UPOV

BMT/2/5
ORIGINAL : English
DATE : March 8, 1994

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

GENEVA

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

Second Session
Versailles, France, March 21 to 23, 1994

BASIS AND USES OF DISTANCES FOR VARIETAL CHARACTERIZATION

Document prepared by experts from France

UPOV
Biochemical and Molecular Techniques Group
Versailles March 21 to 23, 1994

BASIS AND USES OF DISTANCES FOR VARIETAL CHARACTERISATION

Studies for UPOV may involve collecting either or both qualitative or quantitative collection of data. Such information can be grouped, or used to study objects by the calculation of distances, that is, dissimilarity between the objects described. This is the case for analyses of distinctiveness, homogeneity and stability. The notion of distance is also valuable for the study of concepts such as essential derivation of minimal distances.

The *distance* between two individuals, A and B, is any function $\mathrm{d}(A,B)$ of their measure which possesses the following properties:

1. $\mathrm{d}(A,B) \geq 0$, such that a distance is always positive (or zero).
2. $\mathrm{d}(A,B) = 0$, if and only if $A = B$, such that the distance is zero if the measure of the individuals are identical.
3. $\mathrm{d}(A,B) = \mathrm{d}(B,A)$, such that the distance from A to B is equal to the distance from B to A.

Formally, these properties define similarity. A mathematical distance also has the property of triangular inequality (more commonly known in the form "the shortest route between two points is a straight line"). However, this rule has little meaning as concerns taxa.

The effects of the environment on morphophysiological characters, and the choice of locus on genetic characters cause experimental measures of these characters to vary. Thus, only an estimation of distances can be obtained, and it consequently important to determine the accuracy of this estimation. It is also important to be aware that there are always several ways to consider the distance between individuals : coefficient of relatedness and yield have very different meanings, for example. It is thus essential to know what it is one is trying to measure. A measure of distance between two individuals has little meaning unless the characters used in the calculation of this dissimilarity, the method used to calculate the difference and the contribution of error or variability are known.

The relevance of the characters chosen to be measured is the main limiting factor for this approach. It is valuable nevertheless to define what can be expected from the various calculations of distance. This article therefore presents the most widely used distances for continuous characters (particularly phenotypic characters) and then those for discrete characters (particularly genotypic characters). The problems associated with the combination of these two types of distance are addressed and finally some of the classic uses of calculations of distance are presented.

Distance as a tool for continuous measures

Euclidean distance

Data is most often representative of morphological characters. For each individual, we have p observations, each scored $A_1, \ldots, A_p$ for individual A and $B_1, \ldots, B_p$ for individual B. The most simple distance between A and B is defined as:

$$d_1^2(A,B) = (A_1 - B_1)^2 + \cdots + (A_p - B_p)^2 = (A_1 - B_1, \ldots, A_p - B_p) \begin{pmatrix} A_1 - B_1 \\ \cdot \\ \cdot \\ \cdot \\ A_p - B_p \end{pmatrix} \tag{1}$$

This corresponds to the sum of the differences (squared such that all values of distance are positive).

This method is straightforward, but has certain disadvantages. The choice of units indirectly affects the results. It is thus advisable to address this problem by standardising[1] all the raw data, such that the distance is independent of the measurement scale. The resulting distance, calculated from the this standardised data is generally called the *standardised Euclidean distance.*

Mahalanobis Distance

The second problem with Euclidean distances (even standardised) is that variables are not necessarily independent, and thus some of the information is redundant. Furthermore, not all the variables are measured with the same accuracy. It is therefore desirable to accord more importance to the most accurately measured variables and to eliminate redundant information. The Mahalanobis distance does this. It is defined by the equation:

$$d_2^2 = (A_1 - B_1, \ldots, A_p - B_p) W^{-1} \begin{pmatrix} A_1 - B_1 \\ \cdot \\ \cdot \\ \cdot \\ A_p - B_p \end{pmatrix} \tag{2}$$

Where W is the intra-population variance-covariance matrix. The corre-

[1] that is, to divide each data point by the standard deviation of the values of the variable considered for all the individuals

lations and variations of the measured variables are presumed to be identical for every taxon. The difference between this taxa model and the Euclidean model of distance is apparent from the case where one of the variables measured corresponds to the sum of two other variables included in the calculation of distance. The Mahalanobis distance eliminated this redundancy whereas the Euclidean distance does not. There has been much statistical analysis of this model such that, for example, it can be used to construct confidence intervals and tests.

In summary, the Euclidean distance corresponds to a "character by character" approach to distance, and standardisation can be used to give all measures equal weight. In contrast, the Mahalanobis distance involves a multi-character approach to distance and avoids the redundancy of information associated with the Euclidean distance.

Distance as a tool for discrete measures

Certain notations correspond to describing a continuous character in discrete steps. In such cases, the character is ordered and the simplest method is to consider the character as continuous. Nevertheless, it should be noted that such 'discretisation' of a continuous character represents a major loss of information. This paragraph only addresses non-ordered discrete variables, and in particular gentotypic data.

The first step is to calculate the frequency of each of the discrete variables for each object. The Mahalanobis distance cannot be used because the sum of the frequencies must equal one, and this means that the inverse of the intra-population, variance-covariance matrix will not be unique. A variety of alternatives to this matrix have therefore been proposed. For example, Goodman's distance is based on the use of a variance-covariance matrix for a reference population, F_2, and Hanson and Casas' distance uses a matrix based on specific aptitude. All these approaches to calculating distance suffer from the disadvantage that they require extensive experimentation to obtain estimates of the value used to substitute W in the Mahalanobis distance (diallele plan, F_2, populations etc.). There follows a description of types of distance that can be easily used for genotypic data, and do not involve genetic hypotheses.

Rogers' Distance

By analogy with the Euclidean distance, it is intuitive to add together the allelic distances. To avoid problems of sign, the squares of distances are used. This approach is the basis of the Rogers' distance, which is defined by:

$$d_3^2(A,B) = \frac{1}{2}\sum_i (A_i - B_i)^2 \quad i = 1, \ldots, n \; n : \text{nombre d'allèles} \qquad (3)$$

For a given locus, it appears logical increase the weight of differences between allelic frequencies $(A_i - B_i)$ as the rarity of the allele in the two populations increases. Numerous weighting terms have been proposed in the literature, and correspondingly, there are a multitude of distances. In cases of several loci, the best approach is to use the mean of the distances calculated for each locus. It is possible to show that the expectation of all these distances are correlated with the Malécot coancestry coefficient, using hypotheses of panmixia and in the absence of mutations. Theses hypotheses are however restrictive and the resulting distances should be considered as descriptive rather than genetic parameters.

For homozygous lines, most of these distances are equivalent, and the use of weighting terms is not necessary. This is why the Rogers' distance is so widely used for homozygotes.

Nei's distance

Another approach to the calculation of distance is to count the number of mutations that have occurred since the divergence of two taxa. This is the idea behind Nei's distance. It should be noted that this approach is completely different from that of relatedness coefficients. To see this clearly, consider the populations F_2, derived from two unrelated parents. The coefficient of relatedness for the individuals is constant whereas the number of mutations required to pass from one individual to another varies for pairs of individuals. If we define A_i (and B_i) the probability of two identical alleles in populations A (and B) and C_i the probability of identity of an allele drawn from A with one drawn from B, Nei's distance is defined by:

$$d_4 = -\log\left(\frac{C_i}{\sqrt{A_i B_i}}\right) \tag{4}$$

For data from RFLP analysis, or homozygous lines, this corresponds to counting the number of common bands and dividing by the total number of bands (and then calculating the logarithm of this value). Nei's distance is particularly valuable because of its clear evolutionary significance, and its mathematical derivation allows of this its properties to be studied.

It should be noted that this distance is calculated from a small proportion of the total loci. It is thus only an estimation of distance and it is important to assess its accuracy. Resampling methods can be used to calculate the variance of these distances and construct confidence intervals. However, these methods are in practice painstaking. If, for example, the distance between two lines, sharing 75% of their genome in common is calculated using 80 single-locus probes, there is a 95% chance of finding between 14 and 26 discriminative probes.

For homozygous lines, if the signals are single bands, Nei's distance also calculates the number of loci in common. If the probes are single-locus probes, the Rogers' distance also calculates the percentage of the probed genome in common. Recent tools have made possible simpler methods for resampling to calculate confidence intervals, and for testing distances, each with reference to the others. However, these distances is meaningful only if a large number of loci are used for their calculation (at least 50).

Simultaneous consideration of genotypic and phenotypic data

If genotypic data were independent of morphophysiological data, the corresponding distances could simply be added together. However, if the number of loci is large, there is likely to be imbalance of linkage between the loci included in the calculation of genotypic distance and the loci involved in the expression of morphophysiological characters used for the calculation of phenotypic distance. Recent work has show how to allow for these associations, but the main problem is the estimations of such imbalances which depend on the genetic group considered. The bias in the distance due to redundant information is not / constant. Inappropriate appreciation of the correlations will thus result in a poor estimation of distance and consequently erroneous conclusions. It is therefore preferable to calculate these two types of distance independently and use them separately for analysis.

' necessarily

Some common uses of distance

It is first important to consider the pertinence of the characters used to the issue being studied. The same is true for the variability of measures (for example as a result of environmental problems for morphophysiological characters, or due to sampling problems for genotypic characters). The calculated distance is only an estimation of the true distance between the individuals. It should be remembered that a distance can be calculated in a variety of ways (we have seen the difference between the relatedness coefficient and the number of mutations since divergence of two individuals for example). Distance is thus subjective and therefore what it is that is to be calculated must be clear when choosing the appropriate approach to use.

The current literature contains many descriptions of dendograms. The aim of/is to classify individuals. However, the conclusions can only be used for groups of individuals, and not for two by two comparisons. There are numerous methods for associating groups, and different aggregation criteria will give different final results. If we consider the distance in kilometres between two towns (which is in itself subjective, because we could use journey

/ dendograms

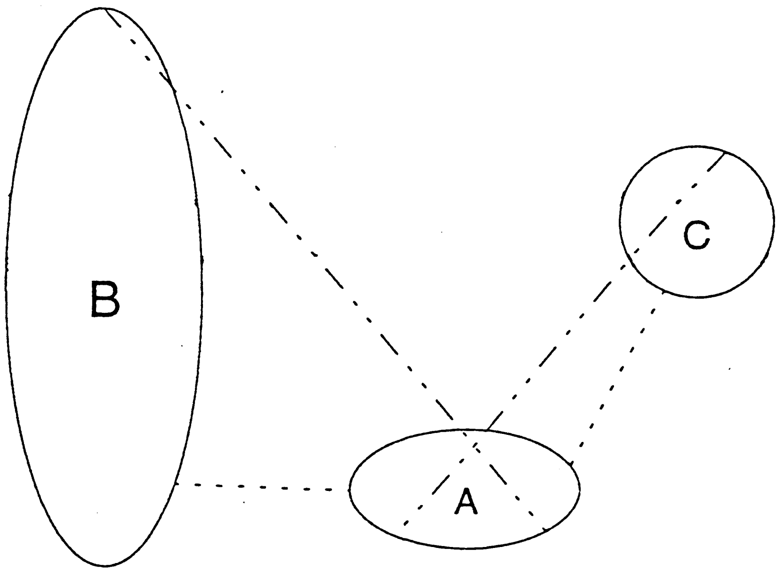
time) there is a difference between considering the closest points (which tends to group together sprawling towns), the points furthest apart (which tends to group small towns) and the centre (which tends to favour dense towns) (see figure 1). This example can easily be transposed to issues associated with varieties. It is thus important to decide what to favour, and what to disadvantage.

A classic method is to use several aggregation criteria. The grouping common to different analyses are termed stable groups. The individuals which behave differently in different analyses can then be considered in greater depth.

In conclusion, it is important to know exactly what is being investigated or tested, and the limits inherent in the method chosen. It is extremely hazardous to draw conclusions from a calculation of distance without quantifying the risk of error.

Avner Bar-Hen
March 1994

Figure 1: Illustration of the different classifications resulting from different aggregation criteria



Aggregation criteria

- - - The closest points : d (A, B) < d (A, C)

— - The furthest points : d (A, B) > d (A, C)

[End of document]