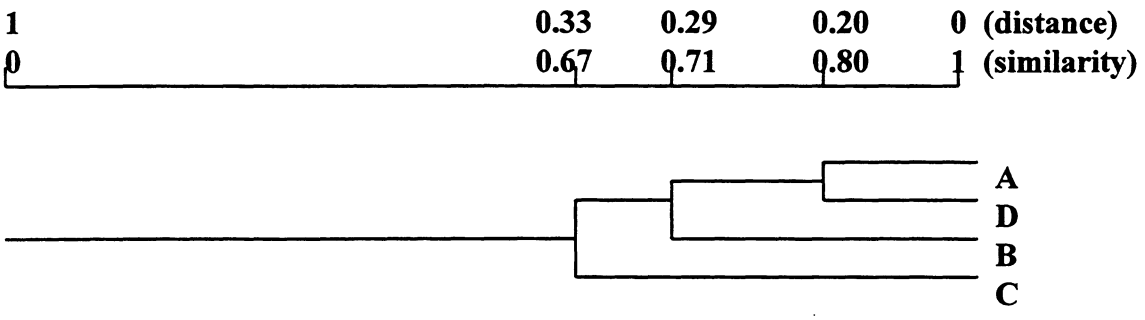
$d_{B(AD)} = \min(d_{AB}, d_{BD}) = \min(0.29, 0.54) = 0.29$ and similarly
$d_{C(AD)} = \min(d_{AC}, d_{CD}) = \min(0.43, 0.38) = 0.38$

Dice distance 1 - NL_{xy} after 1st cycle:

	B	*C*	(*AD*)
B	0		
C	0.33	0	
(*AD*)	0.29	0.38	0

In the next cycle, (*AD*) and *B* are grouped into one cluster, since $d_{B(AD)} = 0.29$ is the shortest distance. We see here an effect known as "chaining", i.e. clusters formed at one step are likely to be involved in the next clustering step. Finally, we compute

$d_{C(ABD)} = \min(d_{AC}, d_{BC}, d_{CD}) = \min(0.43, 0.33, 0.38) = 0.33$



4.3 Complete linkage (furthest neighbor)

As with average linkage and single linkage, *A* and *D* are fused in the first step. The complete linkage distance $d_{B(AD)}$ is the largest distance between *B* and members of (*AD*), i.e.

$d_{B(AD)} = \max(d_{AB}, d_{BD}) = \max(0.29, 0.54) = 0.54$

and similarly

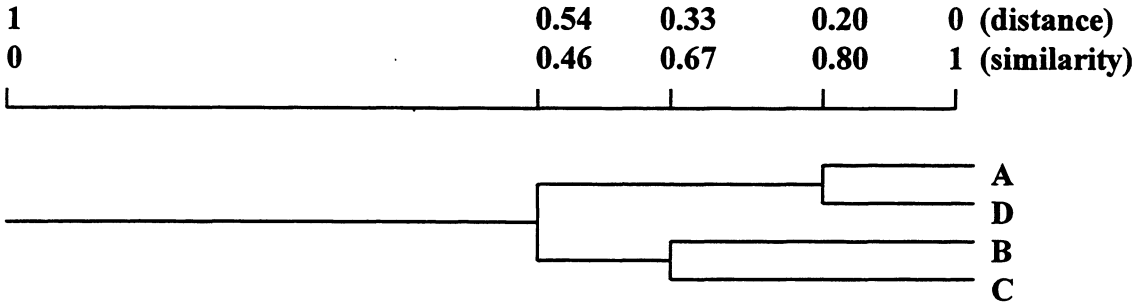
$d_{C(AD)} = \max(d_{AC}, d_{CD}) = \max(0.43, 0.38) = 0.43$

Dice distance 1 - NL_{xy} after 1st cycle:

	B	*C*	(*AD*)
B	0		
C	0.33	0	
(*AD*)	0.54	0.43	0

Next, *B* and *C* are clustered, because d_{BC} is the shortest distance. The complete linkage distance between (*AD*) and (*BC*) is the maximum of all pairwise distances, where one individual is from (*AD*) and the other from (*BC*):

$d_{(AD)(BC)} = \max(d_{AB}, d_{AC}, d_{BD}, d_{CD}) = \max(0.29, 0.43, 0.54, 0.38) = 0.54$



4.4 Other methods

UPGMA, single linkage and complete linkage have four important properties, abbreviated under the acronym SAHN (Sneath and Sokal, 1973: 214; Rohlf, 1994: 8-21), which they share with a large number of other methods:

(1) Sequential (S): A recursive sequence of operations is applied to arrive at the final partition. By contrast, simultaneous methods, such as ordination techniques (e.g. principal component and principal coordinate analysis, multidimensional scaling), involve one single nonrecursive step yielding the final arrangement of cultivars.

(2) Agglomerative (A): Starting with each cultivar as a separate set, cultivars are sucessively grouped, arriving eventually at a single set containing all cultivars. By contrast, divisive methods start out with all cultivars in on single set, subdividing this into subsets.

(3) Hierarchic (H): The number of subsets is reduced at every step, thus creating a taxonomic hierarchy. Any hierarchy can be drawn as a rooted tree or dendrogram. Nonhierarchical techniques include various ordination methods, in which cultivars are projected into two- or three-dimensional space. Nonhierarchical graphs among cultivars are not rooted, i.e. they do not have a beginning from which branches diverge (e.g. minimum spanning trees).

(4) Nonoverlapping (N): A cultivar belonging to one clustering set may not belong to any other set. If the clustering method is hierarchical this means that the classification must be

nested. In overlapping methods, cultivars are allowed to be assigned to more than two clusters. If, for example, the parents of a hybrid are in two different clusters, then the clusters might be considered to overlap, with the hybrid as an object which belongs to both clusters (Pankhurst, 1991: 67).

A very popular SAHN technique not mentioned so far is the Ward method. It fuses subsequent entities such that the sum of squared Euclidian distances within a cluster increases by the smallest amount (Clifford and Stephenson, 1975: 113).

4.5 Choice of clustering method

When faced with a set of data the question arises as to the choice of clustering method. One approach is to try different methods and see to what extend they agree. Broad agreement among different classifications of the same objects is expected if a natural grouping really exists and if the classifications reflect the true associations (Digby and Kempton, 1987).

Summarizing empirical results on hierarchical methods, Everitt and Dunn (1982: 87) state that
(a) no single method is best in every situation
(b) the mathematically respectable single linkage is, in most cases, the least successful for the data used (mainly due to the "chaining" effect), and
(c) group average clustering and the Ward method do fairly well, overall.

In a study with maize inbreds Mumm and Dudley (1994) compared (i) single linkage (ii) UPGMA (iii) Unweighted Pair Group Method Using Centroids (UPGMC) (iv) complete linkage (v) Ward. UPGMA (based on Jaccard coefficient) showed grouping most consistent with pedigree data. Wilkie et al. (1993) employed (i) single linkage, (ii) complete linkage and (iii) UPGMA based on Rogers' distance to study taxonomic relationships within the genus *Allium*. Similar results were obtained with all methods.

Jain and Dubes, (1988; cited in Mumm et al. 1994) list three types of validation for cluster analyses, which are useful in practice:
- External (compare distance matrix to external information not used in clustering, e.g. pedigree relationships)
- Internal ("cophenic" correlation between original distance matrix and pairwise distances implied by the phenogram; assesses the degree to which original distances are preserved through the clustering process)
- Relative (compare agreement of different classifications).

References

Armstrong J, Gibbs A, Peakall R, Weiler G (undated) RAPDistance Programs; Version 1.03 for the analysis of patterns of RAPD fragments. Manual. Australian National University, Canberra

Backeljau T, Bruyn LD, De Wolf L, Jordaens K, Van Dongen S, Verhagen R, Winnepenninckx B 1995 Random amplified polymorphic DNA (RAPD) and parsimony methods. Cladistics 11: 119-130

Bernardo R 1993 Estimation of coefficient of coancestry using molecular markers in maize. Theoretical and Applied Genetics 85: 1055-1062

Blackith RE, Reyment RA 1971 Multivariate morphometrics. Academic Press, London
Clifford HT, Stephenson W 1975 An introduction to numerical classification. Academic Press, New York

Cooke RJ 1995 Varietal identification of crop plants. pp 33-65 *In* Skerritt JH, Appels R (eds) New diagnostics in crop sciences. Wallingford

Digby PGN, Kempton RA 1987 Multivariate analysis of ecological communities. Chapman and Hall, London

Dunn G, Everitt BS 1982 An introduction to mathematical taxonomy. Cambridge University Press, Cambridge

Felsenstein J 1982 Numerical methods for inferring evolutionary trees. Quarterly Reviews in Biology 57: 379-404

Gower JC 1985 Measures of similarity, dissimilarity, and distance. pp 397-405 *In* Kotz S, Johnson NL, Read CB (eds) Encyclopedia of Statistical Sciences, Vol 5

Jain AK, Dubes RC 1988 Algorithms for clustering data. Prentice-Hall, Englewood Cliffs, NJ

Link W, Dixkens C, Singh M, Schwall M, Melchinger AE 1995 Genetic diversity in European and Mediterranean faba bean germ plasm revealed by RAPD markers. Theoretical and Applied Genetics 90: 27-32

Loarce Y, Gallego R, Ferrer E 1996 A comparative analysis of the genetic relationships between rye cultivars using RFLP and RAPD markers. Euphytica 88: 107-115

Melchinger AE, Messmer MM, Lee M, Woodman WL, Lamkey KR 1991 Diversity and relationships among U.S. maize inbreds revealed by restriction fragment length polymorphisms. Crop Science 31: 669-678

Melchinger AE 1993 Use of RFLP markers for analysis of genetic relationships among breeding materials and prediction of hybrid performance. pp 621-628 *In* Buxton DR (ed) First International Crop Science Congress of the Crop Science Society of America, Madison

Monte JV, McIntyre CL, Gustafson JP 1993 Analysis of phylogenetic relationships in the *Triticeae* tribe using RFLPs. Theoretical and Applied Genetics 86: 649-655

Morell MK, Peakall R, Appels R, Preston LR, Lloyd HL 1995 DNA profiling techniques for plant variety identification. Australian Journal of Experimental Agriculture 35: 807-819
Mumm RH, Dudley JW 1994 A classification of 148 U.S. maize inbreds: I. Cluster analysis based on RFLPs. Crop Science 34: 842-851

Mumm RH, Hubert LJ, Dudley JW 1994 A classification of 148 U.S. maize inbreds: II. Validation of cluster analysis based on RFLPs. Crop Science 34: 852-865

Nei M 1972 Genetic distance between populations. American Naturalist 106: 283-292

Nei M 1987 Molecular evolutionary genetics. Columbia University Press, New York

Nei M, Li W-H 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences (USA) 76: 5269-5273

O'Donoughue LS, Souza E, Tanksley SD, Sorells ME 1994 Relationships among North American oat cultivars based on restriction fragment length polymorphisms. Crop Science 34: 1251-1258

Pankhurst RJ 1991 Practical taxonomic computing. Cambridge University Press, Cambridge

Rohlf FJ 1994 NTSYS-pc. Numerical taxonomy and multivariate analysis system. Manual. Applied Biostatistics Inc., Setauket

Song K, Osborn TC, Williams PH 1990 Brassica taxonomy based on nuclear restriction fragment length polymorphisms (RFLPs). Theoretical and Applied Genetics 79: 497-506

Sneath PHA, Sokal RR 1973 Numerical taxonomy. WH Freeman and Company, San Francisco

Swofford DL, Olsen GJ 1990 Phylogeny reconstruction. pp 411-501 *In* Hillis DM, Moritz C (eds) Molecular systematics. Sinauer, Sunderland.(2nd edition 1996)

van de Ven WTG, Duncan N, Ramsey G, Phillips M, Powell W, Waugh R 1993 Taxonomic relationships between *V. faba* and its relatives based on nuclear and mitochondrial RFLPs and PCR analysis. Theoretical and Applied Genetics 86: 71-80

Vierling RA, Nguyen HT 1992 Use of RAPD markers to determine the genetic diversity of diploid wheat genotypes. Theoretical and Applied Genetics 84: 835-838

Weir BS 1996 Genetic data analysis II. Sinauer, Sunderland

Wilkie SE, Isaac PG, Slater RJ 1993 Random amplified polymorphic DNA (RAPD) markers for genetic analysis in Allium. Theoretical and Applied Genetics 86: 497-504

Appendix A

To give an example for deriving allele frequencies and band frequencies from banding patterns, consider a monomeric single-locus enzyme showing triallelic variation in a crosspollinating population of a diploid species. Each allele produces a polypeptide chain, leading to an enzyme. The enzyme from all three alleles are functionally equivalent (isozymes), they only differ in the polypeptide chain size, shape, and charge, and hence in their movement on an electrophoretic gel. Thus, for each allele there is a band position on the gel. The banding pattern of a sample of 10 plants from a given population might look as follows:

Table A: Banding patterns from a hypothetical isozyme system

	Individuals										
band position	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	band frequency
1			---	---		---		---		---	0.5
2	---			---		---					0.3
3	---	---	---		---		---		---	---	0.7
Genotype	23	33	13	12	33	12	33	11	33	13	

Individual I_1 is heterozygous, showing bands in positions 2 and 3. It may be scored "23". By contrast, I_7 is homozygous with only one band in position 3. Thus, it is scored "33". The other individuals are scored in a similar fashion. To obtain the frequency of allele 1 in the sample, simply count the number of 1's on the row of scores and divide this count by twice the number of sampled individuals. In the example, the frequency is 6/20 = 0.30.

Allele	allele frequency
1	$x_1 = 0.30$
2	$x_2 = 0.15$
3	$x_3 = 0.55$

A simpler, but less informative, analysis of the banding pattern is obtained by computing band frequencies (see Table A). Note that in this example the rank order of band frequencies is the same as that for allele frequencies.

When the population is homogeneous and homozygous (e.g. inbred lines), the band and allele frequencies at a given band position will be either 1 or 0. Finally, it is noted that allelic analysis becomes more complicated with polyploids. With enzymes, further difficulties arise, when they are polymeric and/or goverened by multiple loci.

Appendix B

A problem with the interpretation of banding patterns is the treatment of monomorphic bands, i.e. of bands, which are common to all genotypes. The following example of five genotypes with four monomorphic band positions and a total of eight band positions (Table B) is due to Dr. W. Link (University of Göttingen, Germany; also see Link et al., 1995).

Table B: Hypothetical banding pattern of five genotypes (*A, B, C, D, E*) with monomorphic bands

Genotype					
A	*B*	*C*	*D*	*E*	band position
----	----	----	----		*a*
----	----	----			*b*
----					*c*
----					*d*
----	----	----	----	----	*e*
----	----	----	----	----	*f*
----	----	----	----	----	*g*
----	----	----	----	----	*h*

From this banding pattern the Jaccard coefficient may be computed with monomorphic bands included. Consider similarities between *A* and *B* and between *C* and *D*:

$J_{AB} = 6/8 = 0.75$ $J_{CD} = 5/6 = 0.83$

If, by contrast, we ignore monomorphic bands, the similarities are

$J_{AB} = 2/4 = 0.50$ $J_{CD} = 1/2 = 0.50$

Firstly, the similarity is numerically larger when monomorphic bands are included. Secondly, in this example similarities J_{AB} and J_{CD} are the same when monomorphic bands are excluded, whereas with monomorphic bands the similarity between genotypes *C* and *D* is larger than that between *A* and *B*. The example shows that it does make a difference whether or not monomorphic bands are included. It is not easy to decide which of the two strategies is preferable. Also, which bands are interpreted as monomorphic, depends on the genotypes included. Had we included a sixth genotype *F* with missing bands in positions *e* to *h*, all band positions would have been heteromorphic.

[End of document]