

Quantitative Characters in Computer Taxonomy

W. BRYCE KENDRICK

(Plant Research Institute, Canada Department of Agriculture, Ottawa, Canada)

ALTHOUGH a man can visually appreciate a very complex concept almost instantaneously, his much more limited capacity for verbal communication forces him to describe what he sees in a series of words, some of which convey more information than others. The taxonomist often finds himself in an analogous position. He may be able to assimilate the 'facies' of an organism at a glance, but in order to interpret to others what he sees, he must mentally dissect the organism and describe it as a series of characters, some of which may have greater significance than others (Kendrick and Proctor, 1964).

This enforced analysis of an organism is both a limitation and an invaluable feature of taxonomy. On the one hand, there are some subjectively appreciated features of organisms which, though they may be of diagnostic importance, are difficult to circumscribe for the benefit of other scientists. On the other hand, many taxonomically important differences between organisms such as fungi lend themselves to description, not in such vague terms as 'bigger than x' or 'narrower than y', but as more or less precise measurements. It is with such quantitative differences that this paper is concerned.

A character may most broadly be defined as: any attribute of an organism that can be detected and described. In taximetrics, a more specific definition is required, and for most of the numerical work carried out so far, such a definition may be as follows: any attribute referring to form, structure, or behaviour, which can occur in any one organism as one of two or more mutually exclusive states. The chief limitations to the amount of information about a given organism which can be collected are, theoretically, those in man's knowledge and insight, but practically, those in his patience; any living organism, no matter how apparently simple, possesses a very large number of characters.

The characters of any organism can be divided into two main types; qualitative and quantitative. Qualitative characters are those in which differences do not lend themselves to numerical expression; they may be straightforward presence or absence features, or multistate characters, as for example leaf shape. A quantitative character may be defined as: any character which varies from one organism to another in

a way which can be counted or measured and expressed numerically; for example, number of leaflets, weight of fruit, length of leaf, thickness of stalk. One of the weak spots in taximetrics is the subdivision of such characters into states, and their subsequent encoding for the computer. When dealing with quantitative characters, it is difficult to devise methods that are both objective and logical. The taximetrician has, however, a few guide lines to follow. One useful way of treating any group of organisms for a potentially multistate quantitative character is to ensure that each of the states prescribed contains the expression of that character in at least one organism. This may most easily be explained by a simple example. Consider a group of three organisms in which a quantitative character has values of 0.5, 3.5, and 5.5 units respectively. It is theoretically possible to divide this character into states in an infinite number of ways. For practical purposes, however, this number is quite limited. The character may be allotted two states, each covering a range of three units:

0—3	3—6
0.5	3.5 5.5

In this case the first organism is distinguished from the other two, but these are not distinguished from each other. If such a distinction appears desirable or necessary, three states may be allotted:

0—2	2—4	4—6
0.5	3.5	5.5

It is also possible to allot six states to this character:

0—1	1—2	2—3	3—4	4—5	5—6
0.5			3.5		5.5

but three of these states are vacant and therefore redundant, and while additions to the group might fall into the empty states, it is impossible on the evidence provided by the three known organisms to predict whether this further subdivision would be of any practical value. Accordingly in this instance the second method should be used. Objectively applied, the approach explained above will preclude unnecessary subdivision of characters.

In the example given above, the steps into which the total range of the character has been divided are of equal magnitude, that is, are arrived at by an arithmetic progression in which successive increments are equal. This may be satisfactory when the significance of variation at opposite ends of the range is not known. However, under certain circumstances, a small amount of variation at the lower end of the range may be just as significant as a much larger variation at the upper end of the range, for example when one is comparing dwarf and giant varieties of one taxon, or taxa of different orders of size, such as herbs

and trees. To deal with these eventualities a sliding scale is needed in which the steps increase in size toward the upper limit of the range of the character. At least two methods are available. The first is a progression in which successive increments are themselves incremented by a fixed amount; in the following example, each step is larger than the previous one by a value of unity. 0-1, 1-3, 3-6, 6-10, 10-15, 15-21. The second alternative is a geometric progression, in which successive increments are multiplied by a fixed number; in the following example, each step is larger than the previous one by a factor of two. 0-1, 1-3, 3-7, 7-15, 15-31, 31-63. The techniques of taximetrics are still largely empirical, and at present the individual investigator is free to select, from the three types of progression given above, that which seems best suited to his material.

Another suggestion for the subdivision of characters is that the number of states allotted to any character should not normally exceed six. An important reason for imposing this limit is the existence in all living organisms of natural variability, which would tend to obscure finer subdivisions. Allowing the recognition of even six states is sometimes a little dangerous, because in the case of a character expressed in the highest state, a genetically or environmentally induced reduction in magnitude of 16% will be sufficient to cause the character to be expressed in the next lower state.

Most scoring techniques in current use score quantitative characters 'non-additively', that is, a complete character difference is scored both when a pair of organisms have their respective expressions of a character in adjacent states, and when these expressions lie at opposite ends of the total range of states. This is obviously an unsatisfactory state of affairs. It is clearly desirable that some acceptable means of expressing the magnitude of differences in quantitative characters should be available, particularly in cases where large numbers of taxonomically important quantitative characters are present.

Sneath (1962) suggests 'additive' scoring as one way of doing this. His proposal involves scoring a six-state quantitative character as in Table 1. While this system does make allowances for magnitude of quantitative differences, it also introduces an anomaly sufficiently serious to prevent its general adoption. A consideration of Table 1 will show that what was originally a single unweighted qualitative character can now have a score of six, whereas no unweighted qualitative character can have a score higher than unity. This would place a completely unnatural emphasis on quantitative characters. It is also true of this method that if similarity coefficients are used, then the coefficient between two large organisms, that is, organisms which express many quantitative characters in their higher states, will be higher than that between two smaller, but otherwise equally similar organisms. If matching coefficients are used, this second objection does not arise.

Kendrick and Proctor (1964) suggest that it might be better to give fractional values to differences in quantitative characters, allowing a maximum possible score of unity, but because of practical difficulties

associated with the introduction of fractional scoring, they do not test the proposal.

TABLE 1
Additive scoring of a six-state quantitative character

Character expressed in state	State					
	1	2	3	4	5	6
1	+	-	-	-	-	-
2	+	+	-	-	-	-
3	+	+	+	-	-	-
4	+	+	+	+	-	-
5	+	+	+	+	+	-
6	+	+	+	+	+	+

It has subsequently been realised that the use of fractions and the probable attendant programming difficulties could in fact be avoided by the use of the number 60. The score for any unweighted character, qualitative or quantitative, previously unity, is now set at 60. Quantitative characters can score a maximum of 60 if their expression lies in the highest state. If they are expressed in states other than the highest, they score proportionately less, depending on the number of states recognised. 60 is divisible by 2, 3, 4, 5 and 6, and it is therefore possible to use units of 30, 20, 15, 12 and 10, to represent the states of quantitative characters having 2, 3, 4, 5, and 6 states respectively. It is of interest to note here that if only four states were necessary in an analysis, then 12 could be substituted for 60 as the base number, because it is divisible by 2, 3, and 4. The fractions of 60 outlined above are scored additively, and Table 2 summarizes the basic concept of what may be termed 'additive fractional' scoring.

There are at least three ways in which comparisons between organisms can be made using additive fractional scoring. These may be explained by reference to three organisms, A, B, and C, and their expressions of a quantitative character known to be divided into six states. A possesses the character in its first or lowest state, B in its second state, and C in its fifth state.

	1	2	3	4	5	6
A	+	-	-	-	-	-
B	+	+	-	-	-	-
C	+	+	+	+	+	-

Method 1. Only positive similarities are recognized, and similarity coefficients are derived. The portion of the coefficient derived from other features is represented in each case as m/n .

	1	2	3	4	5	6
A	10	—	—	—	—	—
B	10	10	—	—	—	—
C	10	10	10	10	10	—

$$S_{AB} = \frac{10 + m}{20 + n}$$

$$S_{AC} = \frac{10 + m}{50 + n}$$

$$S_{BC} = \frac{20 + m}{50 + n}$$

Method 2. Both positive and negative similarities are allowed, and matching coefficients are derived.

$$M_{AB} = \frac{50 + m}{60 + n}$$

$$M_{AC} = \frac{20 + m}{60 + n}$$

$$M_{BC} = \frac{30 + m}{60 + n}$$

Method 3. If the members of a pair have different expressions of the character, the organism with the higher expression is allotted a score of 60, the score of the other organism being adjusted in proportion. If both organisms express the character in the same state, both score 60. This concept may be readily understood by reference to the diagrams below.

	1	2
A	30	—
B	30	30

	1	2	3	4	5
A	12	—	—	—	—
C	12	12	12	12	12

	1	2	3	4	5
B	12	12	—	—	—
C	12	12	12	12	12

$$M_{AB} = \frac{30 + m}{60 + n}$$

$$M_{AC} = \frac{12 + m}{60 + n}$$

$$M_{BC} = \frac{24 + m}{60 + n}$$

One of the objections to the earlier additive scheme proposed by Sneath (1962), as has already been noted, is that two large organisms would score a higher similarity than two smaller, but otherwise equally similar, organisms. Although reduced to a fractional difference, this difficulty remains in method 1 above. A second anomaly in method 1 is that characters absent from both organisms of any pair are not taken into consideration when the similarity coefficient is being formed. The variable numbers of such negative matches create an undesirable instability in the total number of characters on which a comparison is made. Methods 2 and 3 are ways of removing the inequality introduced by differences in size, and at the same time ensuring that the total number of characters on which a comparison is made remains relatively constant from one pair of organisms to the next. While neither of these methods is perfect, both have some theoretical advantages. Method

3 is logical in that every comparison of two organisms for a given character uses the maximum expression of the character found in that pair as its frame of reference; in effect the two organisms are treated as if no other organisms existed. Method 2 compares organisms in terms of the total expression of each character found in the group under consideration. The relative merits of these methods are open to discussion; however, method 2 is compatible with present scoring techniques, and allows a straightforward extension of the matching principle to cover additive fractional scoring.

With the adoption of this technique, the *states* of quantitative characters can be regarded as *subcharacters*, to distinguish them from the mutually exclusive states of qualitative characters. These subcharacters can be treated by the computer in the same manner as qualitative characters. The main points differentiating characters from subcharacters are as follows: first, unweighted characters have a score of unity (expressed as 60 for the purposes of the new scoring method), while individual subcharacters have fractional scores (expressed as values between 10 and 30 for the purposes of the new scoring method); and second, each character is usually examined before inclusion in the analysis to ensure that its variation is not too strongly correlated with that of any other character already included, while the subcharacters of any given character are completely dependent from right to left — for example, subcharacter 6 of a six-state character cannot be present unless subcharacters 1-5 are already present.

It is now apparent that the definition of a character given earlier does not cover analyses in which additive scoring is used. An amended definition might be as follows: any attribute of an organism referring to form, structure, or behaviour, which can be present in any one organism as one of two or more mutually exclusive states, or as one or more mutually dependent subcharacters.

Kendrick and Proctor (1964) discuss the application of taximetric procedures to the Fungi Imperfecti and find that the most logically satisfying methods are those in which 'primary' characters are weighted according to the number of 'secondary' characters used to describe them, and a matching coefficient is derived after all redundancies have been removed. However, integer non-additive scoring was used for the quantitative characters, and despite the generally satisfactory nature of the classification derived from these methods, certain rather anomalous groupings remained. Part of the dendrogram derived from that study is reproduced here as Fig. 1. Five species of the hyphomycete genus *Verticicladiella* are represented as follows: isolate 1, *V. penicillata*; isolates 8-13, *V. abietina*; isolates 14-20, *V. procera*; isolate 21, *V. wagnerii*; isolate 22, *V. serpens*. As a taxonomist assessing similarity intuitively, I consider *V. abietina* and *V. procera* to have more affinity with one another than with any of the other species shown. From Fig. 1, however, it would appear that *V. procera* is most similar to *V. wagnerii*, and *V. abietina* most similar to *V. penicillata* and *V. serpens*. In addition, *V. abietina* apparently contains two groups, each of three isolates, which

appear to be linked with each other at almost the same level as that at which *V. penicillata* and *V. serpens* are separated, that is, at the species level. The split in *V. abietina* appears to be engendered by dimensional and host differences — the various parts of the conidiophore in isolates 11-13 are consistently larger than those of isolates 8-10. It may be noted that the 8-10 group were isolated about five years before the 11-13 group. As conidiophores in this genus often degenerate with increasing age of the culture, this may explain the phenetic differences. Alternatively, isolates 8-10 may constitute a race peculiar to their shared host genus, *Picea*. Whatever gave rise to the differences existing between

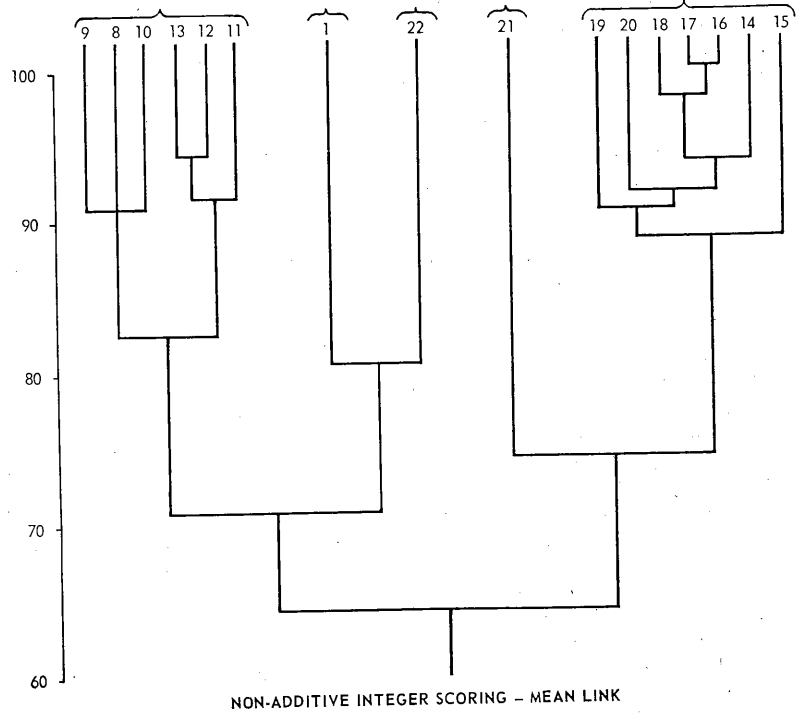


FIG. 1.

the various isolates of *V. abietina*, I do not believe that these differences constitute sufficient reason for the erection of two species rather than one. On the other hand, *V. penicillata* (isolate 1) and *V. serpens* (isolate 22), while they may be critical species, do appear from the available evidence to represent distinct taxa.

Having examined the various putative anomalies produced by the weighted method with matching coefficients and integer non-additive

scoring of quantitative characters, it remains to compare this method with the newly proposed modification. A dendrogram derived from the second method of additive fractional scoring is reproduced as Fig. 2 for direct comparison with Fig. 1*. It may be seen that, with the new methods, *V. abietina* (isolates 8-13) is no longer linked primarily with *V. penicillata* and *V. serpens* (isolates 1 and 22), and that *V. wagnerii* (isolate 21) is now the only organism to be interposed between *V.*

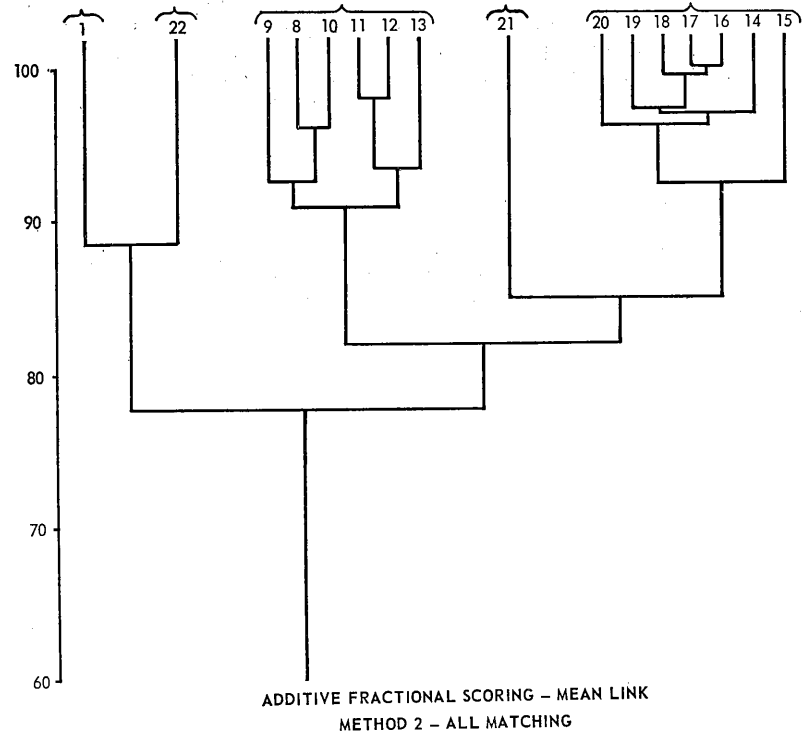


FIG. 2.

abietina and *V. procera*. It is also noticeable that while the *V. abietina* isolates are still divided into two groups, these groups are much less distinct than before.

It appears then, that the new methods effect an improvement in the

* These dendrograms were constructed by a programme written by Mr. A. Bickle of the Statistical Research Service, Research Branch, Canada Agriculture, Ottawa. The programme employs the mean of coefficients between groups to determine linkage levels. Details of the programme will appear in *Taxometrics*.

arrangement of the species, at least as seen through the eyes of the 'intuitive' taxonomist. This practical demonstration helps to confirm that the theoretical advantages of additive fractional scoring discussed earlier in this paper are of real value. Thus both logical and intuitive approaches suggest that this method of scoring is a worthwhile refinement when large numbers of quantitative characters are encountered.

ACKNOWLEDGMENTS

I am indebted to Dr. L. K. Weresub and Miss J. Proctor for invaluable discussions during the preparation of the manuscript. I would also like to thank Miss Proctor and the staff of the Statistical Research Service, Canada Department of Agriculture, for their unfailing cooperation, and Dr. P. Robinson and Dr. M. K. Nobles for many helpful criticisms.

REFERENCES

- KENDRICK, W. B., and PROCTOR, J., 1964. Computer taxonomy in the Fungi Imperfecti. *Can. J. Bot.*, 42: 65-88.
- SNEATH, P. H. A., 1962. The construction of taxonomic groups. In: G. C. Ainsworth and P. H. A. Sneath (eds.), *Microbial Classification (XII Symp. Soc. gen. Microbiol.)*, 289-332. Cambridge University Press, Cambridge.