Methods of Data Analysis

Xavier Perrier, Albert Flori and François Bonnot

Analysis of data on diversity consists mostly of analysis of crossing tables of taxonomic units—species, populations, cultivars, etc.—and of variables that characterize the diversity. These variables are quantitative agronomic characters, morphological descriptors, mostly qualitative, biochemical markers (isozymes), or molecular markers (RFLP, RAPD), often coded as presence or absence. The taxonomic units are generally observed individually, but they can also correspond to populations; the observations are thus of means of characters, of allele frequencies, and so on. In any case, the analysis of these data aims to discover a possible structure of taxonomic units by analysis of resemblances or dissimilarities between these units.

This objective comes under classification, a vague term that has two meanings in current usage. The first involves the separation of objects into homogeneous groups. This ambition has been that of all the great taxonomists, from Aristotle to Linnaeus. The second meaning comes under the action of logical ordering. This is essentially to do with relations of order between taxonomic units, that is, with relations of filiation between units in the sense of evolution, which is phylogenetics.

Both these tasks may result in graphic representations in the form of a tree. The informative content of each is, however, fundamentally different. In the first case, the nodes of the tree correspond to concepts of grouping. In the second, the nodes are the parental units that are not observed but are supposed to exist, for example, ancestral forms that have disappeared. The links thus include the temporal dimension of evolution.

The problem of diversity treated here comes partly from these two processes, since we wish to define groups of comparable units, as well as to describe the relations of parentage between these groups.

Two approaches can be taken to treat evolutionary organization. The first, called phenetic, describes an organization from objective measures of dissimilarities between the units and refers the introduction of genetic hypothesis to the interpretation. These dissimilarities are estimated overall for all the characters observed and the number of characters gives relevance to the measures. It is often extremely difficult to detect the organization of units by direct examination of their dissimilarities. The object of the analysis is thus to find a close but easily readable representation of them: a factorial plane or a tree structure. The advantage of readability is gained at the cost of a certain loss of information, a loss that one of course seeks to minimize.

The second approach, called cladistic, directly looks at observed characters and attempts to distinguish identical characters inherited from a common ancestor, which are only informative, from phenomena of accidental convergence. It relies on a model of genetic transformation and was established initially on morphological characters that were known to be monogenic and for which the order of different modalities in terms of evolution was known. This approach is now most often used for genome sequencing data provided with an explicit genetic model, of the type developed by Jukes and Cantor (1969) or derived from that type.

The cladistic or related methods such as the methods of compatibility or likelihood are not described and are not within the scope of this work. The reader is referred, for example, to Darlu and Tassy (1993). The nature of markers used often allows only the use of phenetic approaches relevant to the analysis of dissimilarities.

The object of this chapter is to present relevant methodological aspects of these approaches. The base data will be a table of taxonomic units described by markers of diversity. There may also be a stacking up of tables when the same set of units is characterized by several series of markers of various kinds that cannot be grouped together. These tables can be considered separately at first, but researchers ultimately try to analyse them simultaneously.

This chapter is made up of three major parts.

The first involves the initial step of any analysis, the definition of a measure of the resemblance or dissimilarity between individuals. The choice of this measure is of primary importance and corresponds to a deliberate choice of perspective on the data. When the individuals are populations and the observations are of allele frequencies, various measures of dissimilarity, called genetic, can be used—e.g., Nei distances, Gregorius distances. Their properties have often been studied, for example in Lefort-Busson and de Vienne (1985). When the data are of individuals, described by a series of characters, the nature of variables—quantitative, qualitative, binary—determines the usable types of dissimilarity. It will be shown, on the example of molecular markers of RFLP or RAPD type, that knowledge of the biological characteristics of markers helps us choose an index of dissimilarity.

The second part describes factorial methods. Once a two-by-two dissimilarity matrix between individuals is established, it is analysed so that a simplified but faithful representation can be found. If the dissimilarity is Euclidean, factorial methods of analysis can be used. The bases of principal methods are summarized, particularly techniques of simultaneous analysis of several tables. If the dissimilarity is not Euclidean, methods of multidimensional scaling can be used to find a decomposition on a fixed number of axes. These iterative methods require considerable time for calculation when there are a large number of units to be treated. They have not been covered in this work, and the reader is referred, for example, to Escoufier (1975).

The third part concerns tree representations. An evolutionary process, by accumulation of transmissible mutations, will produce an organization that can be represented in the form of a tree. This representation, because of its biological foundation, is particularly relevant for the analysis of diversity, even if the modalities of evolution are generally more complex. Various methods for inferring trees will be presented, as well as techniques that allow the construction of synthetic trees from several types of markers.

Factorial methods and tree methods constitute two very different approaches to the representation of diversity. They must be considered complementary rather than concurrent. Factorial methods aim above all to make an overall representation of diversity that is as far as possible free of individual effects. On the other hand, the more commonly used tree methods tend to represent individual relations faithfully. They are thus two different ways of viewing the data.

The main statistical software proposes various methods of factorial analysis. The tools linked to analysis of dissimilarities and to their tree representation are less frequent. The Ntsys software is for instance a complete and easily accessible software (Rohlf, 1987). Various non-available functions or partly modified methods have been grouped in a specialized software available from the authors (Darwin, dissimilarity analysis and representation; Perrier, 1998).

CHOOSING A DISSIMILARITY INDEX

The aim of a dissimilarity index is to define a means of measuring the resemblance (similarity) between two individuals or, on the other hand, the difference (dissimilarity), since one can go from one to the other simply by a linear transformation.

There are, in fact, a large number of measures of dissimilarity and, depending on the data, various choices are available. Some of these dissimilarities have mathematical properties useful for the analysis or are stable in the face of data errors, which are relevant arguments for their use. The final choice depends on the real informative content of the markers used and to the point of view one wishes to take towards this information.

Finally, certain dissimilarities, such as ultrametric or additive distances, have special properties and can be represented in the form of a tree. There is, of course, no reason for the measure established on the data to have these properties. The object of tree construction methods is to transform the measure established on the data into a tree-representable dissimilarity, with the least possible deformation.

Definitions and Properties

The most general mathematical object representing the difference is a dissimilarity; this is a function d of the set of pairs (i,j) of individuals in the set of positive or null reals, symmetrical ($d(i,j) = d(j,i)$) and such that $d(i,i) = 0$ for any i. This definition is relatively simple and covers a large number of possible measures. On the other hand, it opens up only a few mathematical properties and it is necessary to add other constraints to acquire certain useful properties.

A distance, sometimes called a metric, is a particular dissimilarity obtained by adding the condition $d(i,j) = 0 \Rightarrow i = j$ and above all the triangular inequality between three individuals, $d(i, j) \leq d(i,k) + d(j,k)$. This natural condition translates simply into the possibility of representing any triplet of points by a two-dimensional triangle. This very useful property allows us especially to avoid the problem of negative edge lengths in the construction of a tree.

City block and Euclidean distance

An important group of distances is made up of Minkowski distances of order $p(p \geq 1)$. A distance belongs to this group if there exists an integer K and a series of K values x_{ik} applicable to each individual i, the distance thus being written as $d(i,j) = (\Sigma_k \mid x_{ik} - x_{jk} \mid^p)^{1/p}$.

The only cases used in practice correspond to the values 1 and 2 of p. When $p = 1$, the index is known as the city block, $d(i,j) = \Sigma_k \mid x_{ik} - x_{jk} \mid$. For $p = 2$, the usual Euclidean distance is found as follows: $d(i,j) = (\Sigma_k \mid x_{ik} - x_{jk} \mid^2)^{1/2}$. Of course, by definition, city block or Euclidean distances are obtained when we calculate a distance from a table of individuals × variables by summing, the absolute values or the squares of the deviations between i and j. However, certain indexes, for example those calculated on data of presence or absence, can be of either of these two types without that following evidently from the mode of construction.

The city block distance and Euclidean distance belong to the same group and are linked by certain relationships. It can be shown, for example, that any Euclidean distance is a city block distance. It has also been shown that the square root of a city block distance is a Euclidean distance.

Power transformation

Power transformations of a dissimilarity, for certain powers, give it properties of a distance or even of a Euclidean distance. First, we must recall that two dissimilarities d and d are equivalent in order if $d(i,j) \leq d(k,l) \Leftrightarrow d(i,j) \leq d(k,l)$. These two dissimilarities arrange the pairs of individuals in the same fashion and are thus of comparable interpretation. In particular, if d' can be written as an increasing monotone function of d, then d' and d are equivalent in order. This is the case of power functions $\alpha(\alpha \geq 0)$; d^{α} and d are equivalent in order.

It can be shown that, if d is a dissimilarity, one can always find a value α between 0 and 1 such that then for any λ $(0 \leq \lambda \leq \alpha)$, d^{λ} is a distance that, by the preceding property, is equivalent in order to d. Thus, if the informative content of markers results in the choice of a measurement that is only a dissimilarity, a power transformation can give it the properties of a distance that is useful for methods of classification.

Similarly, it can be demonstrated that if d is a distance, one can always find a value α between 0 and 1 such that for any $\lambda(0 \leq \lambda \leq \alpha)$, d^{λ} is a Euclidean distance equivalent in order to d. A power transformation can thus allow the passing from a distance to a Euclidean distance just as one passes from a dissimilarity to a distance. A useful application involves factorial analysis. Although they apply only to Euclidean distances, these methods can be used for any dissimilarity after an appropriate power transformation. This transformation seems preferable to the usual technique of adding a constant, possibly very large, to each pair distance, a technique less meaningful with respect to the initial data.

In some cases, the α values of this transformation are known; for example, it has already been indicated that the square root ($\alpha = \frac{1}{2}$) of a city block distance is a Euclidean distance. Often it is not known how to establish this value, which must be estimated numerically on the data.

Ultrametric and additive tree distance

In adding supplementary constraints to the definition of a dissimilarity, we can give it the property of being a distance that can be represented as a tree. The classification methods all aim to approach as closely as possible, in terms of some criterion, the dissimilarity observed by one of these representable distances.

The best known of these dissimilarities is the ultrametric, a distance verifying the inequality $d(i,j) \leq \max[d(i,k), d(j,k)]$ for any three individuals i, j and k. This property expresses that, among the three distances, the two largest are equal: if $d(i,j)$ is the smallest, then $d(i,j) \leq d(i,k) = d(j,k)$. Any triplet of points thus forms a sharp isosceles triangle. This property allows a representation in the usual form of a dendrogram (Fig. 1a), which is a tree that has a particular point, the root, located at an equal distance from all the leaves of the tree. It is shown that an ultrametric is a Euclidean distance.

In the 1970s, the distance called the additive tree distance was proposed. It verifies the inequality $d(i,j) + d(k,l) \leq \max[d(i,k) + d(j,l), d(i,l) + d(j,k)]$ for any four individuals i, j, k and l. This property, called the four points property, is an extension of the ultrametric condition to four points. It expresses that, among the three sums of distances two by two, the two largest are equal. This property allows a tree representation. It is verified on an example of four points (Fig. 2) that the condition is effectively respected since $d(i,j) + d(k,l) = a_i + a_j + a_k + a_l$ and $d(i,k) + d(j,l) = d(i,l) + d(j,k) = a_i + a_j + a_k + a_l + 2a_c$. This tree can be represented in a hierarchical form such as a dendrogram (Fig. 1b). However, in the absence of an objective root, a radial representation is often preferred (Fig. 1c).

The ultrametric condition is shown to be only a particular case of the four points condition. The ultrametric appears thus as an additive distance of a particular tree subjected to a more narrow constraint. This supplementary constraint is that, among three individuals, the two largest distances are equal, a constraint absent in additive tree distance, which allows for unequal branch lengths. This lesser constraint thus enables a more faithful representation of initial dissimilarities.

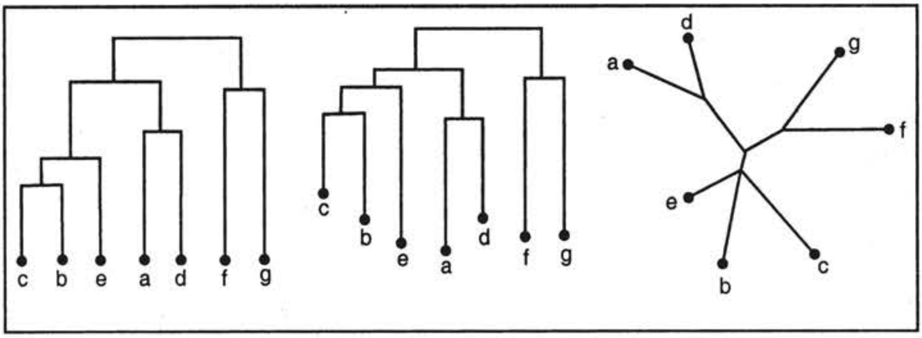

Fig. 1. Ultrametric (a) and additive distance in hierarchical representation (b) or radial representation (c).

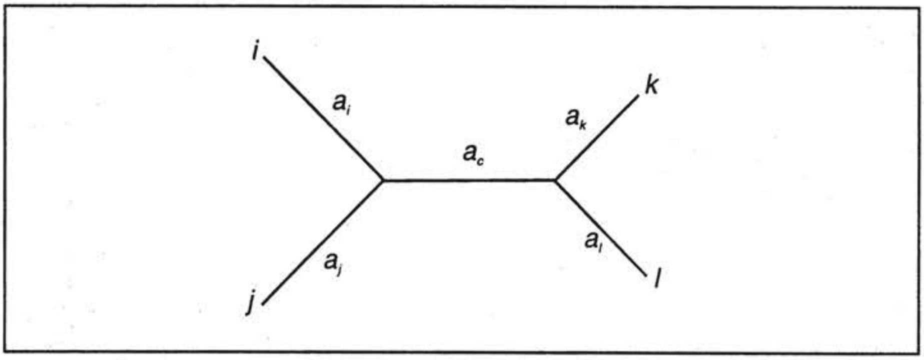

Fig. 2. Tree with four leaves and edge lengths.

INCLUSION RELATION BETWEEN FAMILIES OF DISSIMILARITIES

Different families of dissimilarities have been established by adding supplementary constraints from the initial definition of a dissimilarity. This induces relations of inclusion between the different families (Fig. 3). The ultrametrics, for example, are Euclidean distances and particular additive distances. They are themselves city block distances, which in turn are distances, a particular case of dissimilarities. When we look at Fig. 3 from top to bottom, we find indexes supporting conditions that are increasingly strong and thus less and less apt to describe accurately the relations between individuals. Conversely, from the bottom to the top, the indexes lose their properties of representability. The ultrametrics and the additive distances can be represented in the form of a tree in two dimensions, and the Euclidean distances and city block distances can be represented in spaces of higher dimension. On the contrary, distances and dissimilarities have no useful properties for this purpose. Thus, among the representable distances, the city block distance appears to be the least constrained. In the absence of a contrary indication, the choice must logically include such an index.

In practice, a factorial analysis on a Euclidean distance and a tree representation are often used in parallel. The level equivalent to the Euclidean distance and the additive distance can be noted on the graph. The coherence would thus require that the tree representation be founded on an additive distance rather than on an ultrametric, as is often the case.

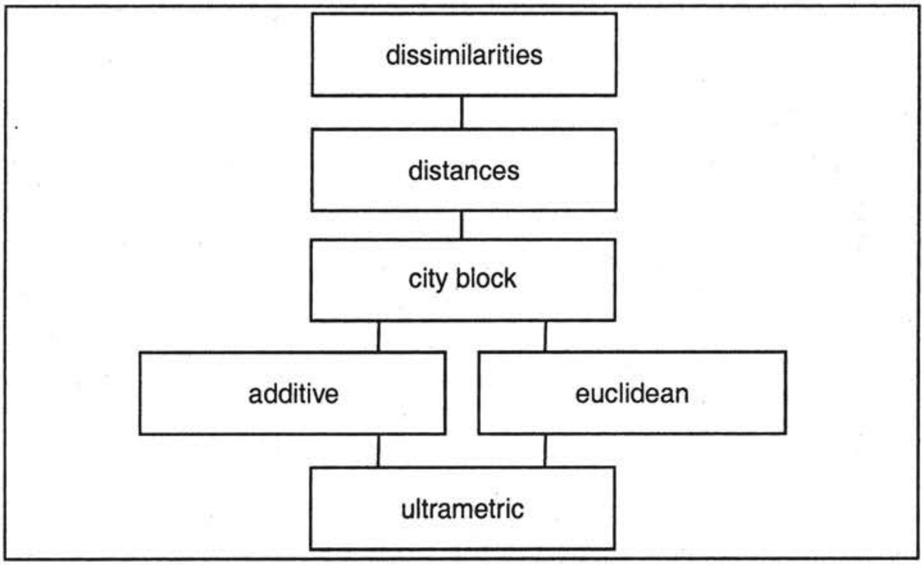

Fig. 3. Relations between groups of dissimilarities (Critchley and Fichet, 1994).

Measures of Dissimilarities on Quantitative Variables

The dissimilarity most often used on quantitative data is the Euclidean distance $d(i,j) = [\Sigma_k (x_{ik} - x_{jk})^2]^{1/2}$ (x_{ik} is the value of the k^e variable for the individual i). It is classical in statistics and can be treated by factorial methods. If the variables are of different nature, it is often useful to reduce them and, to render the value of the distance independent of the number P of variables, one often balances it with $1/P$. From this we get the general expression $d^2(i,j) = (\Sigma_k [(x_{ik} - x_{jk})/\sigma_k]^2)/P$.

The mathematical advantage of the city block distance, $d(i,j) = \Sigma_k \mid x_{ik} - x_{jk} \mid$, has been emphasized. It has been observed that the Euclidean distance, when the differences are squared, gives a heavy weight to significant deviations. This is not always inherently justified; the city block distance, which gives the same weight to all the differences, thus seems preferable. This distance is generally balanced by the number P of variables and, if the order of magnitude of variables is very different, it is necessary to bring them to the same scale, for example by relating them to their amplitude.

Measures of Dissimilarities on Presence/Absence Variables

The descriptors of genetic diversity are often binary variables, which represent the absence or presence of a character. For biochemical or molecular descriptors, these characters in general code the presence or absence of a band on an electrophoresis gel. We conventionally use 0 to denote absence and 1 to denote presence.

Several measures of resemblance between individuals, or association coefficients (Sneath and Sokal, 1973), have been defined on binary data. They have most often been proposed by researchers in a particular discipline—botany, zoology, palaeontology—and are justified mainly because they accurately translate the idea these researchers have of the resemblance between their subjects of study. We will limit ourselves here to discussing indexes for which we can explain the logic of construction and which present a reasonable behaviour (Beaulieu, 1989). For example, if a new marker is added, the dissimilarity between two individuals must increase (or decrease) if these two individuals take different (or identical) values for this marker.

For two individuals i and j, a is the number of markers that are simultaneously present in i and j. Similarly, d is the number of absences in common, b is the number of presences in i and absences in j, and c is the number of absences in i and presences in j. Table 1 lists 13 of these indexes. They are customarily expressed in the form of the similarity S, dissimilarity being obtained for these indexes by $D = 1 - S$.

The general principle of construction of all these indexes is the same, the similitude is estimated by the number of agreements. However, this value does not have an absolute meaning and must be reported with a basis of

Table 1. Major indexes of similarity on presence/absence variables

	Authors	Expression	Properties*			
			(1)	(2)	(3)	(4)
S1	Russel and Rao	a/P	y	y	y	
S2	Simpson	$a/\min[(a+b), (a+c)]$	n		n	
S3	Braun-Blanquet	$a/\max[(a+b), (a+c)]$				
S4	Dice	$a/[a+(b+c)/2]$				
S5	Ochiai	$a/[(a+b), (a+c)]^{1/2}$	n		n	S7, S8
S6	Kulczynsky 1	$(a/2)([1/(a+b)]+[1/(a+c)]$	n		y	
S7	Jaccard	$a/(a+b+c)$	n		n	
S8	Sokal and Sneath un2	$a/[a+2(b+c)]$	y	y	y	S4, S8
S9	Kulczynski 2	$a/(b+c)]$	y	y	y	S4, S7
S10	Sokal and Michener	$(a+d)/\mathrm{P}$				
S11	Rogers and Tanimoto	$(a+d)/[a+d+2(b+c)$	y	y	y	S11, S12
S12	Sokal and Sneath un1	$(a+d)/[a+d+(b+c)/2]$	y	y	y	S11, S12
S13	Sokal and Sneath un3	$(a+d)/ (b+c)$	n		n	S10, S11

*(1) The associated dissimilarity is (y) or is not (n) a distance. (2) It is (y) or is not (n) a city block distance. (3) Its square root is (y) or is not (n) Euclidean. (4) It is equivalent in order to the indexes indicated (absence of mention indicates that the response is not presently known).

comparison. The indexes differ in their mode of estimating the number of agreements and in the choice of the basis of comparison.

The estimation of agreements depends on the meaning assigned to the absence modality. If only modality 1 is considered informative, modality 0 expresses mainly an absence of information; then the number of agreements is a, the number of presences in common (indexes S1 to S9). If 0 and 1 are informative and can be considered as two modalities of a qualitative variable, then the number of agreements is $a + d$, the number of presences and absences in common (indexes S10 to S13). The choice between these two attitudes depends entirely on the nature of characters analysed and is a prerequisite to any reflection on the choice of a dissimilarity index. However, it is not always easy to separate these two points of view. Regarding genetic markers of diversity, it is clear that biological knowledge of the markers being considered will enable us to choose the most appropriate model.

Agreements estimated by presences in common: Indexes S1 to S9

The numerator of these indexes is always a. The denominator should be an estimation of the number of agreements that two identical individuals will present. For index S1 this number is P, the number of variables. This choice is not judicious since two individuals can thus be identical only if they have the value 1 for all the variables. It is better to estimate the denominator from numbers of presences in i, $a + b$, and in j, $a + c$. Of course, these two values are not equal, and a consensual expression must be defined for them. We can take the minimum (S2), the maximum (S3), or a series of values between these two extremes: the arithmetic mean (S4), the geometric mean (S5), and the harmonic mean (S6). The knowledge of characters can sometimes guide the choice of one of these indexes. However, in the absence of a clear justification, we prefer to retain the neutral behaviour of the Dice index (S4) based on the arithmetic mean.

Another approach is to consider that the basis of comparison is the number of presences found in i or in j, $a + b + c$, whence the Jaccard index (S7). With the S7 index, as with S1, a is compared to the number of variables, but the double absences are treated as missing data. This point of view is, in a good number of cases, highly reasonable and explains the success of the S7 index, which is certainly one of the most widely used.

Index S8, which is used often, and index S9 seem difficult to justify. Note that indexes S4, S7 and S8 can be written in the form $S = a/[a + \beta(b + c)]$ with, respectively, $\beta = ½$, $\beta = 1$, and $\beta = 2$. We could construct new indexes by choosing other values for β.

Agreements estimated by presences and absences in common: Indexes S10 to S13

Indexes S10 to S13 are obtained by extension of indexes S1, S8, S4 and S9, simply replacing a by $a + d$. On the evidence, these indexes must be symmetrical in a and d, the notation 0 and 1 being purely arbitrary. The extension of other indexes leads to indexes that are non-symmetrical in a and d, or sometimes non-symmetrical in b and c, and thus unacceptable. The most appropriate in many cases are index S10, an extension of S1 but also of S7, the Jaccard index, and S12, an extension of S4 based on the arithmetic mean. Indexes S12, S10 and S11 can be written in the form $S = (a + d)/[(a + d) + \beta(b + c)]$ with, respectively, $\beta = ½$, $\beta = 1$, and $\beta = 2$.

Properties of indexes S1 to S13

Complementary to the logic of construction, the possession of particular properties can guide the choice of an index. Table 1 summarizes certain properties of dissimilarities associated with the proposed indexes. A primary characteristic is the fact of being a true distance. It can also be shown that some of these indexes are city block distances. Although none of them is a Euclidean distance, it is known that city block distances have Euclidean square roots. Index S5, without being a city block distance, also has a Euclidean square root. These indexes, after transformation, can thus be treated by factorial analysis.

Finally, several of these indexes, which have comparable modes of construction, are equivalent in order. Between several indexes giving identical orders, those having useful properties are preferred, S10 or S11, rather than S12, for example.

The algorithms of tree construction are highly sensitive to small variations of dissimilarities. The behaviour of these different indexes faced with data errors is thus an important criterion for selection. This behaviour can be addressed by studying the properties of a matrix subjected to a random noise exchanging the values 1 with 0 and inversely, with a probability *t*. It is thus possible to estimate the expectation *D′* of the index considered for a noise of rate *t* of data.

Figure 4 presents, for indexes S10, S11, and S12, symmetrical in 0 and 1, the value of *D′* as a function of *D*. The rate of error used is 0.10, but that value does not change the meaning of conclusions that may be drawn. For index S11, data errors may lead to serious overestimations of small values of dissimilarity, while large values are relatively not modified. Index S12 has an inverse behaviour in leaving the small values relatively not modified and underestimating the large values. Index S10 has a more neutral behaviour; the bias is null for $D = 0.5$ and increases symmetrically. It is thus a useful compromise. However, the most widely used algorithms of tree construction are agglomerative; the smallest dissimilarities determine the first groups, which then determine the entire tree. It is thus preferable to minimize the noise on the small values, as with index S12.

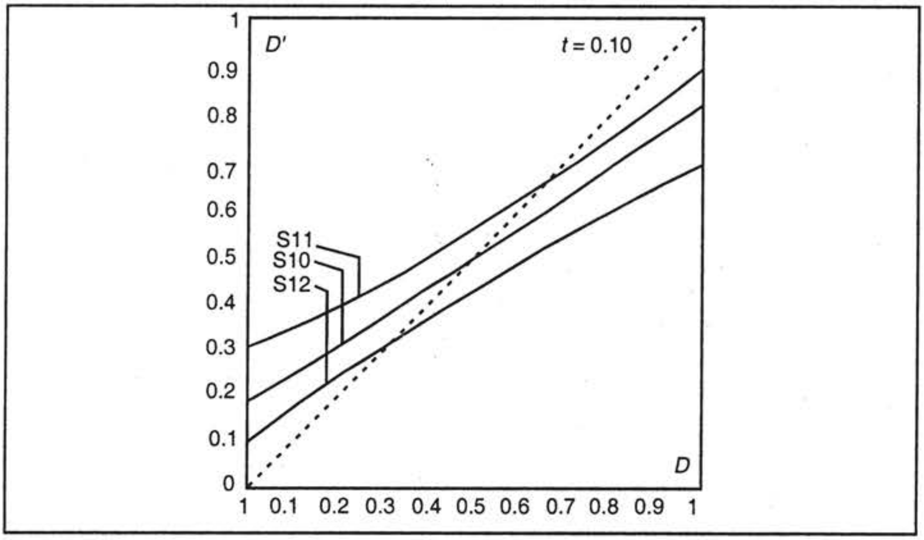

Fig. 4. Estimation by simulation of dissimilarities *D′* associated with S10, S11 and S12 for an error rate of 0.10 as a function of initial dissimilarity *D*.

The study of indexes that estimate the agreements by presences in common is more complicated. The distortions observed are greater than for the symmetrical indexes. They can be very high (up to 50% for the small values of *D*, even with rates of error of 0.05), when the frequencies of 0 are high (Fig. 5). This situation is frequent with highly polymorphic molecular

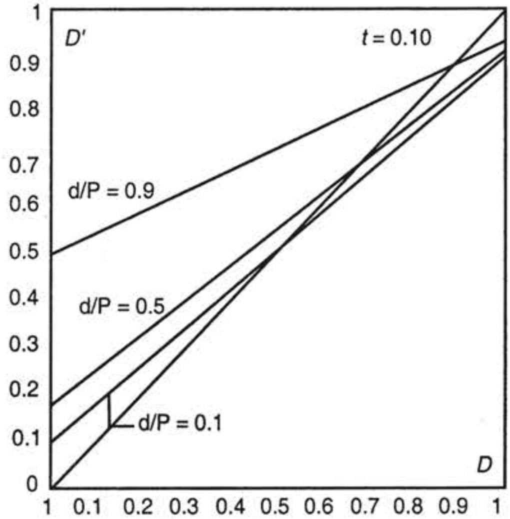

Fig. 5. Estimation by simulation of dissimilarities *D'* associated with index S7 (Jaccard) for an error rate of 0.10 as a function of initial dissimilarity *D*.

markers, for which the number of 1 for each individual is very low. This type of matrix must therefore be used with some caution, and the stability of structures must be verified by comparing results obtained on the complete matrix and free of particularly rare markers. Such markers lead to an increase in the number of absences in common for most of the pairs of individuals. Indexes S4, S5 and S6, corresponding to three types of mean for the denominator, have a closely comparable behaviour. They tend to give smaller distortions of small values of dissimilarity than the Jaccard index, S7, which in turn is more favourable than S8. These two indexes are thus preferred.

Measures of Dissimilarities on Qualitative Variables

Here the characters are of qualitative variables that present a finite number of modalities: petal colour, form of stigmata, and so on. Intuitively, two individuals are closely related if they have the same modality for a large number of variables.

The usable indexes correspond to the generalization of indexes presented for the binary values symmetrical in 0 and 1, which are in fact qualitative variables of two modalities.

The number of variables in agreement is m, the number of variables in disagreement is u, and P is always the number of variables. The corresponding indexes of similarity are written as follows:

$m/(m + u) = m/P$ (Sokal and Michener);
$m/(m + 2u) = m/(P + u)$ (Rogers and Tanimoto);
$2m/(2m + u) = 2m/(P+ m)$ (Sokal and Sneath).

The dissimilarity is obtained directly by $d = 1 - s$.

The χ^2 distance is often used on a binary table. For one variable, and for each of these modalities, a binary is created that takes the value 1 if the individual presents this modality, 0 if not, and the classic χ^2 is calculated on this new data matrix. This distance does not give the same weight to all the modalities, particularly to rare modalities, which are weighted heavily. Although such an effect can be sometimes desirable, it often is not.

Choice of Index of Similarity on Biochemical and Molecular Markers

The genotype resemblance between two individuals can be measured by a genetic similarity defined from allele forms of genes observed. Certain types of markers (e.g., isozymes, microsatellites) generally give direct genetic information and allow coding in alleles by identification of the allele composition of each locus. The genetic similarity, as defined here, can thus be directly estimated. Other markers do not allow access to all of the genetic information and only one phenotype is observed. The genetic similarity cannot be calculated directly, but the characteristics of these markers be used to define the most relevant similarity index. Moreover, it is useful to evaluate the order of magnitude of the error related to this loss of information.

Multiple allele markers

The genetic similarity T_{ij} between two individuals i and j is defined *a priori* as the mean on L loci of the ratio of the number n_{ls} of alleles for the locus l present simultaneously in the two individuals and of the number n_{lc} of alleles compared: $T_{ij} = (\Sigma_l\, n_{ls}/n_{lc})/L$. If π is the ploidy, n_{lc} would be π for all the loci. When $n_s = \Sigma_l\, n_{ls}$ and $n_c = \pi L$, then $T_{ij} = n_s/n_c = n_s/(\pi L)$.

For a diploid species, each allele of a locus can be coded as a variable taking the values 2, 1 or 0. Following this coding for two individuals i and j, the combination of genotypes of each locus belongs to one of seven groups, named A to G (Table 2). The number of loci of each group is designated I_A to I_G. We can thus write T_{ij} from $n_s = 2I_A + I_B + 2I_E + I_F$ and $n_c = 2(I_A + I_B + I_C + I_D + I_E + I_F + I_G)$.

The parameters $n_{r,s}$ ($r \geq s$) are defined as the number of alleles present r times in one individual and s times in the other. Each of the groups A to G contributes to these parameters. For example, for a locus with a_l alleles, a

Table 2. The seven possible combinations of genotypes of two diploid individuals

	A	*B*	*C*	*D*	*E*	*F*	*G*
Individual *i*	20..	200..	200..	2000..	110..	110..	1100..
Individual *j*	20..	110..	020..	0110..	110..	101..	0011..

pair of group D participates one time at $n_{2,0}$, two times at $n_{1,0}$, and a_l — 3 times at $n_{0,0}$.

The numbers I_A to I_G can be expressed from these parameters $n_{r,s}$, from which we can write T_{ij} as the ratio of $n_s = 2n_{2,2} + n_{2,1} + n_{1,1}$ and $n_c = (2n_{2,2} + n_{2,1} + n_{1,1}) + (n_{2,1} + 2n_{2,0} + n_{1,0})/2$.

In the expression of T_{ij}, therefore, a generalization of S4, the Dice index, can be recognized. The number of agreements is effectively 1 for the genotypes contributing to $n_{2,1}$ and $n_{1,1}$ and 2 for those contributing to $n_{2,2}$. Similarly, the disagreements are1 for $n_{2,1}$ and $n_{1,0}$ and 2 for $n_{2,0}$. It can be noted that for haploid species, some microorganisms, for example, we can directly find the Dice index.

The Dice index belongs to the family of indexes that do not take into account the information carried by double absences. This point of view is logical since the number of double absences does not carry information on the proximity of individuals and depends only on the number of alleles of the loci.

Codominant markers coded in bands

For markers of the RFLP type or the isozymes, it is sometimes impossible to identify alleles belonging to the same locus and there is only coding in bands. For a diploid species, the presence of a band, coded 1, can thus correspond to a homozygous locus, which must be coded 2, or to one of the alleles of a heterozygous locus, normally coded 1. The different groups of genotypes defined for multiple allele markers are not identifiable and the phenotype observed differs from the real genotype (Table 3).

Table 3. The seven possible combinations of genotypes and phenotypes observed for codominant markers

	A	*B*	*C*	*D*	*E*	*F*	*G*
Genotype *i*	20..	200..	200..	2000.	110..	110..	1100..
Genotype *j*	20..	110..	020..	0110.	110..	101..	0011..
Phenotype *i*	10..	100..	100..	1000.	110..	110..	1100..
Phenotype *j*	10..	110..	010..	0110.	110..	101..	0011..

As in the preceding case, the information of double absences is not relevant and, by analogy with the Dice index, one can retain, as a measure of resemblance, an index of the group $S_{ij} = n_{1,1}/(n_{1,1} + \beta n_{1,0})$. The value of β is chosen such that, under certain hypotheses, S_{ij} more closely approaches T_{ij}.

For an autogamous diploid species that will be homozygous for all the loci, $S_{ij} = T_{ij}$ for $\beta = ½$; S_{ij} is thus directly the Dice index.

For a given species, a value of t, the rate of average heterozygosity, is generally known. Considering that this value may be applied at each locus, we can define the value of β, function of t, that annuls $S_{ij} - T_{ij}$. The expression of this difference is complex and the relations between β and t have been researched numerically for different distributions of allele number per locus, the maximum being fixed at 7 (Fig. 6). The distributions I, II and IV correspond to extreme distributions, and the distribution III reproduces a distribution observed on hevea (M. Seguin, personal communication).

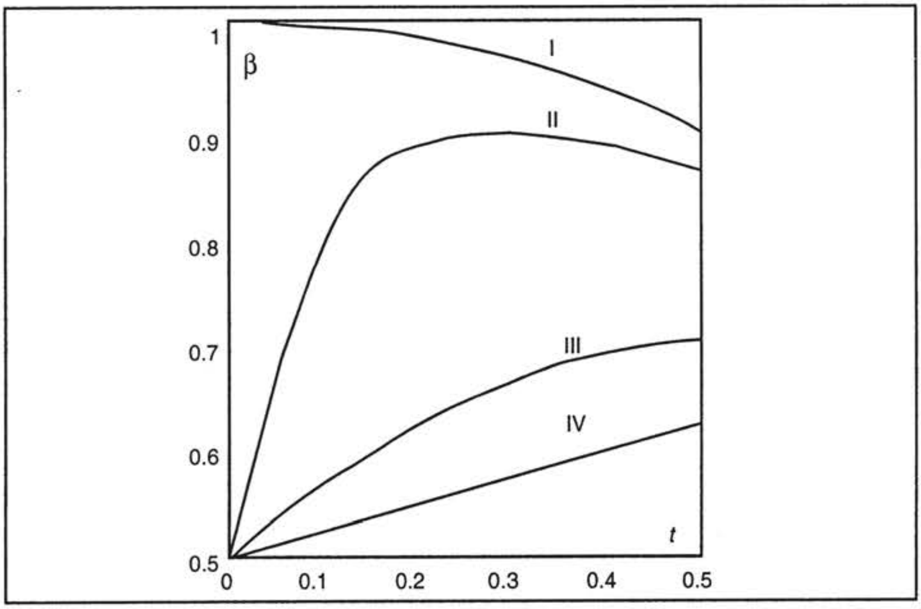

Fig. 6. Values of the index β of the dissimilarity $D = a/[a + \beta(c + d)]$ such that D_{ij} is equal to the genetic dissimilarity T_{ij} calculated from the complete genetic information, as a function of rates of heterozygosity t and for various distributions of number of alleles:
I : all the loci have 2 alleles;
II : loci with 2 or 3 alleles in proportions 0.9 and 0.1;
III : loci with 2 to 7 alleles in proportions 0.27, 0.32, 0.19, 0.11, 0.07 and 0.03;
IV : all the loci have 7 alleles.

For loci with a small number of alleles (I or II), the genetic similarity is approached, for low heterozygosity, only for values of β close to 1.

For a more realistic distribution as distribution III, one can propose to fix the value of β at 0.5 for $t \leq 0.1$, at 0.6 for $0.1 < t \leq 0.3$, and at 0.7 for $t > 0.3$.

The order of magnitude of the difference has been evaluated by simulation for different rates of heterozygosity and a series of values of b between ½ and 1, from a population of 40 individuals described by 200 loci. The deviation $S_{ij} - T_{ij}$ is composed of a systematic bias, without consequence on the methods of analysis, and of a true error that is the only relevant point here. This error is expressed by the half-deviation between the lowest and the highest of the values observed in the simulations; it is expressed as a percentage of the mean genetic similarity. For a type III distribution, this maximal error is of the order of 3% for t close to 0.2, but it reaches 5% for highly heterozygous species.

It can be demonstrated, moreover, that this error depends mostly on t and very little on β. In practice, we can most often retain $\beta = 0.5$, corresponding to the Dice index.

Dominant markers coded in bands

Dominant markers, RAPD or AFLP, can be considered loci with two alleles, since a single allele can be located, the other being a null allele that is not materialized on the gel. These markers are said to be dominant, because it is impossible to know the number of copies of an allele for a particular locus and thus to distinguish the homozygotes from heterozygotes.

Among the types *A* to *G* defined previously for a diploid species, only *A*, *B*, *C* and *E* exist, since the loci have only two alleles. The two alleles do not play an identical role. Genotypes *A* and *B* must be subdivided according to the phenotypes, distinguishing the located allele from the null allele. The impossibility of coding into alleles leads, as for RFLP markers, to a phenotype designated theoretical in Table 4. As the null allele, noted by a point, cannot be located, only one of these allele states is thus readable, from which we get the phenotype that is actually read.

Table 4. Genotypes, theoretical phenotypes, and phenotypes observed for dominant markers

	A1	*A2*	*B1*	*B2*	*C*	*D*
Genotype *i*	20	02	20	02	20	11
Genotype *j*	20	02	11	11	02	11
Theoretical phenotype *i*	10	0.	10	0.	10	1.
Theoretical phenotype *j*	10	0.	1.	1.	0.	1.
Read phenotype *i*	1	0	1	0	1	1
Read phenotype *j*	1	0	1	1	0	1

The number *L* of loci is the number *P* of different bands observed in the population. We must note immediately that $n_{0.0}$, the number of double absences, here has a clear meaning. It corresponds to the number of homozygous loci for the null allele in the two individuals and carries as much information as the number of double presences. It will be logical to retain an index symmetrical in 0 and 1. The denominator of the genetic similarity T_{ij} is equal, in this particular case, to *P*. This value leads us to retain the Sokal and Michener index, here expressed by $S_{ij} = (n_{1,1} + n_{0,0})/P$.

The difference $S_{ij} - T_{ij}$, which measures the error of estimation of genetic similarity, is expressed simply by $(I_{B1} - I_{B2})/P$. This difference is nil if the frequencies of genotypes (2,0) and (0,2) are equal, that is, if the frequencies of the two alleles are identical for all the loci; it remains low if they tend to be identical, on average, for all the loci. On the other hand, if the species is highly homozygous, the proportions of genotypes *B* are low and $I_{B1} - I_{B2}$ remains low.

For more heterozygous species, this difference can be high since the genotypes *B* can represent up to 50% of the population. As before, numerical simulations are used to estimate the order of magnitude for different levels of heterozygosity and for different proportions of the two alleles of a locus. The true error is, as before, the demi-amplitude of extremes observed related to the mean genetic similarity. It increases with heterozygosity and reaches 13% for $t = 0.5$ and no allele deficit, goes beyond 18% if the non-identifiable alleles are in a majority, and falls to 10% in the opposite case. In any case, the imprecision induced by the loss of genetic information is clearly greater than for the codominant markers.

In conclusion, it must be emphasized that the informative content of markers must determine the choice of an index of similarity, a criterion that is often missing in most publications in this field.

METHODS OF FACTORIAL ANALYSIS

Methods of factorial analysis aim to discover the strongest structures in populations that are being studied and to eliminate the occasional peculiarities that hamper the general perception of phenomena. They can therefore be very useful in a study of the diversity of a species or a population.

The aim of these methods is to produce a geometric representation of measures of differences between units. Such a representation allows us to manipulate specifically the notion of diversity. We thus get a hierarchical composition of the diversity, which allows us to distinguish the basic tendencies of various characteristics.

The use of these techniques requires a knowledge of the principles they are based on and the conventions they use. After a general presentation of factorial analysis emphasizing its conventional aspects, we will see how it integrates specially in three widely known applications: principal coordinates analysis (PCoA), principal components analysis (PCA), and multiple correspondence analysis (MCA).

Then we will discuss the treatment of multiple tables, a problem found particularly in the study of a population described by several types of markers. When there is information coming from different points of view about the individuals, we can use multiple factorial analysis (MFA) to reveal the major characters, which are manifested in a consensual manner, and those that on the other hand are specific to one type of measure.

Factorial Analysis

CLOUD AND POINTS OF DISPERSAL

If the distance chosen to measure the differences between the units studied is Euclidean, the group of units can be represented in the form of a cloud of points in a vast space. In this space one can mark out a point B, called the mean point, or barycentre, which is defined as being as close as possible to all the points of the cloud. It corresponds to a neutral 'mean type' of which each real unit differs by its own characteristics.

In these conditions, the total diversity of the population is represented by the dispersal of the cloud around B. Factorial analysis quantifies this dispersal by the inertia of the cloud in relation to B, that is, by the sum of squares of distances from each point to B.

Principal axes of inertia and coordinates of individuals

In general, the space in which the cloud is found is a space of high dimension. Practically, this means that in leaving B one can go in a multitude of directions. In other words, there are many ways for an individual to be distinguished from the mean type and all the units observed are different from it, each in its own way. The aim of factorial analysis is to bring any peculiarity back approximately to the composition of a small number of directions that are very commonly used and independent from each other.

To determine the 'frequency of use' of a direction *U*, one constructs a criterion by summing for all the individuals a value that is large enough so that the angle that *U* makes with the specific direction of the individual is acute and so that the individual is at a great distance from the mean point B. This criterion is called the inertia of the cloud along *U*. The direction that maximizes this criterion is called the principal axis of the inertia of the cloud. Factorial analysis successively produces other axes by choosing for each the direction that maximizes the criterion of inertia among the orthogonal directions of the axes already identified. The orthogonality of axes ensures the independence of variations summarized by them.

The result of this process is a system of independent axes successively explaining the maximum of variation. The coordinates of each unit can be calculated on these axes and they can thus be placed in the orthogonal system. The set of coordinates of all the individuals on an axis is called a *factor*.

From the factorial coordinates, the total dispersal of the population can be reconstituted. However, and this is the essential advantage of factorial analysis, the selection of just the primary factors will enable generally the reconstruction of a large part of the total dispersal. Several relatively empirical criteria have been proposed to determine the number of axes to be conserved (Saporta, 1990). However, it is advisable to remain circumspect and the choice of axes must be reasoned case by case.

In the absence of other information, the coordinates of individuals on the axes have little significance. For a factorial analysis to be interpretable, there must be an 'external' source of information available that one can link to other factorials. This external information may not be explicit and may appear only when one observes that such an axis resembles or, on the contrary, opposes certain recognizable individuals. Most often, however, an axis will be interpreted through observation that it is linked to variables measured on the individuals, and that they have participated or not participated in the calculation of the distance matrix.

Only when an axis is interpreted and when the dispersed part attached to it is comprehensible do we consider that there is a factor that must be taken again into account in the elaboration of the diversity.

Graphic representations

To provide a visual representation of diversity, but also to facilitate the interpretation of axes, the methods of factorial analysis propose graphic representations of individuals and, if necessary, of variables. The graphs are critical inputs of these methods. They allow us to see, literally, the oppositions, groupings, and trends that are difficult to perceive from enumerated statistics. However, each type of graph must be read according to particular rules, and it is advisable to master those rules to avoid any chance misreading.

Classic Methods of Factorial Analysis

In practice, factorial analysis of the kind discussed here is only one step in a process that starts with a table of data, operates possible coding, calculates distances if necessary, and finds factorial axes produced with interpretative aids adapted to the nature of the data. It is the entire process that forms a 'method' of factorial analysis. According to the type of data available, the most commonly used methods are: PCoA for a table of distances, PCA for a table of quantitative variables, and MCA for a table of qualitative variables.

Principal coordinates analysis

Principal coordinates analysis is used to treat a matrix of distances calculated with an original index, chosen in a manner adapted to its own data. It is, however, necessary that the distance obtained be a Euclidean distance or a distance rendered Euclidean by transformation.

In this case, two individuals i and j can be represented by two points e_i and e_j of a space of unknown dimension K such that the square of the distance d_{ij} can be written as follows: $d_{ij}^2 = (e_i - e_j)'(e_i - e_j)$.

In terms of calculation, it is demonstrated that the successive axes of the factorial analysis and the inertia of the cloud on each axis correspond to vectors and values of the matrix W of scalar products between elements. If the origin of the coordinates is placed at the midpoint of the cloud, the scalar products $w_{ij} = <e_i; e_j>$ are entirely determined by the d_{ij} according the formula of Torgerson: $w_{ij} = -(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d^2)/2$.

The diagonalization of this matrix W allows extraction of the eigenvectors and eigenvalues associated with them.

The number of eigenvectors associated with a non-null eigenvalue is the dimension K of the space containing a cloud of points.

The PCoA is the simplest use of factorial analysis. In these conditions, the only aid to interpretation available is the graph of individuals on any choice of two factorial axes.

Principal components analysis

Principal components analysis treats quantitative variables and imposes a Euclidean distance between individuals. The Euclidean distances that can be calculated differ only in the weight assigned to the variables. The most common weightage system consists of bringing all the variables to the same scale and weighing them by the inverse of their standard deviation. The PCA is thus said to be normal, or centred-reduced.

As with PCoA, the coordinates of individuals on the principal components can be calculated from the matrix of scalar products. This can be obtained immediately from a table of centred data.

The interpretation of axes provided by a PCA is more rich because correlations can be calculated between the factors and the variables of the table. From examination of these correlations for each axis, we can reveal the variable or variables most closely linked to the axis. The bundle of variables that vary conjointly can thus be revealed; the conjoint variation allows a distinct discrimination of individuals.

Multiple correspondence analysis

Multiple correspondence analysis (MCA) allows the study of individuals described by qualitative variables with modalities. This method calculates distances between individuals by the χ^2 distance on the binary table (see above). That is, two individuals are distant from each other to the extent that they present numerous disagreements and that their disagreements cause one or another to adopt rare modalities.

This method is also known as multiple correspondence analysis (MCA), correspondence analysis on a binary table (CA on BT), or simply correspondence analysis (CA).

The interpretation of axes provided by factorial analysis of this matrix of distances is deduced from relations that they have with three types of objects: individuals, variables, and modalities. The individuals provide the same interpretations as in other methods of factorial analysis.

As with PCA, one can evaluate the link between the variables of the table and each axis. For a qualitative variable, this link is increased by the ratio η^2, which is equal to the part of the total inertia of the axis reconstituted by the sum of squares of the intermodality deviations. One factor can thus be perceived as a numerical synthesis of a certain number of variables that vary conjointly.

A qualitative variable is a more complex object than a quantitative variable and the diversity that it induces manifests itself on as many axes as the variable has modalities. A more refined interpretation of axes and of liaisons between variables must thus be looked for at the level of modalities. However, the choice of the χ^2 distance determines that the considerable contribution of a modality to an axis can only be due to the rarity of this modality and thus to the eccentricity of the small number of individuals that possess it. When this is not the case, one axis reveals a set of modalities in mutual association, that is, present or absent simultaneously in a large number of individuals.

Regarding graphic representations, the χ^2 distance means that when the individuals and the modalities are plotted on a single graph, an individual is close to modalities that it possesses and a modality is close to the individuals that possess it.

Treatment of Multiple Tables and Multiple Factorial Analysis

Conjoint analysis of several tables

Faced with data of various kinds, we may wish to combine the information contributed by the different tables in such a way as to clarify the finer structures in the population.

An organization of individuals that appears during the analysis of a certain table and that is clearly due to variations of certain important characters of this table may sometimes manifest itself only during analysis of another type of data. One approach could be to make a conjoint analysis of a supertable juxtaposing the different data. All the possible oppositions can thus be made apparent; those that are common to all the types of data and those that, on the contrary, are specific to a given table can be revealed.

There are, however, two major obstacles to the immediate analysis of such a supertable:

(1) The type of variable—qualitative or quantitative—is not necessarily the same for all the tables. It is advisable to analyse the qualitative tables beforehand by MCA and the quantitative tables by PCA.
(2) All the tables do not have necessarily the same inertia and do not clearly express the same organization of the diversity. The most highly structured tables may thus influence the results excessively, and the input of the more loosely structured tables may go unnoticed.

A useful result allows us to overcome the first obstacle: it can be demonstrated that one can obtain the same factors as those produced by MCA by making a PCA, weighted appropriately, of the table formed by the indicators of the qualitative variables, that is, of the table coded in a binary form. In the case of molecular data, this means that each marker must be represented by two columns: one indicating its presence and the other its absence.

The second obstacle is overcome by adopting a weighting of tables that reduces the influence of tables that are highly structured. For that purpose, we must quantify beforehand the part of the total inertia of the table that corresponds to its interpretable structure. Such a quantification is relatively arbitrary and this is one of the points that distinguishes the different methods of analysis of multiple tables. The Statis method (Lavit, 1988) uses as a measure the sum of squares of eigenvalues of the separate analysis of the table. The MFA (Escofier and Pages, 1993) prefers the largest of these eigenvalues. Without going into detail, we can say that the first method can reduce the significance of a table when, by chance, the last eigenvalues are high. The second may on the other hand give more weight than should be given to a table in which the first eigenvalues are not very different from each other.

Multiple factorial analysis

Multiple factorial analysis is thus a particular PCA, in which all the variables in a table are weighted by the inverse of the variance of the principal axis of inertia of the separate analysis of the table.

As does PCA, MFA can be used to locate individuals that resemble each other for the set of variables and thus to make a typology of individuals. It also enables us to attribute the passage from one class of this typology to another to the conjoint variation of some variables. It is these common directions of variation that MFA produces as axes of inertia, under the name of axes of global analysis.

However, the organization of data in the form of juxtaposition of tables enriches the interpretation of MFA. On the one hand, MFA provides indexes for each axis of overall analysis that allow us to determine whether it is due only to variables of a single table or whether it is common to several types of data. On the other hand, the MFA calculates the correlations between the axes of global analysis and the axes that will produce separate analysis of each table. When the preceding criterion indicates that a certain axis of the global analysis is common to several tables, this allows us to determine at what character precisely within each type of measure the common axis is linked.

Finally, from these correlations, we can find out, for a certain axis of a separate analysis whose meaning is known, the number of the axis of the global analysis to which it essentially contributes. If this global axis is common to several tables, that signifies that the partial axis manifests itself in a consensual manner. On the other hand, if the global axis is specific, that signifies that the partial axis corresponds to a character that is proper to the type of observations of separate tables.

MFA also has properties useful in terms of representation of individuals. In particular, it is shown to offer an interesting compromise between the quality of representation of clouds corresponding to each separate table and the proximity of these partial representations for a single individual.

TREE REPRESENTATION OF DISSIMILARITIES

The principle of any tree representation is to approach as closely as possible the dissimilarity δ, chosen for its relevance in describing the relationships between individuals, by a distance d that can be represented as a tree, that is, an ultrametric or an additive tree distance (Barthelemy and Guenoche, 1988).

To find the exact solution, we must enumerate all the possible tree configurations possible for n individuals. For each possible tree structure, the edge length of the tree is estimated in terms of least squares. Finally, we retain the tree that allows us to minimize the sum of squares of deviations between initial dissimilarity and distance reconstituted in the tree. It is shown, by recurrence, that the number of different binary trees constructed over n individuals is $\Pi_{i=3,n}(2i-5)$. For $n = 10$, there are more than 2×10^6 different trees and for $n = 20$ there are more than 2×10^{20}. It is thus impossible to enumerate all the trees when n goes beyond some tens, even with the most powerful computers.

In such situations, the only possibility is to construct, from reasonable heuristics, solutions that will be the best possible but that can never be guaranteed to be optimal. Various heuristics have been proposed, permitting the construction of algorithms of more or less great complexity, the complexity of an algorithm measuring the increase in calculation time when n increases. In the scope of this work, we have discussed only the methods of grouping that are of sufficiently low complexity to treat some hundreds of individuals.

The individuals studied are often described by several types of variables that cannot be combined in the calculation of a single dissimilarity. Each set of variables is thus treated separately and the presence of common structures in the different trees is examined. Two methods of constructing synthetic trees, consensus trees and the common sub-trees, will be presented.

Methods of Grouping

Methods of grouping are iterative methods that proceed by successive ascending agglomerations, constructing the tree step by step. Initially, the matrix treated has as many elements as individuals in the population studied and the tree has a star structure. At each iteration, two elements, individuals or groups already formed and defined as neighbours, are joined to form one group. They are chosen so that the tree that traces their grouping optimizes a fixed criterion. This group becomes a new fictive element that replaces the two combined elements; the matrix is updated and thus reduced by one unit. The process is reiterated until all the individuals are united in a single group.

The various methods of this type are characterized by different choices at three key points of each iteration: the selection of elements to be joined, which depends on the definition of 'neighbourhood' used, the updating of the dissimilarity matrix by calculation of a dissimilarity between the group formed and the other elements, and the construction of lengths of the two edges derived from the two combined elements.

Definition of neighbourhood

The most natural definition of neighbours is the two individuals or groups that have the least dissimilarity. The elements *i* and *j* are defined as neighbours if $\delta(i,j)$ is the smallest dissimilarity.

This criterion allows us to find the tree solution in a theoretical case in which the initial distance is already an ultrametric. But it does not necessarily allow us to find the right structure if the initial dissimilarity is an additive tree distance. In the example in Fig. 2, on just four points, we can imagine that the distance $d(i,k) = a_i + a_c + a_k$ is the smallest of distances even though *i* and *k* are in two pairs opposed by the central edge. This criterion necessarily groups *i* and *k* and does not allow us to find the true tree. For that, other criteria must be used that take all of the distances into account to judge a neighbourhood.

In the context of genetic diversity, Saitou and Nei (1987) proposed a criterion of neighbourhood based on the principle of parsimony, which is at the basis of the phylogenetic approach. The objective is to create a tree that will be of minimal total length. It can be considered that an edge represents a number of mutational events and, by virtue of the basic principle that evolution proceeds always by the simplest genetic modification, the number of events, and thus the total length, must be minimized. These considerations lead to a definition of relative neighbourhood, defined by minimizing a criterion $Q(i,j)$, function of $\delta(i,j)$ and of the average of dissimilarities of *i* and *j* at the $n - 2$ other elements *k*:

$$Q(i,j) = \delta(i,j) - (\Sigma_k\, [\delta(i,k) + \delta(j,k)])/(n - 2)$$

It can be demonstrated that this criterion has properties of optimality in terms of least squares and that its domain of application goes beyond the scope of the phylogenetic reconstruction. It can be interpreted, very generally, as a weighting of the dissimilarity between two individuals by their dissimilarities to other individuals. Two related individuals that differ considerably from other individuals resemble each other more than two related individuals that are equally related to other individuals. This attitude is justified in many cases and explains the success of the method.

Sattath and Tversky (1977) adopt a very different approach. They start from the characterization of an additive tree distance by the four points condition. For any four individuals *i, j, k* and *l*, if $d(i,j) + d(k,l)$ is the smallest of three sums of distances two by two, then the two largest are equal: $d(i,k) + d(j,l) = d(i,l) + d(j,k)$. The initial dissimilarity δ is not an additive tree distance but it is always possible to form the sums of dissimilarities two by two. Among these three sums, one is the smallest, $\delta(i,j) + \delta(k,l)$, for example; the pairs of points (i,j) and (k,l) are thus considered good candidates to be neighbours. They are assigned a score of 1, while other pairs of points, (i,k), (j,l), (i,l) and (j,k), are attributed a null score. All the quadruplets of points that can be formed on the n points are scanned and the individuals of the pair that has the largest total score are considered neighbours. This definition of topological neighbourhood is ordinal in nature, since only the order of the three sums is important, and not directly the dissimilarity values that constitute them. It seems particularly appropriate when the dissimilarity values are somewhat marred by errors.

Updating dissimilarities

When two groups, each possibly composed of a single individual, are combined to form a new group, it is necessary to define a dissimilarity between the new fictive element created and the other elements present at this iteration. Given i and j the groups combined to form a new element s, c_i and c_j the numbers of these groups, and k another element, the most natural definition of $\delta(s,k)$ is the arithmetic mean of dissimilarities between k and the individuals constituting i and j: $\delta(s,k) = [c_i\delta(i,k) + c_j\delta(j,k)]/(c_i + c_j)$. It corresponds to an unweighted criterion since all the individuals play the same role.

Often, an arithmetic mean is used that is said to be weighted, in the sense that a different weight must be attributed to individuals of the groups in order that the mean does not depend on the numbers of the groups and is written: $\delta(s,k) = [\delta(i,k) + \delta(j,k)]/2$. The choice between these two criteria depends on the nature of the population studied. If the whole arises from a real process of sampling on a given structure of diversity, then the number of individuals in each element has a significance and must be taken into account in the calculation of dissimilarity. If the whole is only a circumstantial group of units that is not representative, the number of each group is only the result of chance and is not to be considered in the calculation of diversity.

In the case of adjustment to an additive distance, the formula $\delta(s,k) = [\delta(i,k) + \delta(j,k) - \delta(i,j)]/2$ is used, or its weighted equivalent. In the reconstruction of edges, this formula can be used to find the correct edge lengths when the initial tree is already an additive distance. This modification has no influence on the choice of subsequent steps.

It has sometimes been proposed, for adjustment to an ultrametric, that criteria called simple link and complete link be used that correspond respectively to $\delta(s,k) = \min[\delta(i,k), \delta(j,k)]$ and $\delta(s,k) = \max[\delta(i,k), \delta(j,k)]$. Except in very particular cases, it is difficult to justify these criteria in problems of diversity. If the simple link has been studied a great deal, it is essentially because it induces particular mathematical properties in the ultrametric produced.

Reconstruction of edges

The elements i and j are combined to form the node s of the tree, and edge lengths $l(i,s)$ and $l(j,s)$ remain to be fixed. The dissimilarity $\delta(i,j)$ can simply be divided equally between the two edges: $l(i,s) = l(j,s) = \delta(i,j)/2$. This mode of calculation corresponds to an adjustment of the initial dissimilarity by an ultrametric.

If one does not impose an ultrametric condition, one accepts that the two edges can have different lengths for, on average, the dissimilarities of $n - 2$ elements k to i and to j to be represented as well as possible. The difference of edge lengths, designated e, is the sum on all the elements k of $[\delta(i,k) - \delta(j,k)]/(n - 2)$. From this we get the formulae to calculate the edge lengths $l(i,s) = [\delta(i,j) + e]/2$ and $l(j,s) = [\delta(i,j) - e]/2$. This mode of calculation corresponds to the adjustment of the initial dissimilarity by an additive tree distance.

These formulae of edge reconstruction do not guarantee that the estimations will be the best. Fortunately, the edge lengths are not involved in the subsequent steps. It is thus preferable to reestimate them afterwards, when the tree topology has been established. We know how to estimate the length of each edge of a given tree topology overall, in the sense of least squares. For this, we express the distance in the tree between two individuals i and j as the sum of edge lengths belonging on the way from i to j and we expect that this distance will be the closest, in the sense of least squares, to the initial dissimilarity. We thus resolve a system of $n(n - 1)/2$ equations with as many unknowns as edges. This reestimation will cause loss of the ultrametric property.

Some known algorithms

The most widely known algorithms correspond to certain compatible combinations among the different possible combinations of modes of definition of neighbourhood, various formulae for updating dissimilarities, and modes of estimating the edge lengths.

The set of methods often described as hierarchical clustering correspond to the definition of neighbourhood according to the minimal dissimilarity, with an adjustment to an ultrametric, and various formulae are proposed for updating. Among these, formulae of type average or weighted average are the most commonly used and the corresponding methods are frequently referred to as UPGMA and WPGMA, for unweighted or weighted pair group method using average.

The NJtree method (for neighbour-joining) proposed by Saitou and Nei (1987) is often used in genetic diversity. It uses the criterion of relative neighbourhood, the criterion of updating the dissimilarities is that of the unweighted average, and the dissimilarities are adjusted to an additive tree distance. The calculation complexity is of the order of n^3 and allows us to treat matrices of some hundreds of individuals. Several studies by simulation have shown the effectiveness of this method in finding the true tree. Gascuel (1997) proposes a version, UNJ (unweighted neighbour-joining), which uses a criterion of weighted average. The choice between these two methods depends, as has already been mentioned, on the sampling process.

Sattath and Tversky (1977) propose the Addtree algorithm (or method of scores), which uses the average as a criterion of aggregation and makes an adjustment to an additive tree distance. The notion of neighbourhood is that of topological neighbourhood. The complexity of n^4 is higher than that of NJtree. This method, which also has a proven effectiveness, can be preferred for its topological point of view.

Consensus Trees and Common Sub-trees

The same set of individuals is often described by several types of markers of diversity. To each type of marker corresponds a matrix of dissimilarity and an adjustment to a tree. Each type of marker contributes different information; the trees produced are thus different. However, one can hope to build from them a strong structure, a consensus tree, which will be common to different trees.

The problem of comparison of trees is also found in the resampling approaches (bootstrap or jackknife). The object of these approaches is to test the stability of tree representations obtained. The principle is to compare the trees established from several random samples in the set of markers observed (Felsenstein, 1985). If the markers are numerous and of equivalent informative content, molecular markers, for example, each of these samples is recorded to give a single image of diversity. If all the trees obtained are closely related, we conclude the stability of the structure revealed. On the other hand, if the different random samples of markers give very different trees, then the structure obtained must be regarded with the greatest caution. A measure of dispersion around a synthetic consensus tree quantifies the deviations between these trees.

Considering that the principal information on diversity involves much more of the tree structure than the edge length, the methods presented here construct synthetic trees in terms of structure, without the edge lengths being taken into account.

Consensus trees

Various methods of consensus have been proposed on rooted trees. A direct generalization to non-rooted trees does not pose a major problem. The strict consensus only retains the edges that are present in all the trees compared. Since this method may be considered too severe, the majority rule has been proposed, which retains the edges present in more than 50% of the trees. If one compares two trees only, the two types of consensus are equivalent.

A measure of the difference between two trees can be defined as the sum of differences of each to their consensus tree. The consensus tree is always simpler in organization than the initial trees. This organization is quantified by an index of complexity v; the dissimilarity between two trees H and H', of consensus H_c, is thus:

$$d(H,H') = v(H) + v(H') - 2v(H_c).$$

Various definitions of complexity can be considered. The most common way is to measure the complexity of a tree is by its edge number. For the binary trees that have only nodes of the third degree, the complexity of H and H' will be $2n - 3$; the complexity of H_c is the sum of the number of its external edges, n, and of the number of its internal edges, I_c, from which $d(H,H') = 2n - 6 - 2I_c$.

The distance, called the edge distance, thus varies according to the number of internal edges conserved in the consensus tree. If I_c is 0, the consensus is a star, and the trees compared have no common structure. It is the opposite if I_c is $n - 3$; the initial trees are identical. This distance is often standardized by $2n - 6$ to keep the variation between 0 and 1.

The value of a distance can only be interpreted by reference to the value that will be obtained on the trees drawn randomly, thus without any common structure other than a random one. The distribution of this index, for random trees, is known only asymptotically, that is, when the number of individuals tends towards infinity. We can, however, estimate these distributions for finite numbers by simulation and thus construct an approximate statistical test. Table 5 gives the threshold values for different values of n and different thresholds of probability. For example, for a tree of 40 individuals, there is only a 5% chance that trees of distance less than 0.973 will be constructed. This distance seems hardly discriminating and the random hypothesis is often refused even if the trees have only a very small number of internal edges in common.

The edge distance carries the same judgement on all the edges no matter what their position on the tree. The complexity of a tree depends only on its edge number independent of its internal organization. An intuitive idea will

Table 5. Edge distance between trees. Threshold values for different levels of probability. Empirical test established from simulation of 20,000 pairs of random trees with *n* leaves

	Probability (%)			
n	1	5	10	20
20	0.882	0.941	0.941	1
40	0.973	0.973	0.973	1
60	0.982	0.982	0.982	1
80	0.987	0.987	0.987	1
100	0.990	0.990	0.990	1

be to give a different weight to edges according to their aptitude to 'structure' the tree. For example, an external edge that involves a single element is less structuring, and an edge that separates the group into two sub-groups of number $n/2$ is more structuring. One associates at an edge a weight that is the product of numbers of each of the sub-groups defined by this edge. The complexity is the sum of these weights on all the edges. A distance known as bipartition distance is thus expressed, as previously, as a function of the complexity of trees and of their strict consensus. The maximum reached is not simply expressed but can be calculated from n; the distance is standardized by this maximum. As for the edge distance, the distribution under a random model is not known but can be approached by simulation. Table 6 gives the threshold values for different probabilities of an approximate statistical test. It is clearly more discriminating.

Table 6. Bipartition distance between trees. Threshold values for different levels of probability. Empirical test established from simulation of 20,000 pairs of random trees with *n* leaves

	Probability (%)			
n	1	5	10	20
20	0.727	0.753	0.766	0.782
40	0.580	0.600	0.611	0.626
60	0.492	0.511	0.522	0.537
80	0.435	0.453	0.464	0.477
100	0.394	0.412	0.421	0.434

Common sub-trees

It is not unusual, in certain applications, to obtain consensus trees that look like a star, thus without marked structure, while direct examination of the initial trees reveals evident structural analogies.

The consensus tree makes the hypothesis that all the individuals are correctly represented; the exchange of an individual from one group to another is a strong indication that there is no real separation of the groups. For several applications this hypothesis is not entirely acceptable. For example, it has been demonstrated that the dissimilarities estimated from molecular markers can be compromised by a certain error. If the population studied combines units that are genetically closely related, on the infraspecific scale, for example, the imprecision of the measure of dissimilarities may suffice to explain the erratic character of certain units.

Considering that just a few individuals may mask a common structure, it is useful to try to identify these 'fluctuating' individuals rather than allow them to have the same weight as others in the determination of the common structure. The research of structures common to several trees thus needs to be reformulated. The problem becomes that of determining the smallest subgroup that needs to be pruned in each of these trees to obtain identical trees or, inversely, the largest subgroup of individuals having the same structure in all the trees. These individuals form the common sub-tree. In Fig. 7, this common sub-tree and the consensus tree are compared on a simple example.

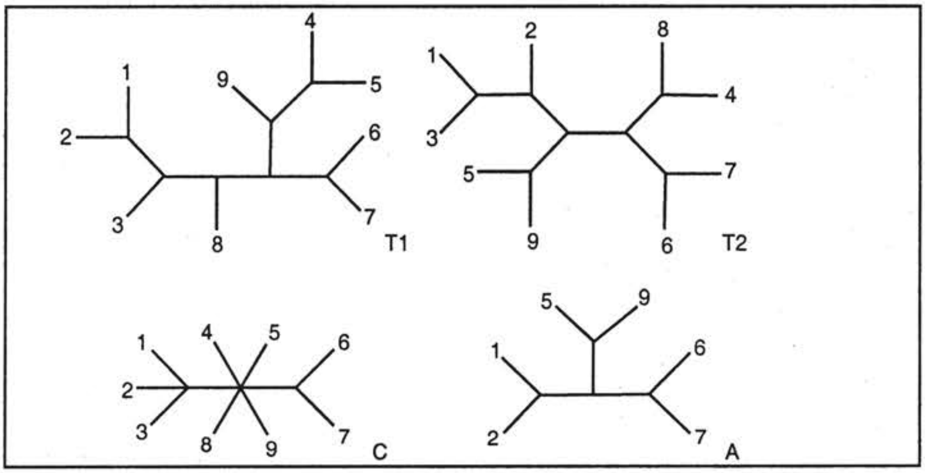

Fig. 7. Consensus tree (C) and common sub-tree (A) of trees T1 and T2.

This approach is much less widespread than the consensus approach and has been the subject of very few studies. The simple statement of the problem masks a relatively complex algorithmic problem. Kubicka et al. (1995) published the basis of an algorithm that gives an exact solution while retaining a sufficiently low complexity to be used in practice. The algorithm relies on the enumeration of all the solutions possible while reducing complexity by limiting the depth of exploration of the branches on a stop criterion.

The order o of common sub-tree, that is, the number of individuals conserved, can be considered as a measure of the resemblance between trees. The maximum order is n, and it is obtained for two identical trees. On the other hand, the minimal value of the order is 3, since it has only a single possible typology of three distinct points, and it is thus common to two trees. From this arises a definition of dissimilarity between trees $D = n - o$, which one can standardize as $n - 3$ to keep the variations between 0 and 1. The practical use of this criterion requires knowledge of the distribution of o under the hypothesis of independence of trees. This distribution can be approached by simulation and has been calculated for binary trees of 20 to 100 leaves (Table 7). It can be used to fix a rule of interpretation. By fixing a threshold of 5%, for example, it can be considered that chance gives less than 5 out of 100 trees of order superior or equal to 10, 13, 16, 18 and 19 when n varies from 20 to 100 by steps of 20.

Table 7. Distribution of number of leaves of common sub-trees of 1000 pairs of random binary trees with 20, 40, 60, 80 and 100 leaves

	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Mean
20	1	40	50	9											7.67
40				5	34	43	17	0	1						10.76
60						2	30	48	17	3					12.89
80								3	25	39	24	7	2		15.13
100									2	19	36	28	11	4	16.39

It is necessary to emphasize that the consensus tree and common sub-tree do not have the same point of view on data. Each is justified according to the interpretation that one can make on the shift of individuals in the initial trees. If it is a matter of accidental exchange of some individuals within a strong common structure, then the common sub-tree appears more adapted. If the exchange really expresses fundamental differences of structure, then the consensus tree is logically more adapted. In practice it is rarely possible to opt boldly for one of these two hypotheses and the two approaches can be implemented in a complementary manner.

CONCLUSION

The dynamism of the field of numerical taxonomy must be emphasized. It was considered, with the work of Sneath and Sokal (1973), that everything, or almost everything, had been said on the subject. Recent theoretical developments, the considerable increase in information technology, and the use of new types of markers have again thrown open this discipline, which is presently highly active at the interface between mathematics and biology.

On reading this chapter, the reader will understand the necessarily joint intervention of these two fields. The question that naturally arises from such a presentation—are some methods better than others?—does not have a real meaning in this field. There is no universal mathematical solution that is valid in all circumstances. Among the various approaches available we can choose the most relevant only on the basis of our knowledge of the plant species studied, its level of diversity, mode of reproduction, heterozygosity, and so on, the nature of the sample analysed, and the characteristics of the markers used.

REFERENCES

Barthelemy, J.P. and Guenoche, A. 1988. *Les arbres et les représentations des proximités*. Paris, Masson, 239 p.

Beaulieu, F.B. 1989. A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, 6: 233-246.

Darlu, P. and Tassy P. 1993. *Reconstruction phylogénétique*. Paris, Masson, 245 p.

Critchley, F. and Fichet, B., 1994. The partial order by inclusion of the principal classes of dissimilarity on a finite set and some of their basic properties. In: *Classification and Dissimilarity Analysis*. B. van Custem ed, New York, Springer, Lecture Notes in Statistics, pp. 5-65.

Escofier, B. and Pages, J. 1993. *Analyses Factorielles Simples et Multiples: Objectifs, Méthodes et Interprétation*. Paris, Dunod.

Escoufier, Y. 1975. Le positionnement multidimensionnel. *Revue de statistique appliquée*, 23(4): 5-14.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4): 783-791.

Gascuel, O. 1997. Concerning the NJ algorithm and its unweighted version, UNJ. In: *Mathematical Hierarchies and Biology*, DIMACS workshop. American Mathematical Society, Series in Discrete Mathematics and Theoretical Computer Science 37, pp. 149-170.

Jukes, T.H., and Cantor, C.R. 1969. Evolution of protein molecules. In: *Mammalian Protein Metabolism*, H.N. Munro, ed., New York, Academic Press, pp. 21-132.

Kubicka, E., Kubicki, G., and McMorris, F.R. 1995. An algorithm to find agreement subtrees. *Journal of Classification*, 12: 91-99.

Lavit, C., 1988. *Analyse Conjointe de Tableaux Quantitatifs*. Paris, Masson.

Lefort-Busson, M. and de Vienne, D. 1985. *Les distances génétiques, estimations et applications*. Paris, Inra, 181 p.

Perrier, X. 1998. Analyse de la diversité génétique: mesures de dissimilarité et représentation arborée. Doct. thesis, Université Montpellier II, Montpellier, France, 192 p.

Rohlf, F.J. 1987. *Ntsys-pc Numerical Taxonomy and Multivariate Analysis System*. New York, Applied Biostatistics Inc., Setauket.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4): 406-425.

Saporta, G. 1990. *Probabilités, Analyse de Données et Statistiques*. Paris, Technip, 493 p.

Sattath, S. and Tversky, A. 1977. Additive similarity trees. *Psychometrika*, 42(3): 319-345.

Sneath, P.H.A. and Sokal, R.R. 1973. *Numerical Taxonomy*. San Francisco, Freeman, 573 p.