TAXON 21(5/6): 567 – 606. NOVEMBER 1972

CHARACTERS AND CLUSTERING IN TAXONOMY: A SYNTHESIS OF TWO TAXIMETRIC PROCEDURES[1]

Pierre Legendre[2,3] *and David J. Rogers*[2]

Summary

The problem of producing a classification from data gathered on specimens has two main components: first the information about the specimens must be structured as characters and character states in such a way that it carries the most information about the taxonomic structure of the objects under study, the mathematical 'noise' being eliminated as much as possible. Then this information must be handled in such a way that a hierarchical partitioning of the objects, called classification, is derived.

This paper presents computer-aided methods for the accomplishment of these steps. These methods were worked out to be both mathematically and biologically sound. The character analysis method (called CHARANAL) uses information theory to measure the amount of information common to pairs of characters, and derives from it various measures for the comparison of characters. The clustering technique presented here (entitled GRAPH), on the other hand, is based upon graph theory, and is intended to represent the thought process of the 'classical' taxonomist. For each method are given a general explanation, a detailed explanation of the mathematics involved, an example, and a section on interpretation of results.

Résumé

Une classification produite à partir des données recueillies sur des spécimens résulte de deux manipulations successives de ces données: il faut d'abord structurer l'information que l'on possède sous la forme de différentes descriptions d'un certain nombre de caractères, et faire en sorte que l'information taxonomiquement significative soit préservée et que le brouillage mathématique soit éliminé le plus possible. Par la suite cette même information doit servir à produire une série hiérarchique de partitions des objets sous étude, ce qui s'appelle une classification.

Les auteurs présentent ici pour ce faire des méthodes d'analyse par ordinateur dérivées de principes biologiques et mathématiques reconnus. Pour l'analyse des caractères (programme CHARANAL), l'information commune à des paires de caractères est mesurée suivant les principes de la théorie de l'information, ce qui permet de dériver différentes mesures pour la comparaison des caractères. La technique de groupement des objects (programme GRAPH) dérive d'autre part de la théorie des graphes et essaie de reproduire aussi fidèlement que possible le cheminement de la pensée du taxonomiste dit 'classique'. À la suite de l'explication générale, on retrouvera une explication détaillée de l'aspect mathématique, un exemple ainsi qu'une section sur l'interprétation des résultats, ce pour chacune des deux techniques.

[1] Paper No. 27 from the Taximetrics Laboratory, Univ. of Colorado, Boulder, Colorado 80302, U.S.A.

[2] Taximetrics Laboratory, Department of Biology, University of Colorado, Boulder, Colorado 80302, U.S.A.

[3] Present address: Centre de recherches écologiques, Université du Québec à Montréal, C.P. 8888, Montréal 101, Québec.

Introduction

Over the past 12 to 15 years, a considerable amount of effort has been expended to understand the methods and thought processes by which taxonomists produce classifications. During this period, several schools and workers have developed various models of various parts of the classification process, but there has been little overall effort to integrate the methods into a flowing step-wise set of processes. The most prominent of these efforts has been in procedures to make the process of clustering more objective. Clustering, which is indeed a major part of the classification process, attemps to place specimens, or taxa, into hierarchical groups, indicating the various levels of relationships between the objects which form clusters. Once clustering techniques were established, it became painfully obvious that there was great need to consider the information, or characters, which were used in the clustering process, because, no matter how powerful the clustering model was, there were frequent cases when the clusters formed did not meet the necessary criteria required by taxonomists to reflect the biological relationships of the organisms comprising the clusters.

Various attemps have been made to improve the clustering process. One obvious attempt is to consider a very large number of characters simultaneously, rather than by making a priori judgements about which characters were "good" or "bad". Hopefully, by describing as many characters as possible, and using all of them simultaneously, one would overcome the serious problems caused by inept weighing of the characters. This led numbers of people to attempt scaling techniques, by which there would be no more, no less, value assigned to each character. It has also become painfully obvious that there are as many dangers in this process as there are in the a priori, intuitive process of selecting characters. For example, in using many characters, one may measure the same character in different ways, thus injudiciously introducing a biased classification towards one small component of the gene pool, and either ignoring, or underemphasizing other important components.

Such processes have frequently led to more and more models to test this or that component of the model, each step generating more and more numbers, and at the same time, separating the taxonomist farther and farther from his initial objective of classification. Since it is obvious that "good" classifications have been produced, some of them well over one hundred years ago, it should also be obvious that the most useful process to build models from would be to examine carefully what a number of taxonomists can agree upon as useful procedures, and to then build models which reflected the best thinking these taxonomists have done. This is a rather difficult process, but one which eventually produces the best results. The difficulty lies in the fact that the "good" taxonomists seldom state the methods by which they reach decisions, but place before us the results of their thinking, and leave to the less-good taxonomist, or non-taxonomist, the enormous job of deciding how decisions were made which produced a good classification.

Basically, it has become apparent that good taxonomists are outstanding pattern analysts. By some process, they have learned to distinguish between patterns which were reliable and predictive, and those which were confusing and unpredictable (in the jargon of the information specialist, "noisy signals"). Furthermore, it is apparent that the basic process used to select a character is a comparative one – comparing a known pattern with a new or undefined pattern. Whether the taxonomist compared only external morphology, or anatomical structure, chemical structure or cytological information, made no real difference. If the taxonomist gives us a character (or a classification) by comparing known against unknown, and reflects in the unknown the same clear differentiation as in the known, he has provided a pattern, reproducable and understandable by others.

It is further apparent that good taxonomists must have a broad-ranging knowledge of the groups of organisms under study (and other, related groups), in order to select patterns which will serve the purpose of producing a good classification. Thus it is not only an insult to well-trained biological taxonomists to attempt to have an "intelligent ignoramus" (Sokal and Rohlf, 1970) make a classification, it is also absurd.

There is still much confusion caused by the failure to separate clearly the methods of thought employed in making models from the multitudinous data confronting the taxonomist. For example, Michener (1970) has no clear definition of the structure of a character in the first part of his paper, but toward the end, he seems to have clarified his own thinking. Mayr (1970) describes the character in terms of its biological nature, but does not tell the basic requirement that a character must serve to make a useful classification. There is a further misconception that all classifications, to qualify as "good", *must* both reflect the modern-day relationships between taxa, and reflect the phylogeny of the group of organisms. The dualism inherent in this requirement frequently is the cause of poor classifications.

In this paper, we attempt to describe together two separate models, which have been previously described separately (Estabrook, 1966, 1967; Wirth *et al.*, 1966), in hopes that their juxtaposition herein provides better insight into the methods employed by taxonomists. The first model (the analysis of characters) reflects the processes of sorting and sifting of information as a taxonomist begins his process of classification, starting with many potential characters, and eventually choosing from the many potential ones those which will clearly aid in the process of classification, and rejecting, or modifying the poor characters which do not provide useful patterns. The second model (clustering) indicates the continuing process of synthesizing the individual characters chosen from the first model, in two basic steps. The first part of the clustering requires that the same measure of similarity between all the objects be provided, and the second part is that which is formally the clustering. In the second part, all the measures of similarity derived in the first part are used to place together in clusters, hierarchically and in a non-overlapping manner, all the objects for which similarities have been derived. Our philosophy with respect to the building of the models described below was that they must reflect the thinking of taxonomists, they must be mathematically sound, they must provide information to the taxonomist to aid in his decision-making, and they must be made into practical computer programs.

Section I: Character analysis

What are the requirements of a character for a classification? A character, at a minimum, must partition the objects under study into nonoverlapping groups. To accomplish this requirement, a character must be structured as a rule which associates with each organism (specimen) in a collection under study one member of a set of nonoverlapping descriptions called character states. A necessary property of the states of any character is that they are independant: in other words, for any given state of a character, we have to be able to say that any one organism falls in this state, or not. On the contrary, the characters do not have to be independant from each other. The information contained in one character may be a partial or complete redundancy of the information contained in another character. A complete redundancy of the information in two characters would mean that one of them will contribute as much as both to the classification. Independance of two characters indicates that they contribute differently, then complementarily, to the classification. In summary, a character is a single basis for comparison defined over all the objects under study. This definition allows the taxonomist to employ his full range of biological knowledge in the construction of a classification. For a more detailed discussion on the properties of the characters and the character states, see also Estabrook (1967) and Estabrook and Rogers (1966).

It is evident that the task of defining which characters have to be used in order to work out the classification of a group of organisms (hereafter referred to as *objects*), and which states have to be defined for each character, requires a good biological knowledge of the objects under study: one has to be careful with such phenomena as phenotypic plasticity and minor allelic differences.

The competent biologist knows from his own experience, and that of other specialists, the most useful characters for the classification of the objects under study. These fundamental characters are those which will have a higher weight in his classification. However, those using numerical methods in taxonomy usually work with a relatively large number of characters, the inter-dependency of which is not always clear.

Considering all the information (structured in terms of characters and character states) available on the objects under study, two types of structure can be recognized. The *intrinsic structure* of the information refers to that part of the information which has a taxonomic value because it reflects the affinities and differences between the objects. It is the portion of the information that one wants to use in order to establish the hierarchical grouping of the objects called a classification. The *extrinsic structure* of the information, on the other hand, is that added by the taxonomist by his action of defining the characters and the character states. There is a certain amount of unescapable "noise" introduced into the classification process that can distort the resulting classification, the amount of which has to be reduced as much as possible. The way to reduce this "noise" is to conduct a study of the structure of the information. At the end of such a study, referred to hereafter as character analysis, a redefinition of the states and an evaluation of the characters will be possible. To summarize the questions of interest with regard to the information content of characters (Rogers and Appan, 1969, pp. 615-616):

1) What is the taxonomically significant information content of each character?

2) What is the amount of correlation between characters, as present in the various objects?

3) Is there any redundancy as a consequence of the description in different ways of the same genetic cause?

4) Should the character states be redefined?

SYMBOLS USED IN CHARACTER ANALYSIS

H: amount of entropy, or confusion, equal to $-\sum_{i=1}^{n} p_i \log_2 p_i$

p_i: the probability of an object being in state *i* of the character
I, J, K: names of three characters
$J(a) = J_3$ means that the character J assigns to object *a* the third state of character J
S: set of 'good objects' under study, or objects for which the information about the two characters compared is known
J_h^{-1}: the subset of S that includes all the objects to which state *h* of character J has been attributed
$p(J_h)$: the probability of state *h* of character J
C[A]: the number of objects in the set A
$C[J_h^{-1}]$: the number of objects to which state *h* has been attributed
C[S]: the total number of 'good objects'
H(J): amount of entropy in character J
$p(J_3/I_1)$: probability of finding an object to which state 1 of character I *and* state 3 of character J have been attributed, among those objects coded in state 1 of character I
$H(J/I_1)$: conditional entropy remaining in J for the objects to which state I_1 has been attributed
H(J/I): the total conditional entropy remaining in character J after observing all the states of character I
D(I,J): measure of independence (distance) of the characters I and J
$p[I_g \cdot J_h]$: the probability of choosing an object with state *g* of character I *and* state *h* of character J *in* the set of 'good objects'
$H[I \cdot J]$: total amount of information possessed by both I and J
S(I,J): similarity of characters I and J equal to 1 - D (I,J)

5) Which characters are of diagnostic value?

6) Which characters can be eliminated because they are of marginal interest?

Various techniques have been used in the past to measure the amount of information held in common by two characters, the most common of which are statistical correlations. Estabrook (1967, p. 86) mentions some other techniques that have been used in connection with this problem.

Correlation studies made with the usual parametric statistics require that the characters be ordered (an ordered character is one in which the states can be associated with the real numbers, or placed in succession on an axis). However, many of the characters used in biology are of the non-ordered type: there is no taxonomic significance, for instance, in ordering the various colours of the human hairs on a wave length axis. Consequently, no regression is possible with such a character.

The computer-aided method described below, called CHARANAL, measures how the information is distributed between the states of each character, and also in the comparison of pairs of characters. It is applicable to ordered and non-ordered characters, as well as to combinations of both types. It cannot be associated with parametric statistics, because it does not compare measures abstracted from the data distributions, such as means or standard deviations. Instead, it studies the actual distributions of the data. In this sense, it is related to non-parametric statistics, except that it uses information theory instead of the theory of probabilities. This method will be discussed after a short presentation of the concept of entropy in information theory.

Entropy

The concept of entropy was first developed as a quantitative formulation of the second law of thermodynamics. Entropy is a measure of the disorganization of a closed system: so, the higher the entropy is, the lower is the amount of work that can be done by this system. Accordingly, the second law of thermodynamics becomes, in the formulation of Boltzmann (1896, p. 60): a closed system can vary only if its probability (entropy) is increased by the variation.

The transfer of the entropy concept to information theory was gradual. The first step was accomplished by Boltzmann (1896, pp. 41-42) who found the entropy to be proportional to the logarithm of the number of alternatives possible for a closed system, when all the known information has been recorded. In other words, the entropy is proportional to the logarithm of the amount of information that is missing.

This concept was developed by various authors; among them Shannon (1948, pp. 419-420), who derived from the premises of the theory of information, the equation

$$H = - \sum_{i=1}^{n} p_i \log p_i$$

where H is a measure of the uncertainty, or choice; p_i is the probability of the various events i, or the frequency attached to each piece of information i. He recognized that his equation was similar to the equation of entropy of Boltzmann (1898, pp. 219-221) and concluded that H corresponds to the entropy of an information system.

Probability distribution

The example introduced here will be carried all through the following discussion. Two correlated metric characters will be analyzed in twelve populations of trouts. Table I defines the character states and Table II describes the 60 objects (specimens) with the two characters. The first two digits of the object number refer to the population.

An example using non-ordered, qualitative biological descriptors can be found in Estabrook (1967).

A character (denoted by a capital letter, e.g., I or J) has been defined as a function which assigns to each object under study (denoted by a small letter, e.g., a) one and only one state (denoted by function symbols with a subscript) of this character. For example, $J(a) = J_3$ means that character J assigns to object *a* the third state, or description, of character J. The set of objects under study will be referred to as the set S. J_h^{-1} will designate the subset of S that includes all the objects to which state *h* of character J has been attributed.

A probability distribution can be associated with each character over the range of its states: the probability of each state will be equal to the relative

Table I. *Definition of the character states.*

Character I = 1	Head length / standard length
State 1	0.220 - 0.245
State 2	0.251 - 0.273
State 3	0.276 - 0.290
State 4	0.294 - 0.300
State 5	0.306 - 0.310
Character J = 2	Orbit length / standard length
State 1	0.049 - 0.052
State 2	0.062 - 0.063
State 3	0.067 - 0.088
State 4	0.091 - 0.096

Table II. *Description of the objects by characters 1 and 2.*

Object number	State for character I=1	State for character J=2	Object number	State for character I=1	State for character J=2
011	1	1	071	1	3
012	1	1	072	2	3
013	1	1	073	2	3
014	1	1	074	2	3
015	1	1	075	2	3
021	2	3	081	2	3
022	2	3	082	3	3
023	2	3	083	2	3
024	3	3	084	2	3
025	2	3	085	2	3
031	4	4	091	2	3
032	4	4	092	2	3
033	5	4	093	2	3
034	5	4	094	2	3
035	4	3	095	2	3
041	2	3	101	2	3
042	3	3	102	2	2
043	2	3	103	2	2
044	3	3	104	2	2
045	2	3	105	3	3
051	3	3	111	2	3
052	3	3	112	3	3
053	3	3	113	2	3
054	2	3	114	3	3
055	3	3	115	3	3
061	2	3	121	3	3
062	3	3	122	3	3
063	3	3	123	3	4
064	3	3	124	4	4
065	2	3	125	4	4

frequency of this state as found in the set S of objects under study. For example, referring to Table II, state 2 of character 1 has been attributed to 29 of the 60 specimens. The relative frequency of this state is 29/60 = 0.48333. This frequency can also be considered as the probability of randomly choosing, from the set of objects under study, an object to which the state 2 has been attributed. To generalize, the probability of each state *h* of character J can be noted as

$$p(J_h) = \frac{C[J_h^{-1}]}{C[S]}$$

where C[A] denotes the number of objects in (or cardinality of) the set A. $C[J_h^{-1}]$ is the number of objects to which state *h* has been attributed and C[S] is the total number of objects in the study (actually, if the state of an object is not known for a given character, the program is built in such a way that C[S] will be the cardinality of the set of "good objects", or objects for which the information is known). Table III shows the probability distribution associated with characters I = 1 and J = 2 of the example.

TABLE III. *Probability distribution associated with characters I=1 and J=2.*

$p(I_1) = .10000$	$p(J_1) = .08333$
$p(I_2) = .48333$	$p(J_2) = .05000$
$p(I_3) = .30000$	$p(J_3) = .75000$
$p(I_4) = .08333$	$p(J_4) = .11667$
$p(I_5) = .03333$	

Unconditional entropy

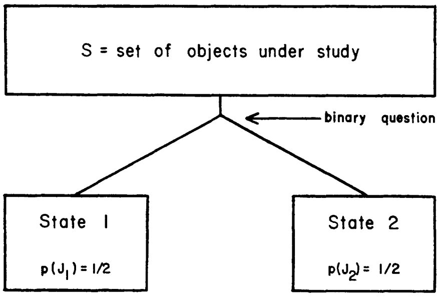
An information-theoretic measure of entropy can now be assigned to each of these probability distributions. It will be referred to as *unconditional entropy*, by opposition to the conditional entropy that will be introduced below. As mentioned above, the entropy measures essentially the difficulty of predicting what state of the character considered has been applied to an object chosen randomly. This measure of entropy is also equal in quantity to the information learned by actually observing the objects: entropy and information will then be considered as synonymous.

This confusing last sentence can be illustrated by a simple example. If half of the objects are found in each of the two states of a character, one can say that there is confusion, or entropy, in this distribution, because one cannot predict with certainty in which state a given object will fall, by observing *only* the probability distribution on the two states. By comparison to that, if one knows in which state each object has been classified, the confusion is removed, and we can say that the *amount* of information gained by observing the objects is equal to the *amount* of confusion, or entropy, that the probability distribution presented before. This is why, when talking about this *amount*, we can refer to it as an amount of entropy or of information necessary to remove the confusion.

Now, if one makes his observation by asking yes-no questions, the average minimum number of these questions *necessary* to find the proper assignment of each object can be seen as a measure of the confusion, or entropy of the system. The entropy will then depend on the number of states and on the distribution of the objects in the various states of the character. A few examples can clarify the process.

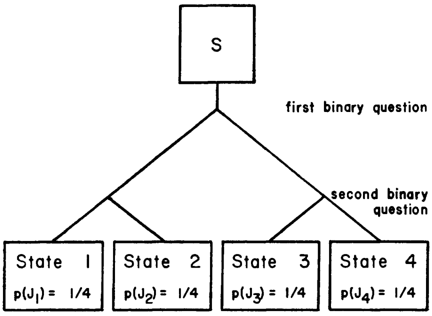
1) If all the objects are in the same state of the character, everything is already known about the distribution of the objects in the states of this character. The number of questions necessary is then 0.

2) With a two-state character, if one of the states has been attributed to half the objects and the other state to the other half, it will be necessary, for each object, to ask exactly one yes-no question of the type: "Has the state 1 been attributed to this object?", in order to know the probability distribution of the character:

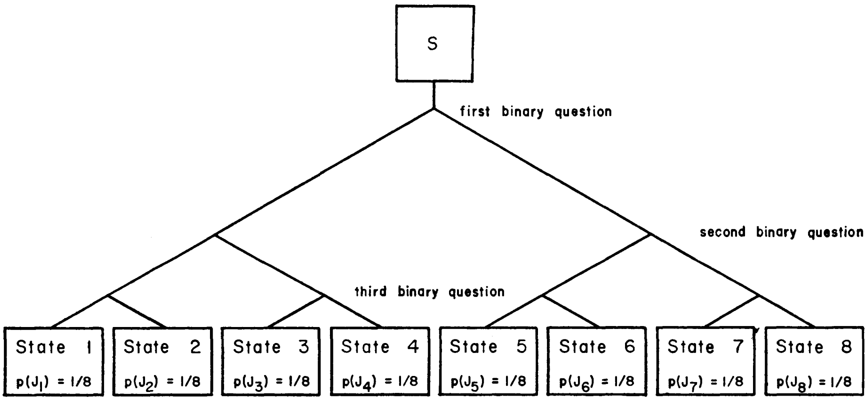


The entropy associated with this character is then 1.

3) The same process, applied to a character of four states between which the objects are distributed equally, would give an entropy of 2, since exactly 2 binary questions have to be asked for each object:



4) The same process is applicable in the case of an eight-state character with equal distribution of the objects between the states:



The total entropy in this character will then be [3 questions x 8(1/8 of the objects)] = 3.

To generalize, the entropy associated with a character for which the objects are evenly distributed between the states is the logarithm base 2 of the number of states: $\log_2 1 = 0$; $\log_2 2 = 1$; $\log_2 4 = 2$; $\log_2 8 = 3$; etc.

Note: The presentation of the entropy concept in information theory made above, using the actual minimum average number of yes-no questions that have to be asked in order to remove the confusion of the system, as a measure of entropy, gives only an approximation of the entropy. It has been used in order to give the reader a conceptual understanding of what entropy is. But this method is accurate only in the cases where the number of states in the character is one of the integer powers of 2, like 1, 2, 4, 8, 16, . . . , the objects being, again, equally distributed between the states. In the other cases, there is a slight departure between the yes-no and the logarithmic method of calculating the entropy, as shown in Table IV. This is due to the fact the yes-no questions have their optimal effect only when they can divide the objects under study into two equal groups, which cannot be the case when the number of states is not an integer power of 2, (the objects being again, equally distributed between the states) or when the number of objects in the various states is not such that the binary questions will divide them into equal groups, as in number 5 below. In the other cases, because the binary questions are a little less efficient, on the average, more questions than $\log_2$ (number of states) are necessary in order to remove the confusion. For a justification of the measure of the information-theoretic entropy by the logarithm of the number of states, see Shannon (1948) or any textbook of information theory.

As the entropy of a character depends on the number of states in the character, the total entropy of the character with four states above can be divided between the four states. To each state will be attributed $1/4 \log_2 4$,

TABLE IV. The minimum average number of yes-no questions necessary to remove the confusion of the system is equal to $\log_2$ (number of states) only when the number of states is equal to an integer power of 2 (values in boldface), the objects being equally distributed between the states. In the other cases, the former is larger.

Number of states	$\log_2$ (number of states)	Average minimum number of yes-no questions
1	**0.00000**	**0.00000**
2	**1.00000**	**1.00000**
3	1.58496	1.66666
4	**2.00000**	**2.00000**
5	2.32193	2.40000
6	2.58496	2.66666
7	2.80735	2.85714
8	**3.00000**	**3.00000**
9	3.16993	3.22222
10	3.32193	3.40000
11	3.45943	3.54545
12	3.58496	3.66666
13	3.70044	3.76154
14	3.80735	3.85714
15	3.90689	3.93333
16	**4.00000**	**4.00000**

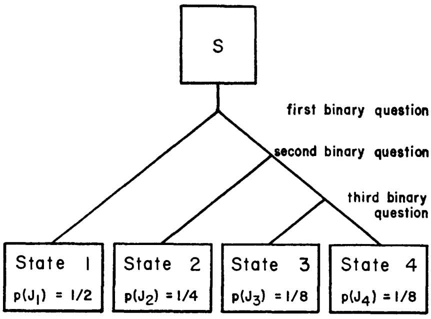
that is equal to $1/4 \log_2 (1/4)^{-1}$. The entropy of the character is then:

$$\sum_{\text{the 4 states}} 1/4 \log_2 (1/4)^{-1}$$

For the eight-state character above, to each state is attributed one eighth of the entropy of the character, or $1/8 \log_2 8$, that is equal to $1/8 \log_2 (1/8)^{-1}$. The entropy of this character is then:

$$\sum_{\text{the 8 states}} 1/8 \log_2 (1/8)^{-1}$$

5) Suppose now that the number of objects to which each state is attributed varies from one state to the other. An example would be a four-states character in which state 1 has been attributed to 1/2 of the objects, state 2 to 1/4, state 3 to 1/8 and state 4 to 1/8. The less evenly the probabilities of the different states are distributed, the more information one has, then the less should be the entropy: consequently the entropy of this character should be lower than 2 (see no. 3 above), that is the maximum amount of entropy that can be found in a four-states character. In asking the yes-no questions, it is economical to isolate half of the objects with the first question, then half of the remaining objects with the second question, and use a third question for the two last groups of 1/8 of the objects, following this pattern:



Half of the objects require 1 question, 1/4 require 2, and the two groups of 1/8 require 3 each. So, the total unconditional entropy of this character is

$(1/2 \times 1) + (1/4 \times 2) + (1/8 \times 3) + (1/8 \times 3) = 1.75$

$= 1/2 \log_2 2 + 1/4 \log_2 4 + 1/8 \log_2 8 + 1/8 \log_2 8$

$= 1/2 \log_2 (1/2)^{-1} + 1/4 \log_2 (1/4)^{-1} + 1/8 \log_2 (1/8)^{-1} + 1/8 \log_2 (1/8)^{-1}$

$$= \sum_{h=1}^{4} p(J_h) \cdot \log_2 [p(J_h)]^{-1}$$

As $\log_2 x^a = a \log_2 x$, the general form of the expression defining the entropy can be written

$$H(J) = - \sum_{h=1}^{j} p(J_h) \cdot \log_2 [p(J_h)]$$

where j is the number of states of character J.

The unconditional entropy of character 1 of the example is then, using the data of Table III:
$-[.10000 \log_2 (.10000) + .48333 \log_2 (.48333) + .30000 \log_2 (.30000) + .08333 \log_2 (.08333) + .03333 \log_2 (.03333)] = 1.82257$
The maximum amount of entropy that can be found in a five-states character is $\log_2 5 = 2.32193$

Similarly, the unconditional entropy of character 2 of the example is
$-[.08333 \log_2 (.08333) + .05000 \log_2 (.05000) + .75000 \log_2 (.75000) + .11667 \log_2 (.11667)] = 1.18773$
The maximum amount of entropy that can be found in a four-states character is $\log_2 4 = 2$.

Conditional entropy

As pointed out above, the entropy of a character is equal to the information that can be gained by observing the character over the objects. The entropy of character J is the maximum amount of information that can be obtained about J. Now suppose that instead of J, I is observed in order to learn about J; the relevant information will be obtained only insofar as I and J contain information in common. Knowing the probability distribution of I, it is possible to establish the *conditional probability distribution* of J, and then calculate the *conditional entropy* of J that it determines. If I and J share no common information, the conditional entropy of J is equal to its unconditional entropy; but in the other event, the conditional entropy of J is lowered by an amount equal to the information shared in common by the two characters.

The first step is to build a matrix in which the frequency of the objects, in each of the states of character J, will be established for the various states of character I. This is done in Table V for the example.

The first line of the Table V, for instance, represents all the objects that are in I_1 and gives their frequency distribution on the character J. State J_1 has been attributed to 83.333% of these objects, and state J_3 to 16.667%. This is why this row, as well as the others, sums to 1. In other words, the probability of finding an object to which state 1 of character I *and* state 3 of character J, noted $p(J_3/I_1)$, is the number of objects to which J_3 and I_1 have been attributed (that is the number of objects in the intersection, noted Ω, of the subsets I_1^{-1} and J_3^{-1}), divided by the number of objects to which state I_1 has been attributed (that is the number of objects in the subset I_1^{-1}). It would then be noted

$$p(J_3/I_1) = \frac{C\,[J_3^{-1}\,\Omega\,I_1^{-1}]}{C\,[I_1^{-1}]}$$

The number of objects in the subset I_1^{-1} (Table II) is 6, and the number of objects in the intersection of J_3^{-1} and I_1^{-1} is 1, since states I_1 *and* J_3 have been attributed only to object 071. Then

$$p\,(J_3/I_1) = 1/6 = .16666'$$

The general formula will be written

$$p\,(J_h/I_g) = \frac{C\,[J_h^{-1}\,\Omega\,I_g^{-1}]}{C\,[I_g^{-1}]}$$

TABLE V. Conditional probability distributions: probability of choosing an object with the various states of character J=2 assuming that the object has the given state of character I=1.

States of character I=1	$p(J_1) =$ .08333	$p(J_2) =$ .05000	$p(J_3) =$ .75000	$p(J_4) =$ .11667
I_1	.83333	0.00000	.16667	0.00000
I_2	0.00000	.10345	.89655	0.00000
I_3	0.00000	0.00000	.94444	.05556
I_4	0.00000	0.00000	.20000	.80000
I_5	0.00000	0.00000	0.00000	1.00000

The conditional entropies remaining in character J after observing each of the states of character I can now be established by using the formula of entropy developed above. The conditional entropy remaining in J for the objects to which state I_1 has been attributed is:

$H(J/I_1) = -[.83333 \log_2 (.83333) + .16667 \log_2 (.16667)] = .65002$

For the objects to which state I_2 has been attributed:

$H(J/I_2) = -[.10345 \log_2 (.10345) + .89655 \log_2 (.89655)] = .47983$

And similarly

$H(J/I_3) = -[.94444 \log_2 (.94444) + .05556 \log_2 (.05556)] = .30954$

$H(J/I_4) = -[.20000 \log_2 (.20000) + .80000 \log_2 (.80000)] = .72193$

$H(J/I_5) = -[1.00000 \log_2 (1.00000)] = 0.00000$

In the general form, the entropy remaining in character J after observing each of the states g of character I is determined by the formula

$$H(J/I_g) = -\sum_{h=1}^{j} p(J_h/I_g) \cdot \log_2 [p(J_h/I_g)]$$

where j is the number of states in J.

The conditional entropy $H(J/I_1)$ applies to the proportion of objects to which state 1 of character I has been attributed, that is .10000 of the objects, according to Table III. Similarly, the conditional entropy $H(J/I_2)$ applies to .48333 of the objects. And so on. The total conditional entropy remaining in character J after observing all the states of character I is then the weighed sum, over the frequencies of the various states of I, of the conditional entropies $H(J/I_g)$ found above.

The general formula can be written:

$$H(J/I) = \sum_{g=1}^{i} H(J/I_g) \cdot p(I_g)$$

where i is the number of states in I.

From the results above and the data of Table III, the calculations of the conditional entropy remaining in J after observing I will be performed in Table VI.

Table VI. Conditional entropy remaining in character J=2 after observing character I=1.

$H(J/I_g)$	$p(I_g)$	$H(J/I_g) \cdot p(I_g)$
$H(J/I_1) = .65002$	$p(I_1) = .10000$	.06500
$H(J/I_2) = .47983$	$p(I_2) = .48333$	.23192
$H(J/I_3) = .30954$	$p(I_3) = .30000$	.09286
$H(J/I_4) = .72193$	$p(I_4) = .08333$	.06016
$H(J/I_5) = 0.00000$	$p(I_5) = .03333$	0.00000

$$H(J/I) = \sum_{g=1}^{i} H(J/I_g) \cdot p(I_g) = \quad .44994$$

Similarly, the conditional entropy H(I/J) remaining in character I=1 after observing character J=2 of the example, is 1.08478.

Interdependence of characters

Of the 1.82257 units of information (unconditional entropy) of character I=1, 1.08478 belong exclusively to character I. The remaining .73779 units are shared with character J=2. Similarly, of the 1.18773 units of information of character J, .44994 belong exclusively to it and .73779 units are shared with I. The fraction of information in character I=1 also contained in character J=2 can be found by dividing .73779 by 1.82257, which gives .40481. The same calculation, applied to character J, gives .62117 of the information of J that is shared with I.

A measure of independence can now be constructed from the various entropy measures, as follows:

$$D(I,J) = \frac{\text{info. held exclusively by I + info. held exclusively by J}}{\text{total information possessed by both I and J}}$$

The amount of information held exclusively by I and J are the values H(J/I) and H(I/J) calculated above.

To obtain the value of the total information possessed by both I and J, the procedure will be similar as before. The frequency of each of the possible combinations of states of the two characters is established first (in Table VII for the example). Each of these frequencies is obtained by dividing the number of objects in each of the pairs of states by the total number of objects in the study. Formally, the equation of each frequency, or probability, is

$$p\,[I_g \cdot J_h] = \frac{C\,[I_g^{-1}\,\Omega\,J_h^{-1}]}{C\,[S]}$$

where C[A] is the number of objects in the set A, as before. The symbol $p(I_g/J_h)$ was used before in the sense of "the probability of choosing an object with state *g* of character I *in* the set of objects that have state *h* of character J". The symbol $p\,[I_g \cdot J_h]$ is used here to mean "the probability of choosing an object with state *g* of character I *and* state *h* of character J *in* the set of objects under study".

TABLE VII. Actual number of objects in each of the possible combinations of states of the two characters I=1 and J=2 (above), and the corresponding frequency (below), or probability, when each number is divided by the total number of objects in the study, which is 60 in this example.

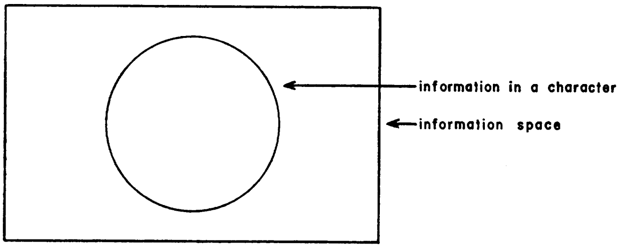
	J_1	J_2	J_3	J_4
I_1	5	0	1	0
I_2	0	3	26	0
I_3	0	0	17	1
I_4	0	0	1	4
I_5	0	0	0	2
I_1	.08333	0.00000	.01667	0.00000
I_2	0.00000	.05000	.43333	0.00000
I_3	0.00000	0.00000	.28333	.01667
I_4	0.00000	0.00000	.01667	.06667
I_5	0.00000	0.00000	0.00000	.03333

Then the total amount of information possessed by both I and J is calculated in the usual way, by applying the formula

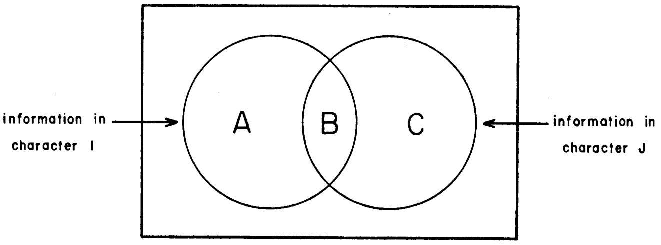
$$H[I \cdot J] = - \sum_{\substack{g=1 \\ h=1}}^{\substack{j \\ i}} p[I_g \cdot J_h] \cdot \log_2 (p[I_g \cdot J_h])$$

The value of H[I · J] in the example is 2.27251.

There is another, easier way to calculate the total amount of information possessed jointly by I and J. The intuitive concept of *information space* has to be introduced in order to visualize it. The information space corresponds to all the information possessed by the set S of objects under study and can be represented as a rectangle. The information possessed by each character corresponds to a subset of this space, represented by a circle in this rectangle. Such a representation is known as a Venne diagram:



Characters I and J have some information in common, as shown above. They can be represented as follows:



The amount of information in common is labelled B, and the letters A and C are attached to the part of the information held exclusively by I and J, respectively. The total amount of information possessed jointly by I and J is then, obviously, A+B+C. In the example, A = 1.08478, C = .44994 and B, the amount of information shared by I and J, is .73779, as shown at the beginning of this section. The total A+B+C is then equal to 2.27251, that is the value H[I · J] found above.

The independence of the characters I and J can now be calculated by using the expression above, the formal expression of which is

$$D(I,J) = \frac{H(J/I) + H(I/J)}{H[I \cdot J]}$$

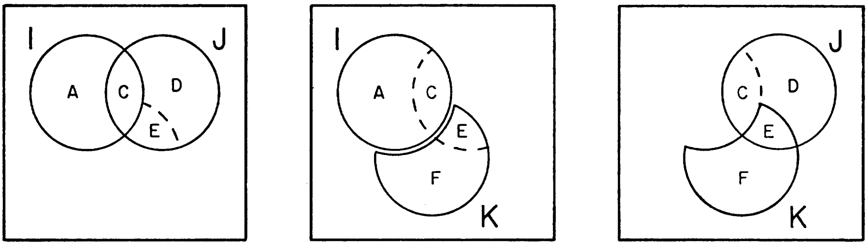
Its value for the example is

$$D(I,J) = \frac{.44994 + 1.08478}{2.27251} = .67534$$

It is easy to see that when the value of D is 0, the two characters contain exactly the same information. When it is 1, the characters are completely independent. D(I,J) can be understood as a function of the two variables I and J that associates with these variables a value, called the distance between I and J, that has the following properties:

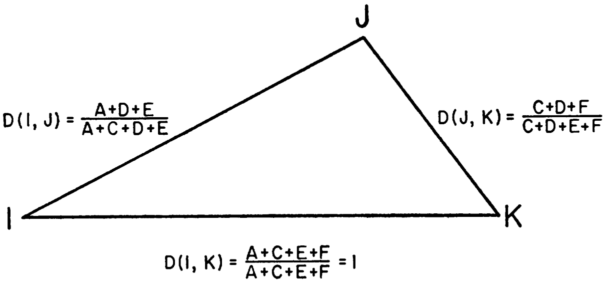
$D(I,J) \geq 0$ (= 0 if and only if I = J)
$D(I,J) = D(J,I)$
$D(I,J) + D(J,K) \geq D(I,K)$
Because of these properties, function D is called a metric.

The two first properties have been explained above. The third one can be illustrated by the following example, which considers characters I and J as above, and also a character K that has no information in common with I, but some with J:



$$D(I,J) = \frac{A+D+E}{A+C+D+E} \qquad D(I,K) = \frac{A+C+E+F}{A+C+E+F} = 1 \qquad D(J,K) = \frac{C+D+F}{C+D+E+F}$$

It is easy to show that D(I,J) + D(J,K) is larger than or equal to 1, then larger than or equal to D(I,K). It will be equal to 1 e.g. in the trivial case where C, D and F contain no information, A and E being non-void, that is the case where I and J have already no information in common and K is the same as J. The resulting picture can be represented by a triangle, the sides of which are equal to the measure of the distance between the characters:


J
$D(I, J) = \frac{A+D+E}{A+C+D+E}$
$D(J, K) = \frac{C+D+F}{C+D+E+F}$
I
K
$D(I, K) = \frac{A+C+E+F}{A+C+E+F} = 1$

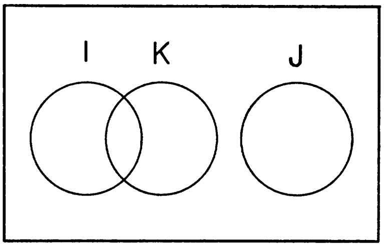
This distance value can be easily converted into a measure of the similarity of the two characters, by defining it as the complement of the distance, or

$$S\,(I,J) = 1 - D$$

In other words, the similarity between I and J is equal to the amount of information in common divided by the total amount of information held in the two characters, since

$$S\,(I,J) = 1 - D = 1 - \frac{A+B}{A+B+C} = \frac{C}{A+B+C}$$

This similarity value is however not a metric, since there are cases where it does not satisfy the third property of a metric. For example, in



it is obvious that S(I,J) + S(J,K) is not larger than or equal to S(I,K), since S(I,J) and S(J,K) are both null. However, the value of similarity will always be included between 0 and 1. In the example, the value of S is

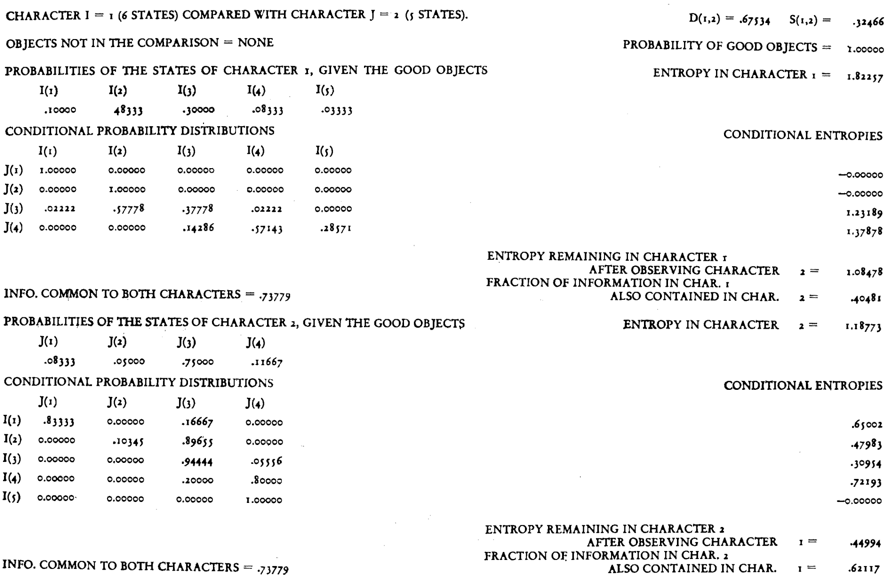
$$S(I,J) = 1 - .67534 = .32466$$

Actual form of the printout

After presenting a listing of the state of each character that has been attributed to each object in the study, as in Table II, the printout presents a complete comparison of all the pairs of characters, performing for each pair all the calculations explained above. As example, the printout of the comparison of characters I=1 and J=2 of the example is presented in figure 1.

The printout is divided in two main parts: study of the information left in character I after observing character J, and study of J after observing I.

Before these comparisons, the title of the page shows what pair of characters is compared and how many states are in each. The next line shows for how many – if any – objects the information is missing regarding one or the other of the characters in the pair under study. As explained above, these objects are eliminated from the comparison. Those remaining are called



```
CHARACTER I = 1 (6 STATES) COMPARED WITH CHARACTER J = 2 (5 STATES).          D(1,2) = .67534   S(1,2) =    .32466

OBJECTS NOT IN THE COMPARISON = NONE                                      PROBABILITY OF GOOD OBJECTS =   1.00000

PROBABILITIES OF THE STATES OF CHARACTER 1, GIVEN THE GOOD OBJECTS             ENTROPY IN CHARACTER 1 =   1.82257
        I(1)       I(2)       I(3)       I(4)       I(5)
        .10000     48333      .30000     .08333     .03333
CONDITIONAL PROBABILITY DISTRIBUTIONS                                               CONDITIONAL ENTROPIES
        I(1)       I(2)       I(3)       I(4)       I(5)
J(1)   1.00000    0.00000    0.00000    0.00000    0.00000                                      -0.00000
J(2)   0.00000    1.00000    0.00000    0.00000    0.00000                                      -0.00000
J(3)    .02222     .57778     .37778     .02222    0.00000                                       1.23189
J(4)   0.00000    0.00000     .14286     .57143     .28571                                       1.37878

                                                     ENTROPY REMAINING IN CHARACTER 1
                                                              AFTER OBSERVING CHARACTER    2 =   1.08478
                                                     FRACTION OF INFORMATION IN CHAR. 1
INFO. COMMON TO BOTH CHARACTERS = .73779                          ALSO CONTAINED IN CHAR.    2 =    .40481

PROBABILITIES OF THE STATES OF CHARACTER 2, GIVEN THE GOOD OBJECTS        ENTROPY IN CHARACTER    2 =   1.18773
        J(1)       J(2)       J(3)       J(4)
        .08333     .05000     .75000     .11667
CONDITIONAL PROBABILITY DISTRIBUTIONS                                               CONDITIONAL ENTROPIES
        J(1)       J(2)       J(3)       J(4)
I(1)    .83333    0.00000     .16667    0.00000                                                   .65002
I(2)   0.00000     .10345     .89655    0.00000                                                   .47983
I(3)   0.00000    0.00000     .94444     .05556                                                   .30954
I(4)   0.00000    0.00000     .20000     .80000                                                   .72193
I(5)   0.00000    0.00000    0.00000    1.00000                                                 -0.00000

                                                     ENTROPY REMAINING IN CHARACTER 2
                                                              AFTER OBSERVING CHARACTER    1 =    .44994
                                                     FRACTION OF INFORMATION IN CHAR. 2
INFO. COMMON TO BOTH CHARACTERS = .73779                          ALSO CONTAINED IN CHAR.    1 =    .62117
```

Fig. 1: Facsimile printout of the comparison of characters I=1 and J=2 of the example, made by CHARANAL.

the "good objects" and their frequency $= \frac{\text{number of good objects}}{\text{total number of objects}}$ is written down after the sentence "probability of good objects =". In this case, no information was missing about characters 1 and 2, and the probability of good objects is 1.00000. Above this line, the distance D(1,2) and similarity S(1,2) values are given for the pair of characters under study, I=1 and J=2.

The second part of the printout is the one for which the most extensive calculations have been done before. The explanations given for this part apply, of course, also to the first one. It starts with the sentence "probabilities of the states of character 2, given the good objects". The values given under it are the values $p(J_h)$ that have been shown in the right hand part of Table III. The value of "entropy in character 2", that occupies the right part of this same line, is the one that has been calculated at the end of the section on unconditional entropy.

The matrix called "conditional probability distributions" has been shown in Table V. The "conditional entropies" represent the amount of information contained in the distribution, on character J=2, of the objects to which each of the states of character I=1 has been attributed. These values were presented in the first column of Table VI. The "entropy remaining in character 2 after observing character 1" is the value of H(J/I) calculated at the end of Table VI.

The "information common to both characters" is the total entropy in character 2 = 1.18773 minus the value of H(J/I) = .44994, that is .73779. The "fraction of information in character 2 also contained in character 1" is the ratio of the information common to both characters divided by the total entropy in character 2. These two values were calculated at the beginning of the section on the interdependence of characters.

Exercise

Taking a pair of 3- to 6-states characters observed on about 50 objects (the data can also be invented for the purpose of the exercise), perform the various calculations explained above and write the results according to the format of the printout shown in figure 1.

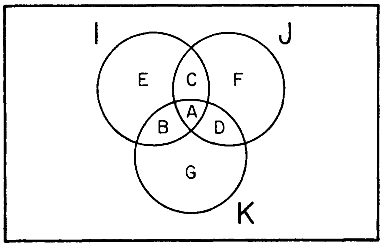
Interpretation

A biologist is looking for correlations that will give the structure of the resulting classification. What it means is that if many characters define the same partition of the objects into subgroups, these characters can be said to give rise to the same classification, then they are coherent with each other. But some attention has to be given also to the relative weight of the various characters.

Much emphasis has been given in the past to the question of weight. Some biologists argue that, in order to obtain a "natural" classification, one should not create artificially a difference in the importance attributed to the various characters. However, the competent biologist knows very well that all the characters do not have the same classificatory value. Characters built on single allele differences do not have the same importance as the characters related to structural or numeric differences between chromosomes, that characterize the species category since they are the basis for intrinsic reproductive isolation.

The task of attributing the weight of the various characters is the most difficult part of the work of the taxonomist. After the structure of the information has been established, the resulting classification can be obtained by a process largely automatic, as will be seen in the next section. There are two ways by which the proper weighing can be accomplished.

The first one will please the tenants of the non-weighing philosophy. Considering three characters I, J, K that are partially correlated as follows:



the area A represents information common to the three characters. If the three are used in making the classification, A will be weighed three times. Similarly, B, C and D will be weighed twice. This differential weighing reflects the biological structure of the study and results in a "natural" classification. But accordingly, no classification using correlated characters can be said to be of the non-weighed type.

The weighing of the characters can be modified also by changing the number of states. Each character makes its contribution to the resulting classification by dividing the group of organisms into subgroups. The more states per character, the smaller and more numerous are the subgroups formed by the character. Consequently, a character with few states will contribute to a major partition of the organisms, and will have more weight in the resulting classification. This can be understood in yet another way: the amount of entropy in a character is proportional to the number of states of the character. A character with two states can have no more than $\log_2(2) = 1$ unit of entropy, but an eight-states character can have up to $\log_2(8) = 3$ units of entropy. This, however, does not change the intrinsic amount of taxonomically significant information carried by the character. So, by increasing the number of states of a given character, without biological reasons, one diminishes the value of the fraction (intrinsic information)/(total entropy) and adds more mathematical "noise" to the contribution of this character to the classification. The extreme situation would be a character for which one state is attributed to each object; its total entropy would be equal to $\log_2$ (number of objects) but its contribution to the resulting classification would be almost nil.

For the worker, the criteria for the number of states to attribute to each character are: 1) how much importance he attaches to the diversity expressed by the character, and 2) how well the character classifies the objects under study. Also, he will tend to restrain the number of states of the most important characters (those applicable for the species, genus or other higher categories).

The first things to look at, on the printout, are the value of distance D(I,J) between the two characters, and also the values of the total entropy in each character and of the amount of information in common. The value of similarity S(I,J) can be used instead of the distance, since it carries the same information. Only the scale in which this information is presented has been changed by the passage from D to S. The figures of entropy remaining in each character after observing the other, and the fraction of information in common, carry the same type of information to the worker. Generally speaking, the following classification can be used (Hawksworth *et al.*, 1968):

the characters are very highly correlated if $D \leqslant .5$ and $S > .5$
highly correlated if $.5 < D \leqslant .7$ or $.5 \geqslant S > .3$
correlated if $.7 < D < < 1.0$ or $.3 \geqslant S > > 0$
unrelated if D is almost 1.0 or S is almost 0

No rigidity is attached to this nomenclature, however.

If the characters are found to be correlated, the matrices can be looked at, in oder to see the correlation more clearly. Table VIII presents the same data as Table V, but the values of entropy are written in boldface, in the conditional probability distribution matrix, when they are equal to or higher than the corresponding value in the unconditional probability distribution of character J, above.

In order to make easier the interpretation of the frequency values given in the matrices, it is possible to obtain a printout page showing the number of objects coded into the various states of each character. This can be done by entering the proper code on one of the parameter cards preceding the data deck. In the example for instance, it can help to know that the distribution of the objects on the various states of character I is 6, 29, 18, 5 and 2, and that there are 5, 3, 45 and 7 objects in the various states of character J, respectively.

The boldface figures of Table VIII have the following meaning: for the one in the upper left hand corner, for instance, it means that to observe state 1 of character I tells us 10 times more, about the probability of an object to

TABLE VIII. Data of table V, presenting the conditional probability distributions of character J=2, modified (see explanations in the text).

States of character I=1	Unconditional probability distribution of J			
	$p(J_1)$ = .08333	$p(J_2)$ = .05000	$p(J_3)$ = .75000	$p(J_4)$ = .11667
I_1	.83333	0.00000	.16667	0.00000
I_2	0.00000	.10345	.89655	0.00000
I_3	0.00000	0.00000	.94444	.05556
I_4	0.00000	0.00000	.20000	.80000
I_5	0.00000	0.00000	0.00000	1.00000

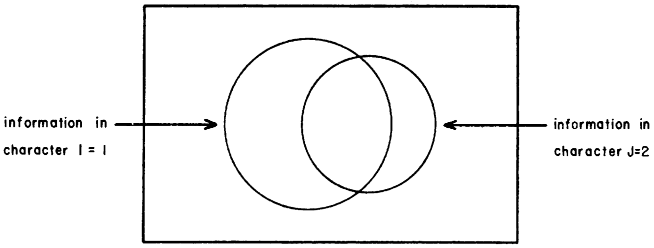
be in state 1 of character J, than we knew only from the knowledge of the unconditional probability distribution of character J. The correlation is clear in this example: the characters are highly correlated, as was already known by the observation of the value of D(1,2), that is .67534 in this case. Referring to what these characters are (Table I), it means that as the value of the ratio head length/standard length increases, also the ratio orbit length/standard length increases; that is, a relation depending only upon growth, and not something that gives any indication on the taxonomic structure of the objects under study.

When a correlation is found, two possibilities exist: either it is a valuable correlation between two characters, corresponding to a genetic correlation, in which case both characters should be kept, since good taxa are based on correlated characters; or the correlation is due to an undesired redundancy, as a consequence of the same structure being unintentionally described twice. A variation of this unintentional redundancy can be found in cases where the correlation depends on the determination of the correlated pattern by external factors such as ecology or growth, as it is the case in the example. This illustrates the contribution that has to be made by the biological knowledge of the material under study, in sorting out the unsignificant variations.

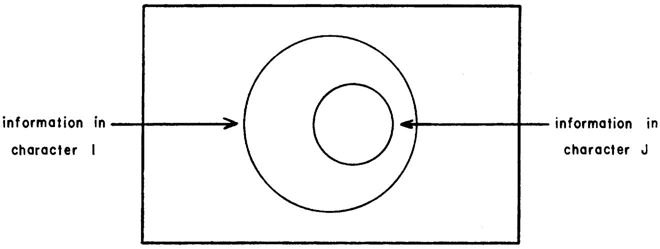
Another variation of the same phenomenon is found in the case where a character is a refinement of the other. If, for instance, the first character partitions the objects in two groups, and the second character divides each group into a number of subgroups, the working taxonomist will have to question the significance of each of these partitions, and eliminate the less significant one. Of course, if he considers the information described by these two characters as very important for his classification, he can keep the two, as he could use the most significant one twice, or three times, in making the classification.

If only one or a few objects do not participate in the general correlation between two characters, it usually means that an error has been introduced during one of the steps of the work of transmission of the information, from the actual observation of the specimen to the transcription on computer cards. If it is not the case, the character states may be redefined in order to improve the correlation, although this is not necessary in a multi-character study.

Different types of correlation can be involved. The value of the distance D(I,J) does not say everything about the type of correlation, since it does not refer to the relative amount of information present in each of the characters compared. The following picture is the illustration of the correlation in the example:



Because of the value of the distance, these characters were classified in the category "highly correlated". But considering the relative surface of the circles, proportional to the amount of information in each of the characters, it is obvious that this correlation would be better classified in the category "very highly correlated", and if one of the characters has to be eliminated from the study, it will be character J=2, since character I=1 contains more information. The extreme case can be represented as



where D(I,J) has a fairly high value, but the correlation is better represented by the fraction of information in character J also contained in character I, that is 1.00000.

If a character shows very little correlation with most of the other characters, and if it is considered as an important character, it will simply contribute to the resulting classification in a different way. It probably means also that not enough characters have been considered in this study. But if such a character represents a property of marginal interest, it is better to eliminate it.

The printout indicates also the necessity of redefining some character states. By looking at the column "conditional entropies" on the printout, it is easy to spot the values that are higher than the total amount of entropy in the character (there is no such value in the printout of figure 1). It means that, say, the observation of state 3 of character 5 gives less information about character 8 than what was already known from the frequency distribution in the various states of this character. If the same situation is found also, say, in characters 3, 6 and 7, where the observation of state 3 of character 5 gives less information than what was already known from the probability distribution of these characters, it would be recommended to divide this state 3 into more states; a revision of all the states can also be necessary, especially if this situation is found in many of the states. On the contrary, if the worker wishes to have fewer states in a character, he can group the states that create the less conditional entropy in most of the other characters, except if these states have a special biological significance.

A classification can be considered itself as a character, since it consists in a partition of the objects under study. The states of this character are of the type: "is a member of taxon X". So, after a classification is reached, CHARANAL can be used in order to calculate the amount of information in each character that the classification has preserved. This gives a measure of the "goodness" of the division of the characters in character states, according to the given classification. If the taxonomist has various classifications in mind, or if he wishes to compare various published classifications, he can calculate, by this method, which of the classifications preserves best the information contained in the various characters.

Still another way to accomplish this – that is limited to ordered characters, however – is to perform a discriminant function analysis, that calculates the optimum weight that has to be attributed to each character in order to obtain the best separation of the taxa. This optimum weight can be compared with the total entropy calculated by CHARANAL for each character, and the characters can be redefined accordingly.

Finally, it is possible to use CHARANAL for the establishment of the classification itself. This can be accomplished by choosing the character that represents best all the other characters, and partitioning the objects according to the various states of this character. Next, each of the groups thus obtained is run again in a separate CHARANAL, from which the first character chosen is eliminated. In each of the groups then, the character that represents best the classificatory power of all the others is chosen again, and its states are used to partition the objects into sub-groups, and so on, until a satisfactory classification is reached.

This can be done easily by asking for a special printout page that gives two values, called SUMRAT and SAMRAT, for each character. This page can be called by entering the proper code on one of the parameter cards preceding the data deck.

The term SUMRAT has been coined from 'sum of ratios'. SUMRAT (I) is actually the sum of the fractions representing the amount of information that I has in common with each of the other characters, divided by the amount of information of the character with which I is compared in the given ratio. Formally,

$$\text{SUMRAT (I)} = \sum_{j=1}^{n} \frac{H(I) - H(I/J_j)}{H(J_j)}$$

where n is the number of characters other than I in the study. Or in other words,

$$\text{SUMRAT (I)} = \sum_{j=1}^{n} (\text{fraction of information of character } J_j \text{ also contained in character I})$$

The term SAMRAT designates a closely related type of sum of ratios. This is why the name coined is also very similar. The difference is that the denominator of the various ratios that are summated is always the same, that is the amount of information in I, instead of the amount of information in the various J_j. So,

$$\text{SAMRAT (I)} = \sum_{J=1}^{n} \frac{H(I) - H(I/J_j)}{H\,(I)}$$

or in other words, $\text{SAMRAT (I)} = \sum_{j=1}^{n}$ (fraction of information of character I also contained in character J_j).

SUMRAT (I) and SAMRAT (I) are both large when much of the information in I is shared with that of other characters. If there is a one character that represents the classifying power of the other characters much better than any other, then the values of SUMRAT and SAMRAT of this character should be the largest. However, the values taken by SUMRAT (I) and SAMRAT (I) are also influenced by the amount of information in I. If I contains a large amount of information, like in a character with many states, that will tend to make SUMRAT (I) larger, but SAMRAT (I) smaller. This factor must be considered when there is a conflict between the indications given by SUMRAT and SAMRAT about which character represents best all the others. The conveniency of the classification obtained with each of the two characters in conflict can also be taken into account, in such a case.

SUMRAT and SAMRAT can also be used to find the best characters to be used for the clustering procedure. However, in this case, they have to be supplemented with biological judgment.

Section ii: Clustering analysis

Once the information about the objects has been properly structured (formed into characters and character states), it is relatively simple to obtain a clustering of the objects that will be used by the biologist as an indication or a basis for his classification. The main problem is to choose the proper model on which the grouping of the objects will be based. A model, in this case, is a series of definitions and rules of inference which reflect as closely as possible what one wants a classification to be. From these, equations are derived that allow one to calculate, from the structured information about the objects, results that will be in accordance with the model.

The computer-aided method explained here is a clustering technique based on a model in graph theory developed at the Taximetrics Laboratory under the direction of the junior author. It is intended to follow as closely as possible the mental process of the classical taxonomist. The reader may refer to Wirth, Estabrook and Rogers (1966), Estabrook (1966) and Estabrook and Rogers (1966) for a different presentation of parts of this method. A flow-chart-like presentation of the algorithm can be found in Estabrook (1966).

The premises of the model on which this technique is based are the following:

1) A classification for a collection of objects is a hierarchical, two-dimensional partitioning of the objects. A partition of the objects is a subdivision of the collection into sub-collections, such that each object is in one and only one sub-collection, for the given partition. The hierarchial partitioning is said to be two-dimensional because it can be represented in a plane, or a

sheet of paper: on one of the axes are presented the various sub-collections of each partition, the various partitions being ordered on the other axis according to their hierarchy. The following example illustrates this definition, with the partitions into genera and into species.

	hierarchical presentation of 2 partitions		Objects
the various sub-collections of each partition	Genus 1	Species 1	7, 12
		Species 2	3, 5, 11
		Species 3	1, 2, 6
	Genus 2	Species 4	4, 9
		Species 5	8, 10, 13, 14

2) For any given partition, two objects at least as similar as the degree considered for this partition should not be placed into different sub-collections (this can be modified later by the taxonomist if, from his knowledge of the evolution of the group, he considers the similarity relation as secondary, for instance in the case of cryptic species).

3) The sub-collections of a given partition should be isolated from one another: that is, there should be some discontinuity between the members of different sub-collections, found in the structure of the information available about the objects.

From these three widely accepted principles will be derived the graph theory model in the section on the clustering technique.

Various techniques for clustering biological objects have been developed in the past. After comparing four of them, Prance *et al.* (1969) concluded that the one explained hereafter gave the most useful results and presented them in the most informative way. Furthermore, the experience of the Taximetrics Laboratory has shown that this method leads to valuable results in botany as well as in zoology, with lower as well as higher organisms, with cultivated as well as wild plants, and in classifications in the fields of ecology, anthropology, geology, psychology and sociology.

The method consists of two main steps. First, a similarity measure is calculated between all the pairs of objects in the study. Second, in decreasing similarity order, the objects that are similar at least to the considered level, or degree of similarity, are connected in various clusters. The clusters become more and more inclusive as the similarity level drops and the procedure stops when all the objects form a single cluster. Useful measures of connectedness and isolation, calculated for each of the intermediate clusters, help the taxonomist to interpret the results.

Example

For the sake of clarity, the example introduced here will be carried through the following sections. Rogers and Fleming (1972) used fifteen characters to describe specimens of the cultivated species *Manihot esculenta*, and through the use of the taximetric methods here explained, they arrived at a classification into 'groups', which correspond to phenons of this widely cultivated crop. Five objects were chosen from each of their groups 1 and 14 to serve here as an illustration of the clustering technique. The descriptions are given in Table IX.

TABLE IX. Description of the objects used in the example (from: Rogers and Fleming, 1971). Each object is described by listing which state of each character has been attributed to it.

Object number	Characters														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
108	2	2	1	2	2	4	2	6	1	4	3	2	2	4	3
114	1	4	1	1	2	3	2	6	2	2	4	2	1	4	3
131	1	5	1	1	2	3	2	6	2	2	4	3	2	4	3
132	1	3	1	1	2	3	1	6	2	2	4	2	1	4	3
242	1	3	1	1	2	2	3	6	2	3	4	3	1	3	3
281	2	2	1	5	2	2	3	5	1	4	3	2	2	4	3
284	2	2	1	5	2	3	2	6	1	4	3	2	2	1	3
330	2	2	1	5	1	2	3	2	1	4	3	1	2	1	1
377	2	2	1	5	2	4	3	5	1	4	3	2	2	1	1
454	1	3	3	1	2	2	2	5	2	2	4	3	1	4	2

Similarity measure

Various similarity measures have been discussed extensively by Sokal and Sneath (1963), chapter 6. The one chosen here is very "natural" in its most simple form: the similarity S(a, b) attributed to the pair of objects *a* and *b* will be the number of characters for which the same state has been attributed to objects *a* and *b*, divided by the total number of characters. For example, the similarity S(108, 281) is calculated as follows: the states of the fifteen characters attributed to each of these objects are:

108: 2 2 1 2 2 4 2 6 1 4 3 2 2 4 3
281: 2 2 1 5 2 2 3 5 1 4 3 2 2 4 3

then for eleven of the fifteen characters, the same state has been attributed to the two objects. S(108, 281) is then equal to 11/15 or 0.733.

This measure of similarity has the following properties:

1) $0 \leqslant S(a, b) \leqslant 1$.

2) S(a, b) = 1 implies that *a* and *b* are maximally similar (identical).

3) S(a, b) > S(c, d) implies that the pair (a, b) is more mutally similar than is the pair (c, d).

4) S(a, a) is always equal to 1.

5) S(a, b) = S(b, a).

With only characters of type 1 (see below), it is easy to see that in this simple form, the measure of similarity could not engender more than N + 1 levels of similarity, where N is the number of characters in the study. In the example, N = 15 then only 16 levels could be produced: S = 15/15, S = 14/15, . . ., S = 1/15, S = 0/15.

Two sophistications have been added to this measure of similarity, however. The first one considers that the taxonomist might wish to say that two objects in different states of a character are partially similar, instead of completely different according to this character.

The character for which a pair of objects in two different states is not necessarily more similar than any other pair of objects in any two different states, is the one for which the simplified equation above is directly applicable. This character is said to be of type 1 (simple character).

A character of type 2 (ordered character) is one in which the states form

SYMBOLS USED IN THIS SECTION

a, b: names of two objects

S(a, b): similarity attributed to the pair of objects a and b

K: name of a character

K_i, K_j: two of the states of character K

i, j: designate two of the states of character K, and also, in the case of an ordered character, the placement of K_i and K_j in the ordered sequence of character states

K_0: state of character K attributed to the objects on which the character was not observable, practically; K_0 is never included in a partial similarity measure

$n(K_i, K_j)$: similarity value calculated, on character K only, for a pair of objects to which states i and j of character K have been attributed, respectively

n(K(a), K(b)): similarity value attributed, for character K only, to the pair of objects a and b; K(a) and K(b) are the states of character K attributed to the objects a and b respectively

f(d, k): function of d and k

d: 'distance' between state K_i and state K_j, in an ordered character, that is equal to $|i - j|$

k: parameter set by the worker, when he uses the partial similarity formula, to indicate the largest 'distance' $|i - j|$ between character states in the sequence, for which he wishes to make a non-0 assignment of partial similarity

M: number of states of a character

$N_{(a, b)}(K)$: a character K of which state K_0 has been attributed to object a, or object b, or both

a logically well ordered sequence, and the similarity for any pair of objects in different character states depends on how far apart in the sequence the two states in question occur. Such a character may also be treated as a type 3 character, as will be seen below. However, whenever the states of a character can be logically ordered in such a way that this ordering reflects a continuous or semi-continuous gradation from state to state along the sequence, an equation of partial similarity can be applied. The following empirical equation is included in the computer program:

$$n(K_i, K_j) = f(d, k) = \begin{cases} \dfrac{2\ (k+1-d)}{2k+2+dk} & \text{whenever } d \leqslant k \\ = 0 & \text{when } d > k, \end{cases}$$

where $n(K_i, K_j)$ is the similarity value calculated for a pair of objects to which states *i* and *j* of character K have been attributed respectively. The subscripts *i* and *j* indicate also the placement of K_i and K_j in the ordered sequence of character states. This similarity value, between 0 and 1, is a function f(d, k) of the "distance" *d* between state K_i and state K_j, expressed by the absolute value $|i - j| = d$, and of the parameter k, set by the taxonomist for each character in one of the computer cards that indicate the largest "distance" $|i - j|$ between character states in the sequence, for which the worker wishes to make a non-0 assignment. Whenever the taxonomist wishes to state that a given character is of type 1 (see above), he makes k equal to 0. The properties of this formula are the following:

1) When *d* is larger than *k*, the objects *a* and *b* in the pair of character states K_i and K_j are considered as not similar at all for this character, i.e., $n(K_i, K_j) = 0$, or in another formulation $n(K(a), K(b)) = 0$ where K(a) and K(b) are the states of character K attributed to the objects *a* and *b*.

2) When d = 0, the value n(K_i, K_j) is found and it is equal to 1, as expected.

3) f(d, k) decreases as *d* increases for fixed *k*.

4) f(d, k) increases as *k* increases for fixed *d*.

Even if this formula is empirical, it reflects the judgments made by those workers who used this method, and it has been proven useful. The values of f(d, k) for the most commonly used values of *k* are given in Table X.

TABLE X. The values of f(d, k) for the most commonly used values of *k* (see equation in the text).

k	d =						
	1	2	3	4	5	6	(etc.)
1	.40	0.00	0.00	0.00	0.00	0.00	0.00
2	.50	.20	0.00	0.00	0.00	0.00	0.00
3	.55	.28	.12	0.00	0.00	0.00	0.00
4	.57	.33	.18	.08	0.00	0.00	0.00
5	.59	.36	.22	.13	.05	0.00	0.00

A few remarks can be made about this equation. First, if M is the number of states of a character that form an ordered sequence, it is advised not to make *k* larger than M – 2: since the largest "distance" *d* is M – 1, if *k* is made equal to M – 1, no pair of states will receive the assignment 0 and this character will not differentiate objects as adequately as it could. M is of course the number of real states of the character, at the exclusion of state K_0, which is never considered in a partial similarity measure.

A discontinuous gradation can be shown in ordered or semi-ordered characters. An example of a semi-ordered character would be the dimension of the eggs produced by various fishes, in which *k* is made equal to 1:

K_1 = up to .7 mm in diameter, incl.
K_2 = larger than .7, up to 1.3 mm incl.
K_3 = larger than 1.3 up to 1.5 mm incl.
K_4 = larger than 1.5 mm
K_5 = (void state)
K_6 = not *logically* applicable because it is a male
K_0 = eggs not observable (young specimen or not in season).

Note the difference between K_6 and K_0. The state 0 of a character, which means "no information available", is never included in the computation of a partial similarity, that is to say that n(K_i, K_j) = 0 if either i or j is 0. Accordingly, an exception has to be made to rule 2 above: when d = 0 and i = 0, then n(K_i, K_j) = 0, instead of 1. The introduction of a void state K_5 was necessary because K_6 should not be made partially similar to K_4. With a void state in K_5 and k = 1, the partial similarity measure will be calculated only between states 1 and 2, 2 and 3, and 3 and 4. If k had been 2, two void states would have had to be created. The same trick can be used whatever the reason is to limit the partial similarity to certain states. Another method to obtain the same result would be to consider this character as one of type 3 below.

The general formula for the similarity measure between objects *a* and *b* now becomes

$$S(a, b) = \frac{\sum\limits_{\text{all characters K}} n\,(K(a), K(b))}{\text{total number of characters}}$$

where n(K(a), K(b)) is the similarity value calculated between objects *a* and *b* for character K.

A partial similarity was defined on character 6 of the example, with k = 1; k = 0 for the other characters. The similarity S(284, 377) between objects 284 and 377 of the example would be calculated as follows:

Characters	Objects		n(K(284), K(377))
	284	377	
1 k = 0	2	2	1.00
2 k = 0	2	2	1.00
3 k = 0	1	1	1.00
4 k = 0	5	5	1.00
5 k = 0	2	2	1.00
6 k = 1	3	4	0.40
7 k = 0	2	3	0.00
8 k = 0	6	5	0.00
9 k = 0	1	1	1.00
10 k = 0	4	4	1.00
11 k = 0	3	3	1.00
12 k = 0	2	2	1.00
13 k = 0	2	2	1.00
14 k = 0	1	1	1.00
15 k = 0	3	1	0.00

S(284, 377) = 11.40 / 15 = 0.76000

In a character of type 3 (matrix character) the states are not logically ordered, but some pairs of objects in non-identical states are judged by the worker to be more similar to each other, according to this character, than some other pairs of objects in another pair of states. In this instance, it is necessary for the worker to make a judgement and decide the value that $n(K_i, K_j)$ takes for each pair of states K_i and K_j, except for the state K_0, "missing information", which has a similarity of 0 with all the other states. 1) This value still has to be between 0 and 1. 2) When i = j, this value has to be 1. 3) If K_i or K_j describes the logical inapplicability of character K, then the assignment should be 0, except when i = j.

Character 14 of the example is a type 3 character. It describes the petiole color as follows:
K_1 = red
K_2 = greenish red
K_3 = reddish green
K_4 = green

The similarity assignment made by Rogers and Fleming (1971) for the various pairs of objects to which the various states of this character have been attributed is best represented by half of a matrix, the other half being symmetrical:

	K_1	K_2	K_3	K_4
K_1	1.00			
K_2	.75	1.00		
K_3	.25	.75	1.00	
K_4	0.00	.25	.75	1.00

The similarity S(132, 242) between objects 132 and 242 of the example would be calculated as follows:

Characters	Objects		n(K(132), K(242))
	132	242	
1 k=0 type 1	1	1	1.00
2 k=0 type 1	3	3	1.00
3 k=0 type 1	1	1	1.00
4 k=0 type 1	1	1	1.00
5 k=0 type 1	2	2	1.00
6 k=1 type 2	3	2	0.40
7 k=0 type 1	1	3	0.00
8 k=0 type 1	6	6	1.00
9 k=0 type 1	2	2	1.00
10 k=0 type 1	2	3	0.00
11 k=0 type 1	4	4	1.00
12 k=0 type 1	2	3	0.00
13 k=0 type 1	1	1	1.00
14 k=0 type 3	4	3	0.75
15 k=0 type 1	3	3	1.00

S(132, 242) = 11.15 / 15 = 0.74333

The similarity assignments in a type 2 character can also be treated in the same way, if the worker considers the similarity assignments given by the formula, and shown in Table X, as not satisfactory. The similarity assignments in a type 1 character, on the other hand, can always be represented by a matrix in the form

	K_1	K_2	K_3	K_4	K_5
K_1	1				
K_2	0	1			
K_3	0	0	1		
K_4	0	0	0	1	
K_5	0	0	0	0	1

It is easy to see that more than N + 1 levels of similarity can be engendered by a type 2 or a type 3 character, although the actual number of levels remains usually small.

The second sophistication added to the formula giving the similarity value is one that causes the similarity value to be calculated only on the characters for which the information is available for the two objects compared. Indeed, the taxonomist often has to work with preserved specimens, some of which may lack information about one or a few characters. If the overall similarity between objects is computed without taking this into account, every missing piece of information about an object will unduly lower the similarity values of comparisons involving this object. This is the problem with many similarity measures. Instead, it is more practical to calculate the overall similarity of a pair of objects only on those characters for which information is present, forgetting the characters for which one or the other object of the pair has been coded in state K_0, "no information available".

This can be done by establishing that $N_{(a,b)}$ (K) is 1 if and only if $K(a) = K_0$, or $K(b) = K_0$, or both, that is to say when the state "no information available" has been attributed to one of the objects *a* or *b*, and $N_{(a,b)}$ (K) is 0 in all the other cases. Then the modified formula describing the similarity between objects *a* and *b* can be written:

$$S(a, b) = \frac{\displaystyle\sum_{\text{all characters K}} n\,(K(a), K(b))}{(\text{total number of characters}) - \displaystyle\sum_{\text{all characters K}} N_{(a,\,b)}(K)}$$

This formula is the one used in the clustering program explained here.

The similarity measures calculated between all the pairs of objects of Table IX can be represented by the half-matrix shown in Table XI. The other half of it is symmetrical, since S(a, b) = S(b, a).

TABLE XI. The similarity measures calculated between all pairs of objects of table IX (the other half of the table is symmetrical).

	108	114	131	132	242	281	284	330	377	454
108	1.00000									
114	.49333	1.00000								
131	.49333	.80000	1.00000							
132	.42667	.86667	.73333	1.00000						
242	.31667	.67667	.67667	.74333	1.00000					
281	.73333	.36000	.36000	.36000	.38333	1.00000				
284	.82667	.46667	.46667	.40000	.31000	.76000	1.00000			
330	.46667	.09333	.16000	.09333	.21667	.66667	.62667	1.00000		
377	.66667	.22667	.22667	.22667	.21667	.80000	.76000	.73333	1.00000	
454	.20000	.62667	.62667	.62667	.65000	.26667	.16000	.06667	.13333	1.00000

In the listing of similarity values corresponding to this table, which forms the first part of the printout, the similarity value is given for each pair of objects in the study, followed by the number of characters used in establishing this similarity value.

Before proceeding to the clustering, the computer has one more intermediate step: ordering the similarity measures obtained before, in decreasing order of similarity. This is done for the example in Table XII. The format of this table is not the one followed by the computer: the presentation in a table is simply better adapted to this paper. If some pairs of objects are identical (S = 1.00000) these pairs are presented at the beginning of the printout. The table does not present either the similarity between an object and itself, which is obviously 1.00000.

Clustering procedure

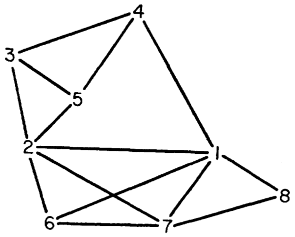
A definition of the cluster concept will be given first, then it will be shown that this definition is in accordance with the three general principles of taxonomy developed in the introduction.

Table XII. Ordered similarity ratios. The data of table XI are re-arranged in decreasing order of similarity.

Similarity values	Pairs of objects similar at the given values
.86667	(132, 114)
.82667	(284, 108)
.80000	(131, 114) (377, 281)
.76000	(284, 281) (377, 284)
.74333	(242, 132)
.73333	(132, 131) (281, 108) (377, 330)
.67667	(242, 114) (242, 131)
.66667	(330, 281) (377, 108)
.65000	(454, 242)
.62667	(330, 284) (454, 132) (454, 114) (454, 131)
.49333	(114, 108) (131, 108)
.46667	(284, 114) (284,131) (330, 108)
.42667	(132, 108)
.40000	(284, 132)
.38333	(281, 242)
.36000	(281, 114) (281, 131) (281, 132)
.31667	(242, 108)
.31000	(284, 242)
.26667	(454, 281)
.22667	(377, 114) (377, 132) (377, 131)
.21667	(330, 242) (377, 242)
.20000	(454, 108)
.16000	(330, 131) (454, 284)
.13333	(454, 377)
.09333	(330, 114) (330, 132)
.06667	(454, 330)

A cluster is a set of connections, and of objects which are interconnected at any given level of similarity, i.e., a group of objects for which there exists at least one continuous pathway of connections joining all the objects. There exists a connection between two objects, at a given similarity level, if these two objects are at least as similar to each other as the similarity level considered. The similarity values for the pairs of objects are determined by using the equation developed in the previous section.

A cluster can be represented by placing the objects anywhere in a plane and drawing lines between the objects that are interconnected at this level, such as:



According to this definition, each object does not have to be connected to each of the others to be member of the cluster. This fits many biological cases, like a species modified along a cline, in which the extremes are not very similar to each other, but they are connected through a pathway of high similarity values. This is similar to the "single linkage" concept of Sneath (1956).

A cluster can also be defined by the relation G_c: for any pair of objects *a* and *b* in the study,

$a\ G_c\ b \leftrightarrow S(a, b) \geqslant c$

this is to be read: "*a* is connected with *b* at level of similarity *c* if, and only if, the over all similarity value between *a* and *b* is larger than or equal to *c*". *c* is a value between 0 and 1, as previously. The properties of symmetry and reflexivity of G_c are clear, since $aG_cb \leftrightarrow bG_ca$, and aG_ca always. This is actually a reformulation of the second premise of the introduction (ref. to Section II, p. 591): every two objects to which a similarity value at least as large as *c* has been attributed, will be included in the same cluster. Furthermore, all the objects at least as similar to another one as *c* will be partitioned into clusters at level *c*. The other objects form single-member clusters at this level. A group of clusters such as defined above by the relation G_c applied to the set U of objects in the study is technically called an *undirected graph* (for more details, see a textbook of graph theory).

The notion of a G_c-chain can now be introduced. A G_c-chain is said to exist from *a* to *b* if there is a series of points $d_1, d_2, d_3, \ldots, d_i$ in U such that aG_cd_1 and $d_1G_cd_2$ and $d_2G_cd_3$ and ... and d_iG_cb. The unique equivalence relation R_c can now be defined: aR_cb if and only if there exists a G_c-chain from *a* to *b*. This means that two objects *a* and *b* will be in relation with each other at level *c* (will be in the same cluster) if there exists a connection between them, a connection that can be established through other intermediate objects. In other words, R_c is a transitive G_c. R_c is an equivalence relation since it has the following properties:

1) it is clearly reflexive: aR_ca always since a G_ca always;

2) since G_c is symmetric, the G_c-chains can be turned around, and then

the existence of a chain from *a* to *b* implies the existence of a chain from *b* to *a*;

3) if a G_c-chain exists from *a* to *b* and another exists from *b* to *d*, then these two chains can be combined; consequently aR_cb and bR_cd implies that aR_cd, or in other words the relation is transitive.

The result of this process is to group the objects in various clusters at any given level of similarity. Consequently the relation G_c defines a partition of the objects at each level of similarity, as was required by premise 1. Furthermore, the partitioning process is hierarchical when considered along the axis of decreasing similarity, since the clusters will connect to each other as the similarity value drops. At high similarity values, there are many clusters, which become fewer and larger as the similarity value drops, finally attaining the point where there is only one cluster which includes all the objects in the study. Furthermore, pairs of points which are in the same cluster at high similarity value remain inseparable for all the lower similarity values.

If G_c together with U is a graph, then each cluster is a connected subgraph. A subgraph of (U, G_c) is a subset of U, called V, together with the relation G_c restricted to the objects included in V. A subgraph (V, G_c) is connected if there exists a G_c-chain from *a* to *b* for all pairs of points *a* and *b*, which are elements of V.

The C-values that will be used in the printout as principal markers are defined as those values of similarity at which at least one of the clusters is modified by addition of at least one new object. The C-values are the only levels of similarity which will be considered in the clustering printout, but internal connections arising between two C-levels will be recorded. A connection is called *external* when it results in the addition of a new object to the cluster, thus modifying the composition in objects of the cluster. An *internal* connection is one that occurs, at a given similarity level, between two objects that were both already members of the cluster. It modifies only the composition in connections, by increasing the tightness of the connections within the cluster. At every C-level, a measure of tightness of the connections of each cluster will be given by a statement about the number of connections actually formed in the cluster at this point, and about the maximum number of possible connections which can be formed between the objects of the cluster. If a cluster has M objects, the maximum number of possible connections is $(M-1) + (M-2) + (M-3) + \ldots + 2 + 1$, which is equal to $M(M-1)/2$ as can be easily proved by induction over M. The measure of connectedness at each C-level can then be easily calculated: it is the ratio of the number of connections actually formed at the given C-level, to the maximum possible number of connections between the objects present in the cluster.

A more sophisticated mathematical discussion of the principles involved in the clustering process has been given by Estabrook (1966).

Every time a cluster is modified by addition of new objects, a measure of isolation, called moat, is also given. It states how "long" on the axis of decreasing similarity the worker will have to wait before the cluster will be modified again by addition of new objects. The statement made on the printout is of the kind: "MOAT = .06667* NEXT PAIRS TO JOIN (284,

* The computing machine runs out 12 decimals, but it was arbitrarily chosen to use 5 of them.

281) (377, 284)" (level 2 in figure 2) meaning that the cluster will be modified only after .06667 units of similarity, and then the modification will consist of the connection of objects 281 and 377 to object 284 that was already a member of the cluster.

The moat can be thought of as a measure of stability as well as a measure of isolation of the clusters. It can also be seen as an indirect measure of isolation between the various clusters formed by each of the partitions of the hierarchical series, at each C-value, thus applying premise 3.

The clustering process can now begin. The first C-level is .86667. Table XII says that the similarity value for the pair of objects (132, 114) is .86667. One cluster is formed at this level. The following data are given in the printout (figure 2).

CLUSTER MEMBERSHIP: 114, 132
C-VALUE: .86667
CONNECTEDNESS:

1 connection formed / $\frac{M(M-1)}{2} = \frac{2 \times 1}{2} =$ 1 possible connection

The G_1-chain defining the equivalence relation R_1 is formed by the pair (132, 114)
MOAT = .06667 NEXT PAIR TO JOIN (131, 114)

This last piece of information means that it will take .06667 units of similarity before this cluster gets modified. At .86667 — .06667 = .80000 units of similarity (level 3 in figure 2), object 131 will form a connection with object 114.

At the end of each C-value, the printout shows how many objects in the study have not yet been included in a cluster. In the example, at the end of the first level, this line is read: SINGLE MEMBER CLUSTERS (8), which is followed by a list of these 8 objects.

Let us consider now level 6 of the example (figure 2). Two clusters are present at this level. For the first one, the following information is given:

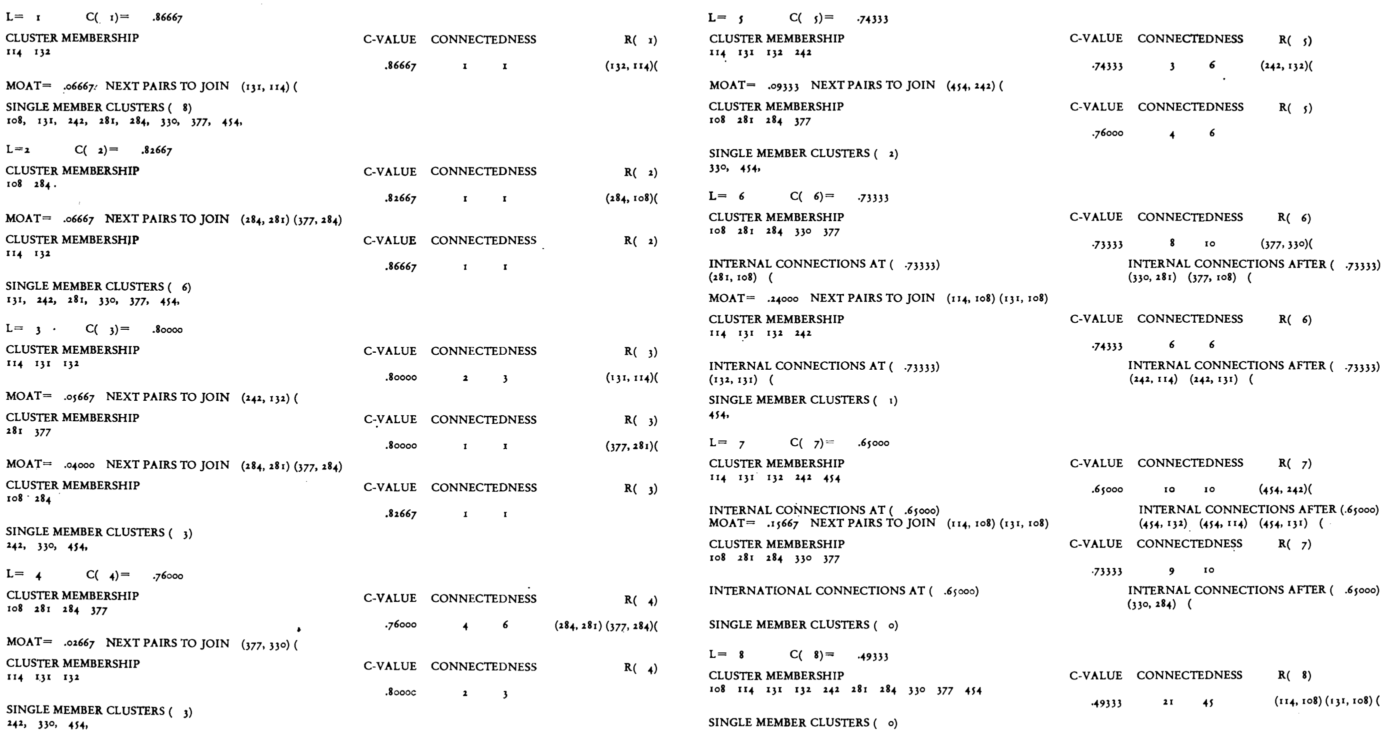
CLUSTER MEMBERSHIP: 108, 281, 284, 330, 377
C-VALUE: .73333
CONNECTEDNESS:

8 connections formed (before, at and after the level; see below) / $\frac{M(M-1)}{2} = \frac{5 \times 4}{2} =$ 10 possible connections

The equivalence relation R_6 is established by connecting the pair (377, 330)
INTERNAL CONNECTIONS AT (.73333) : (281, 108)

INTERNAL CONNECTIONS AFTER (.73333) : (330, 281) (377, 108)
MOAT = .24000 NEXT PAIRS TO JOIN (114, 108) (131, 108)

The internal connections at the level, which is C_6 = .73333 in this case, are those connections happening exactly at the given similarity level, between pairs of objects that were already members of the cluster. The internal connections after the level are those also happening between pairs of objects that were already members of the cluster, and occurring after the given level of similarity (C_6 = .73333 here) but before the next level of similarity (C_7 = .65000 here). It is interesting to know about these connections, since they increase the tightness of the connectedness of the clusters and can help in deciding about systematic questions. However, since the



```
L=  1         C(  1)=    .86667
CLUSTER MEMBERSHIP                                C-VALUE   CONNECTEDNESS            R(  1)
114   132
                                                   .86667       1      1             (132, 114)(
MOAT=  .06667  NEXT PAIRS TO JOIN   (131, 114) (
SINGLE MEMBER CLUSTERS (  8)
108,  131,  242,  281,  284,  330,  377,  454,

L=2           C(  2)=    .82667
CLUSTER MEMBERSHIP                                C-VALUE   CONNECTEDNESS            R(  2)
108   284.
                                                   .82667       1      1             (284, 108)(
MOAT=  .06667  NEXT PAIRS TO JOIN   (284, 281) (377, 284)
CLUSTER MEMBERSHIP                                C-VALUE   CONNECTEDNESS            R(  2)
114   132
                                                   .86667       1      1
SINGLE MEMBER CLUSTERS (  6)
131,  242,  281,  330,  377,  454,

L=  3  .      C(  3)=    .80000
CLUSTER MEMBERSHIP                                C-VALUE   CONNECTEDNESS            R(  3)
114   131   132
                                                   .80000       2      3             (131, 114)(
MOAT=  .05667  NEXT PAIRS TO JOIN   (242, 132) (
CLUSTER MEMBERSHIP                                C-VALUE   CONNECTEDNESS            R(  3)
281   377
                                                   .80000       1      1             (377, 281)(
MOAT=  .04000  NEXT PAIRS TO JOIN   (284, 281) (377, 284)
CLUSTER MEMBERSHIP                                C-VALUE   CONNECTEDNESS            R(  3)
108   284
                                                   .82667       1      1
SINGLE MEMBER CLUSTERS (  3)
242,  330,  454,

L=  4         C(  4)=    .76000
CLUSTER MEMBERSHIP                                C-VALUE   CONNECTEDNESS            R(  4)
108   281   284   377
                                                   .76000       4      6     (284, 281) (377, 284)(
MOAT=  .02667  NEXT PAIRS TO JOIN   (377, 330) (
CLUSTER MEMBERSHIP                                C-VALUE   CONNECTEDNESS            R(  4)
114   131   132
                                                   .8000c       2      3
SINGLE MEMBER CLUSTERS (  3)
242,  330,  454,
```

```
L=  5         C(  5)=    .74333
CLUSTER MEMBERSHIP                                C-VALUE   CONNECTEDNESS        R(  5)
114   131   132   242
                                                   .74333       3      6         (242, 132)(
MOAT=  .09333  NEXT PAIRS TO JOIN   (454, 242) (
CLUSTER MEMBERSHIP                                C-VALUE   CONNECTEDNESS        R(  5)
108   281   284   377
                                                   .76000       4      6
SINGLE MEMBER CLUSTERS (  2)
330,  454,

L=  6         C(  6)=    .73333
CLUSTER MEMBERSHIP                                C-VALUE   CONNECTEDNESS        R(  6)
108   281   284   330   377
                                                   .73333       8      10        (377, 330)(
INTERNAL CONNECTIONS AT (  .73333)                   INTERNAL CONNECTIONS AFTER (  .73333)
(281, 108)  (                                        (330, 281)  (377, 108)  (
MOAT=  .24000  NEXT PAIRS TO JOIN   (114, 108) (131, 108)
CLUSTER MEMBERSHIP                                C-VALUE   CONNECTEDNESS        R(  6)
114   131   132   242
                                                   .74333       6      6
INTERNAL CONNECTIONS AT (  .73333)                   INTERNAL CONNECTIONS AFTER (  .73333)
(132, 131)  (                                        (242, 114)  (242, 131)  (
SINGLE MEMBER CLUSTERS (  1)
454,

L=  7         C(  7)=    .65000
CLUSTER MEMBERSHIP                                C-VALUE   CONNECTEDNESS        R(  7)
114   131   132   242   454
                                                   .65000       10     10        (454, 242)(
INTERNAL CONNECTIONS AT (  .65000)                    INTERNAL CONNECTIONS AFTER (.65000)
MOAT=  .15667  NEXT PAIRS TO JOIN   (114, 108) (131, 108)   (454, 132)  (454, 114)  (454, 131)  (
CLUSTER MEMBERSHIP                                C-VALUE   CONNECTEDNESS        R(  7)
108   281   284   330   377
                                                   .73333       9      10
INTERNATIONAL CONNECTIONS AT (  .65000)              INTERNAL CONNECTIONS AFTER (  .65000)
                                                     (330, 284)  (
SINGLE MEMBER CLUSTERS (  0)

L=  8         C(  8)=    .49333
CLUSTER MEMBERSHIP                                C-VALUE   CONNECTEDNESS        R(  8)
108   114   131   132   242   281   284   330   377   454
                                                   .49333       21     45        (114, 108) (131, 108) (
SINGLE MEMBER CLUSTERS (  0)
```

Fig. 2: Facsimile printout corresponding to the eight C-levels of the example. For the meaning of the various expressions, see text. (In column 4, sixth line from bottom read 'internal' instead of 'international'.

taxonomist is interested, at this stage of the game, in obtaining a clear picture of the partitions, then C-levels are recognized only when there is a change in the object membership of the cluster; internal connections occurring between two C-levels are listed with the activity of the superior C-level.

Internal connections occurring at or after the last C-level (C_8 = .49333 in the example) are not listed, because there are usually too many of them (24 of them, over a total of 45 connections, in the example: see Table XII), and because they are not taxonomically interesting. This is why the clustering process stops after all the objects in the study have been included in the same cluster.

Data of figure 2 can be represented by drawing the corresponding subgraphs, such as in figure 3 which shows the evolution of the clustering procedure for the example. The two clusters of level 7, which join at C_8 = .49333, correspond respectively to parts of groups 1 and 14 in Rogers and Fleming (1971).

One could draw a separate subgraph for those internal connections formed between two consecutive C-levels. However, it has been found more useful to include these connections with either the subgraph of the superior C-level, or the one of the inferior C-level, designating them by a symbol different from the one used for the internal connections formed at the C-level, or not. The decision is left to the worker, according to his needs, as it is the case with all other matters of graphical representation.

Various graphical symbols such as circles, boxes, etc. can be used to point out some aspects of the data (see, for instance, Prance *et al.*, 1969).

By entering the proper code on one of the parameter cards, the worker may obtain on the printout, after the actual clustering of the objects, and for each object in the study, a listing of the ten objects most similar to it, with the corresponding similarity values. Such a listing can be useful in the establishment of the final taxonomic structure. A "skyline" plot, which is also available, summarizes the results of the clustering process by showing the clusters that were formed and the value of similarity at which they occur. This plot also shows the objects' hierarchical relationship and the measure of isolation (moat) of each cluster. The ordinate is the similarity scale, and the objects are distributed on the abscissa.

Interpretation

The only generally valid statement that can be made in this section is that the results of the clustering analysis are an aid for the worker, helping him to discern the taxonomic structure of the objects he studies. But it is a powerful aid. How he will use these results is left to him. It depends mainly on the group of organisms with which he is working. This section is intended to give some hints — and not recipes — on how to use the printout to the best advantage.

Some groups of organisms will easily show their taxonomic structure, simply by drawing the subgraphs. No program exists yet to draw these subgraphs, for practical reasons like the size of the memory of the machines. But also it has been the experience of the Taximetrics Laboratory that 1) the subgraphs can be drawn in very many ways, depending on the group of organisms under study, in order to carry the most information, and that 2) the worker learns very much about the classification of his organisms simply by drawing the subgraphs and thinking about the ways to make them carry as much information as possible.

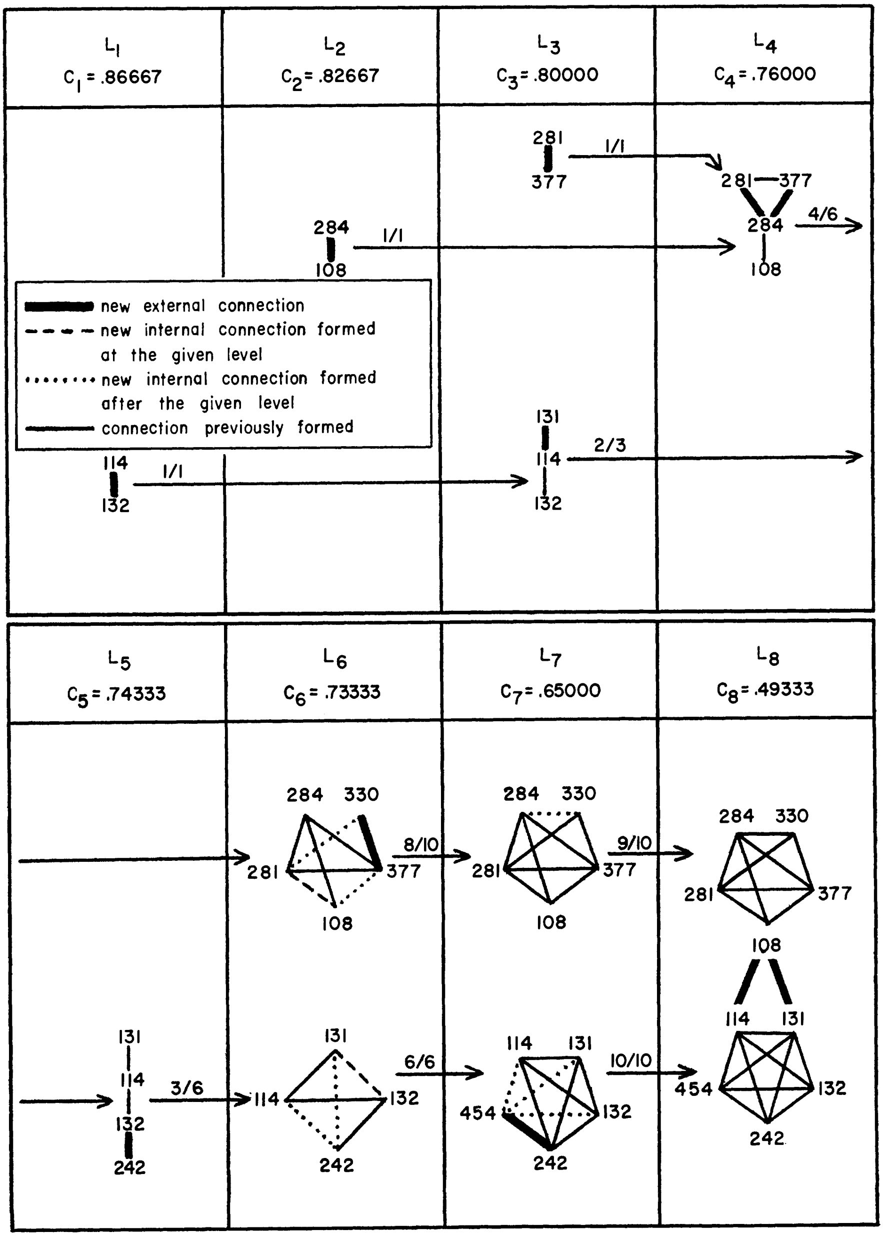


Fig. 3: Subgraphs drawn from the data of figure 2. Lı: the various levels of similarity considered. Cı: the various C-levels, followed by the value of similarity to which they correspond. Each cluster drawn is cumulative of the activity of the previous levels. The fraction shown on each arrow indicates the degree of connectedness of the cluster.

It was mentioned in Section I that, depending on the attitude of the worker, the group of organisms under study, and the knowledge the worker has about the organisms he studies, it is possible to start a study by a character analysis followed by a clustering analysis, or vice versa. It also happens that only one of these analyses gives enough information to allow the worker to make his classification, as it happens also that none of these analyses are necessary. Similarly, it can happen that a skyline plot will give enough information by itself, when in other cases it is not necessary at all. Some workers have best used this pictorial visualization at the beginning of the analysis of the cluster formation, others at the end, as a complement. In all cases, however, the moat values, given on the skyline plot as well as in the clustering analysis, have been seen as of major importance, since they measure the degree of isolation of the clusters.

In another case, the analysis of the number of connections formed between populations of objects, at the various C-levels, has been found to be a valuable complement to the study of the subgraphs, allowing to give a "distance" measure between the various taxonomic categories and between the taxa involved in the study (Legendre, Schreck and Behnke, 1972).

Another question of interest is: how different do species, or genera, have to be? There is no precise answer to this question. It has been the experience at the Taximetrics Laboratory that congeneric species are often 50% to 60% similar to each other. The species level is often found around 75% of similarity, and subspecies around 85% similarity. Large departures from these values have also been found. It obviously depends on 1) the group of objects considered and 2) how well the information about the objects is structured. If the characters are chosen and subdivided into character states in such a way as to be consistent with each other, the partitioning of the objects will be much easier, and the clusters will be more isolated from each other, thus causing the similarity value that corresponds to each taxonomic category to drop. The exaggerated use of partial similarities, with the equation (type 2 character) or with matrices (type 3 character) produces the opposite effect, making more similar to each other objects that belong to different taxa.

After the clustering analysis, it is always important to bring the geographical, cytological and ecological data (when available) into the study, trying to correlate the preliminary taxonomic structure with these data. This is also the time to try to fit into the structure those objects for which a large amount of information is missing, or which are considered as hybrids and then were left out of the analysis in order to clarify the structure first. Those characters that were not considered in the study for various reasons, such as missing information for many of the objects, and also other classifications can be compared with the preliminary structure. This process has been discussed by Rogers and Appan (1969).

One has to be a biologist to make a biological classification, or a sociologist to make a sociological one, since only a specialist (*sensu lato*) can interprete in the correct way the various attributes of the objects under study. An advantage of biology over, for instance, geology, is that one knows that there is a genetic basis for the similarities and dissimilarities observed, and thus there is hope to find a "natural" classification. But one does not have to be a mathematician to understand and use the methods explained here; by applying them, he will realize that they are intended to help him to work according to his own mental process as a taxonomist.

Availability of the programs

Available: Fortran IV listings, CHARANAL and Graph flow-charts.

At IBM: Graph, version of the program written for an IBM 7044 is available from the Program Information Department, IBM, 40 Saw Mill River Road, Hawthorne, N. Y. 10532. Library #3501.

References

Boltzmann, L. 1896 – Vorlesungen über Gastheorie, I. Teil. J. A. Barth, Leipzig. 265.

Boltzmann, L. 1898 – Vorlesungen über Gastheorie, II. Teil. J. A. Barth, Leipzig.

Estabrook, G. F. 1966 – A mathematical model in graph theory for biological classification. J. Theoret. Biol. 12: 297-310.

Estabrook, G. F. 1967 – An information theory model for character analysis. Taxon 16: 86-97.

Estabrook, G. F. and D. J. Rogers 1966 – A general method of taxonomic description for a computed similarity measure. BioScience 16 (11): 789-793.

Hawksworth, F. G., G. F. Estabrook and D. J. Rogers 1968 – Application of an information theory model for character analysis in the genus *Arceuthobium* (Viscaceae). Taxon 17 (6): 605-619.

Legendre, P., C. B. Schreck and R. J. Behnke 1972 – Taximetric analysis of selected groups of western North American *Salmo* with respect to phylogenetic divergences. Systematic Zoology 21(2): 292-307.

Mayr, E. 1970 – Populations, species, and evolution. Harvard Univ. Press, Cambridge. xv + 453 pages.

Michener, C. D. 1970 – Diverse approaches to systematics. Pp. 1-38 *in*: Evolutionary Biology, Vol. 4. T. Dobzhansky, M. K. Hecht and W. C. Steere, editors. Appleton-Century-Crofts, New York. ix + 312 pages.

Prance, G. T., D. J. Rogers and F. White 1969 – A taximetric study of an angiosperm family: generic delimitation in the Chrysobalanaceae. New Phytol. 68: 1203-1234.

Rogers, D. J. and S. G. Appan 1969 – Taximetric methods for delimiting biological species. Taxon 18 (6): 608-624.

Rogers, D. J. and H. S. Fleming 1972 – A monograph of *Manihot esculenta* Crantz with an explanation of the taximetric methods used. Economic Botany (in press).

Shannon, C. E. 1948 – A mathematical theory of communications. Bell System Technical Journal 27: 379-423, 623-656.

Sneath, P. H. A. 1956 – Some thoughts on bacterial classification. J. Gen. Microbiol. 17: 184-200.

Sokal, R. R. and F. J. Rohlf 1970 – The intelligent ignoramus, an experiment in numerical taxonomy. Taxon 19 (3): 305-319.

Sokal, R. R. and P. H. A. Sneath 1963 – Principles of numerical taxonomy. W. H. Freeman and Co., San Francisco. 359 pages.

Wirth, M., G. F. Estabrook and D. J. Rogers 1966 – A graph theory model for systematic biology, with an example for the Oncidiinae (Orchidaceae). Systematic Zoology 15 (1): 59-69.