

COMPLEXITY AND DEPENDENCE IN COMPUTER TAXONOMY *

W. B. Kendrick (Ottawa)

No practising taxonomist seriously believes that all the characters he records are of equal importance. Yet for several years the development of taximetrics proceeded with the working hypothesis that an Adansonian approach, giving equal weight to all characters of an organism, was quite adequate. The validity of this hypothesis has recently been challenged (Kendrick and Proctor 1964, Lockhart 1964) on both logical and intuitive grounds, and although no fully satisfactory method of representing relative importance for the computer has yet been devised, taximetrists should no longer regard equal weighting as axiomatic.

How can relative importance be assessed? Two possible indices of importance are *conservatism* and *complexity*. Although conservatism, often expressed in embryonic and reproductive characters, is of extreme importance in differentiating taxa of very high rank, and is almost always taken into consideration by practising taxonomists, I know of no acceptable and objective method of expressing this quality numerically. Complexity and its corollary dependence, however, are currently the objects of considerable study.

A complex character usually requires amplified description. It is not enough to say 'leaves present'. Those leaves are complicated structures and have a high potential of taxonomically useful information, which is not realized until the leaf is characterized by a number of properties such as shape, length, color, hairiness, etc. This introduces the concept of dependence. The hairs on the surface cannot be present if the leaf is absent.

The whole problem of dependence is in need of exploration in depth. Lubischew (1963) suggests that it may be possible to select a few basic characters from among the enormous number of available characters. These few may be considered as independent variables, all others as dependent ones, with, of course, varying degrees of dependence. The Arthropoda constitute a very instructive example of this principle. All members of the phylum possess in common the following characters: (1) the chitin-protein sclerotized cuticle, (2) the articulated appendages, (3) the moult or ecdysis mechanism, (4) the absence of circular muscles, (5) the absence of cilia, (6) the absence

* Contribution No. 431 from the Plant Research Institute, Canada Department of Agriculture, Ottawa, Canada.

of mucous membranes, (7) feeble capacity for regeneration of appendages. The chitinous cuticle provides the protection from water loss which is necessary for successful adaptation to the terrestrial habitat. The other six characters listed above are direct or indirect consequences of sclerotization.

The chitin-protein sclerotized cuticle is certainly one of the basic characters of the Arthropoda. But chitin is not restricted to the Arthropods. It occurs also in the Porifera, Hydrozoa, Bryozoa, Brachiopoda, Mollusca, Annelida and the Fungi, having presumably arisen independently several times. However, we will not normally need a computer to tell us the difference between an arthropod and a member of another phylum. There will be a more or less automatic selection to ensure that all organisms in any one taximetric analysis share many basic characters. Any character found in all organisms of the group to be studied is automatically excluded from the analysis. This would mean that, in a comparison between arthropods, chitin, the articulated appendages, the moult, etc., would not be considered.

Another kind of dependence may be exemplified by a fungus. The conidia of the genus *Verticicladiella* (Kendrick 1962) are produced by a complex sporogenous apparatus. It may be contended that the conidia are dependent on the sporogenous cell, the sporogenous cell upon its subtending tertiary metula, the tertiary metula on the secondary metula, the secondary metula on the primary metula, the primary metula on the supporting stipe, and the stipe on the assimilative mycelium from which it arose. There is also a physical and nutritional dependence of the fungus on its substrate.

A further example of dependence may be found in the tricarboxylic acid cycle, commonly known as the Krebs cycle, in which a series of enzymes such as aconitase, isocitric dehydrogenase, oxalosuccinic decarboxylase, α -ketoglutaric oxidase, etc., must act sequentially. Each step in this aerobic conversion of pyruvate to carbon dioxide and water is normally dependent on the successful completion of the preceding steps and the consequent provision of the correct substrate. A chemically induced blockage of the chain at any point, or a genetic loss of any of the enzymes, may interrupt the entire cycle. Metabolic pathways like the Krebs cycle are gradually being explored, and our knowledge of dependence in molecular biology is increasing. It is easy to see that the ramifications of this type of dependence will eventually lead us to a consideration of the individual gene or nucleoprotein, a goal which we cannot yet attain in the majority of organisms.

We can now examine the picture in its broadest sense. Life is an extremely complex phenomenon, within which it is impossible in most cases to look upon dependence as a linear process as we did in the case of the fungal conidium, or even a circular one as in the Krebs cycle; rather, the whole story cannot be represented by anything less than a multidimensional reticulum. It is the task of the taxonomist-ecologist to find his way through this web of dependencies and emerge with discrete (?) units called taxa, or phenons, or what you will.

UNEQUAL WEIGHTING OF CHARACTERS

It is now appropriate for us to discuss the basic questions facing the taxonomist who would enlist the services of the impartial computer. How shall we assess dependence? How shall we express degrees and kinds of dependence? How much of all this shall we take into account in our scoring procedures?

The only way in which we can express differences in relative importance for the computer is by giving the characters proportionately more or less numerical 'weighting' according to our evaluation of them. Although a majority of bacteriologists look with disfavour on any form of overt weighting, I suspect that their rejection of the concept arises from the adequacy of equal weighting in bacterial taximetrics. Bacterial taxonomy

has long ceased to be a visual science, if indeed it ever was. Rather, it has contained the germ of Adansonian taximetrics from its infancy. Instead of a visual appreciation of similarity, it employs a battery of tests capable of objective application, although perhaps of rather varied interpretation. Who could decide whether one test was more important taxonomically than another? Although, with the elucidation of various metabolic pathways, the existence of dependent characters even at this level must be admitted, a majority of the diagnostic tests used by bacteriologists can still be treated as independent variables without apparently incurring any severe taxonomic penalties.

The taximetrically inclined entomologist has also managed quite well without the necessity of invoking any unequal weighting of characters, but for different reasons. The insects exhibit a wealth of characters, and a large number of features is readily available for inclusion in a taximetric analysis. This large sample in itself appears to contribute much to the validity and stability of the results. The amount of weighting which might be theoretically desirable may constitute only a small part of the total data, and its absence may not have any significant unfavourable effect on the classification derived from the analysis.

These are perfectly valid reasons for not introducing unequal weighting. As long as an acceptable taxonomy can be derived using the simpler Adansonian approach, unequal weighting is an unnecessary complication. In the fungi, however, equal weighting has been found to be inadequate (Proctor and Kendrick 1963). It has produced unrealistic and contradictory classifications. We discovered that this is due to the presence in the analysis of large numbers of dependent variables — characters which are dependent for their existence on other characters. For convenience in discussion, the dependent variables have been called secondary characters, and the features on which they depend have been termed primary characters. This is the simplest possible level of dependence, but its recognition and the subsequent development of a logical scoring technique to allow for its existence, have done much to render the application of taximetrics a practical proposition in some groups of organisms (Kendrick and Proctor 1964).

PRIMARY AND SECONDARY CHARACTERS

It is rather difficult to arrive at a really satisfactory definition of these terms. What is a primary character? Is it a character which does not depend on any other character? Or is it just a character which needs description by other characters? The latter definition seems to be ruled out by the already admitted existence of further levels of dependence which Kendrick and Proctor did not find it necessary to consider in their published study.

Perhaps a primary character might be defined as a character which often needs description in the form of subsidiary characters, but is not itself descriptive of any higher category of character in the group of organisms under consideration. A primary character is always a qualitative one, and in organisms whose taxonomy is morphologically based, it is usually an organ — conidiophore, spore, flower, leaf — with a definite function — dispersal, reproduction, photosynthesis, etc. The subsidiary status of secondary characters — metula, echinulation, petal, midrib — is easily detected. Some terms such as 'cell' and 'wall' are at first sight difficult to deal with, but while the cell is a basic character, it will normally be present in all the organisms under consideration, and can be disregarded as redundant in the taximetric sense. In the higher fungi one cannot have a spore without a wall, but walls are not restricted to spores. It appears that 'wall', 'cytoplasm', 'nucleus', etc., are secondary characters of the cell, but they too are universal phenomena in most groups, and their presence does not require consideration by the taximetrist. We must not forget that what we are

attempting to do is to provide the computer with an accurate assessment of the relative taxonomic importance of a reasonably large and representative selection of characters for each organism examined. We are not attempting to give a complete exposition of the whole multidimensional reticulum of interdependence.

Kendrick and Proctor discussed in detail only primary and secondary characters. They found that the importance of primary characters, and the subsidiary nature of secondary characters, could be quite satisfactorily represented by assigning to each primary character, in addition to its own score of unity, a weighting equal to the number of secondary characters used to describe it. This weighting can be justified in several ways, at least one of which should convince any individual reader.

(1) Its use in the groups of fungi analyzed has resulted in a sensible classification largely in agreement with that derived by the orthodox approach, whereas an unweighted analysis produced numerous anomalies and an unreasonable taxonomy.

(2) The weighting of primary characters removes a logical contradiction. Differences in secondary characters will naturally lower the similarity or matching coefficient, and since secondary characters may considerably outnumber primary characters, such differences can, in an unweighted analysis, outweigh similarities in primary characters. Where a taxon lacks a given primary character, the associated secondary characters are not considered when making comparisons with other taxa, and hence do not affect the coefficients. It is unsatisfactory that, in an unweighted analysis, it is possible for two taxa both possessing a given primary character, to appear less similar to each other than to a third taxon lacking the primary character entirely, because of differences in secondary features of that character.

(3) Once the existence of dependent characters has been recognized, there are three possible ways of dealing with them. (a) Omit secondary characters from the analysis altogether. This has in fact been suggested by Beers and Lockhart (1962), and may be feasible in analyses of organisms in which few dependent characters have been recognized (bacteria?), or in organisms with very large numbers of available characters (bees?). However, there are many groups such as the fungi and flowering plants in which dependent characters make up a large fraction of the total characters available and cannot readily be ignored. (b) Give secondary characters the same weight as primary characters — the Adansonian approach. Surely this is giving the secondary characters too much importance. (c) Give primary features more weight than secondary characters. This seems to be a reasonable compromise, particularly if the importance of secondary characters relative to primary characters is made self-governing, so that the secondary characters of any given primary character can under no circumstances outweigh that primary character.

METHODS OF NOTATION

In the convention proposed by Sneath (1957) and adopted in most subsequent taximetric studies, each state of every character is represented by one of the following symbols: '+' for present, '-' for absent, and 'NC' for no comparison. Whatever the symbols used, each is represented within the computer by a distinctive digital code. Different computers operate on different systems, but the IBM 16-20, which is widely available, will serve as an example. It uses a 'binary coded decimal' system, otherwise known as the 'seven bit alphameric code'. In this code, all single numbers and letters are represented by seven machine positions which make up one 'column'. Various arrangements of the seven bits can represent any digit from 0 through 9, or any of the 26 letters of the alphabet. The machine codes for some of the symbols used in this paper are given below.

Symbol	Code
0	0 00 1010
1	1 00 0001
2	1 00 0010
A	1 11 0001
B	1 11 0010
C	0 11 0011

Before any character can be included in an analysis, it must exist in at least two alternative states, the simplest case being that of presence or absence. With Sneath's notation adapted for the derivation of matching coefficients,¹⁾ these are scored as follows:

	States	
	presence of	absence of
	Q	Q
Organism with Q present scores	+	-
Organism with Q absent scores	NC	+

Consequently, two computer columns would be required for a two-state secondary character. A multistate secondary character would require one computer column for each state.

Lockhart and Hartman (1963) introduced a simple and space-saving notation which deserves to be widely adopted. They suggested the use of a different symbol, not just for '+', '-', and 'NC', but for every alternative state of a character. This means that a multistate character can be represented completely in only one computer column.

	Symbol
Organism with character R present in state 1 scores:	A
Organism with character R present in state 2 scores:	B
Organism with character R present in state 3 scores:	C
Organism with character R present in state 4 scores:	D

If character R can be absent, we merely add another symbol, E for absence.

¹⁾ A matching coefficient is derived when negative as well as positive matches are counted as similarities. This concept is particularly valid when characters, in order to merit inclusion in an analysis, must be present in at least one member of the group being studied. If flagella are found in some members of a group of bacteria but not in others, then their absence in both members of any given pair of isolates is a similarity worthy of consideration. Other advantages of a matching coefficient are as follows: the number of characters on which a comparison is made varies very little in any one analysis; and it becomes unnecessary to decide which aspect of a character such as "presence or absence of spore curvature" is positive, and which negative.

In Sneath's notation, the chief function of the 'No Comparison' symbol in the scoring of a character for which data were available was to prevent the scoring of more than one similarity or difference between any two organisms for that particular character. Under the new scheme, the 'No Comparison' symbol becomes redundant for that particular purpose, although it will still be needed to deal with cases where no information is available about a particular character in one or more of the organisms to be compared. In the example just given, 'No Comparison' (now perhaps better termed 'No Information') could be represented by F. In some computers the use of more than 4 symbols means that more internal storage space is required. However, if no more than 8 symbols are needed, a reasonable compromise between space-saving and flexibility is still possible. It is probable that 8 symbols would be adequate for most purposes, allotting six symbols to positive states, one symbol to a negative or absence state, and one symbol for 'No Information.' I have previously suggested (Kendrick 1964) that the total range of expression of a quantitative secondary character should not be divided into more than six states, because of the potential instability caused by environmentally induced variation in such characters. However, in the IBM 16-20 computer we could employ any of the 10 numerals or 26 letters of the alphabet without encroaching on additional machine space. This means that any character possessing any reasonable number of states can be represented in only one machine column.

ALTERNATIVE METHODS OF WEIGHTING

Unequal weighting was originally devised not merely to add weight to primary characters, but also to prevent undesirable weighting of secondary characters. Lockhart (1964) with this aspect in mind, has subsequently proposed a modification of the original scheme, using the alphabetic notation discussed above. This modification consisted in assigning to each secondary character of a given primary character an 'absence' symbol, which would be scored in place of the original 'no comparison' symbol whenever the primary character was absent. Thus two characters concerning 'leaf color' and 'leaf shape' would now both possess an extra symbol signifying 'leaves absent'. Now two organisms both lacking leaves will score a similarity, not only on the primary absence but also for every secondary character required to describe the leaves, and two organisms, one possessing leaves and the other lacking them, will score a difference not only for the primary character but also for each secondary character.

Both methods of weighting will give identical coefficients when one or both of the organisms of a given pair lack the primary character (Z in the example below). Where the primary character is present in both organisms (X and Y in the example below), both the numerator and the denominator of the coefficient derived by the original method will be greater than in the Lockhart method by a number equal to the number of secondary characters associated with that primary character. Another way of expressing the difference between the two methods of weighting is as follows: the Lockhart method does not weight a primary character except when it is present in only one member of any pair of organisms being compared; the original method weights primary characters all the time. In the example which follows, the alternative 'weighted' methods are compared with each other and with the 'unweighted' method.

Three organisms, X, Y, and Z are to be compared. X and Y both exhibit a primary character (leaves) which possesses six secondary characters (shape, length, thickness, color, surface, and venation). X and Y express all six secondary characters in different states. Z, lacking the primary character, naturally does not exhibit any of the secondary characters.

Character		Organism		
		X	Y	Z
1ary	leaves	present	present	absent
2ndy	shape	linear	rounded	-
2ndy	length	long	short	-
2ndy	thickness	thin	thick	-
2ndy	color	green	red	-
2ndy	surface	glabrous	pubescent	-
2ndy	venation	parallel	net	-

If we assume that X, Y and Z are identical in all (n) characters other than those associated with leaves, then under the original unweighted system:

Characters														
Organism	Leaf		Shape		Length		Thickness		Color		Surface		Venation	
	Pres.	Abs.	linear	rounded	long	short	thin	thick	green	red	glabr.	pubesc.	par.	net
X scores:	+	-	+	-	+	-	+	-	+	-	+	-	+	-
Y scores:	+	-	NC	+	NC	+	NC	+	NC	+	NC	+	NC	+
Z scores:	NC	+	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC

The computer is programmed to compare each pair of organisms and count the number of + signs in common (matches) and also the number of differences. It then derives a matching coefficient (M) by dividing the number of matches (n_m) by the total number of matches plus differences ($n_m + n_d$)

$$M_{xy} = \frac{n + 1}{n + 7} \quad M_{xz} = \frac{n + 0}{n + 1}$$

If $n = 10$, then

$$M_{xy} = \frac{11}{17} = 0.647 \quad M_{xz} = \frac{10}{11} = 0.909$$

Quite obviously $M_{xz} > M_{xy}$; that is, X and Z apparently resemble each other more closely than do X and Y.

Although this may be a somewhat unlikely example, it clearly illustrates the possible anomaly introduced by equal weighting of primary and secondary characters.

If the weighting scheme of Proctor and Kendrick is applied:

Organism	Leaf		Shape		Length		Tickn.		Color		Surface Venation	
	Present	Absent	1	2	1	2	1	2	1	2	1	2
X scores:	+++++++	-----	+	-	+	-	+	-	+	-	+	-
Y scores:	+++++++	-----	NC	+	NC	+	NC	+	NC	+	NC	+
Z scores:	NC	+++++++	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC

$$M_{xy} = \frac{n + 7}{n + 13} \quad M_{xz} = \frac{n + 0}{n + 7}$$

If $n = 10$, then

$$M_{xy} = \frac{17}{23} = 0.739 \quad M_{xz} = \frac{10}{17} = 0.588$$

$$M_{xy} > M_{xz}$$

The two organisms possessing leaves (X and Y) now reveal a greater similarity than either of them shows with the organism lacking leaves (Z).

Applying the Lockhart weighting scheme:

Organism	Leaf	Shape	Length	Thickness	Color	Surface	Venation
X scores:	A	C	C	C	C	C	C
Y scores:	A	D	D	D	D	D	D
Z scores:	B	B	B	B	B	B	B

Code: Present = A, Absent = B, State 1 = C, State 2 = D.

$$M_{xy} = \frac{n + 1}{n + 7} \quad M_{xz} = \frac{n + 0}{n + 7}$$

If $n = 10$, then

$$M_{xy} = \frac{11}{17} = 0.647 \quad M_{xz} = \frac{10}{17} = 0.588$$

$$M_{xy} > M_{xz}$$

It may be readily determined that the Lockhart scoring of leaf characters requires many fewer machine 'columns' than are required by other notation schemes in current use. Another advantage of Lockhart's approach is that the total score of similarities plus differences between all pairs of organisms is the same. However, it is reasonable to allow changes in the total score such as may occur in the method of Proctor and Kendrick in order to represent primary characters as consistently important in the taxonomic frame of reference. Hence both schemes have considerable validity. Considerations of economy in machine storage space make it desirable to adopt the Lockhart and Hartman alphabetic notation to express the positive weighting of primary characters as practised in the Proctor and Kendrick method of scoring. This is exemplified in the table which follows. The 'No Comparison' symbol now becomes necessary, and is here represented as E.

The example given above would now be scored in 13 columns:

Organism	Leaf	Shape	Length	Thickness	Color	Surface	Venation
X scores:	AAAAAAA	C	C	C	C	C	C
Y scores:	AAAAAAA	D	D	D	D	D	D
Z scores:	BBBBBBB	E	E	E	E	E	E

It is now possible to use the same computer program to compare the different approaches to weighting.

FURTHER LEVELS OF DEPENDENCE

Although only primary and secondary characters have been discussed so far, tertiary and quaternary characters may also be encountered in certain situations. A primary character may have secondary characters to describe it, and certain of the qualitative secondary characters may require tertiary features to describe them. For example, the conidia of *Menispora glauca*, *M. ciliata*, and *M. tortuosa* have, as secondary features, terminal setulae which have the tertiary features of length and position of insertion. As a general hypothesis, it may be suggested that the more complex a primary character is, the more levels of dependence may be required to characterize it.

Kendrick and Proctor have scored primary characters in such a way that when they are absent, their absence weighs heavily — in proportion to the number of secondary characters whose absence is a corollary of the absence of the primary character. It seems reasonable to extend this system to the relationship between secondary and tertiary characters. If a secondary character is absent, then the effect of its absence should be modified by the additional absence of whatever tertiary characters it may possess. Just as a primary character is weighted according to the number of secondary characters it possesses, so we should also weight a secondary character according to the number of tertiary characters it possesses. This procedure would also maintain the balance between the various hierarchies of characters, ensuring not only that differences in secondary characters cannot outweigh similarities in primary characters, but also that differences in tertiary characters cannot outweigh similarities in secondary characters.

The following example, based on an excerpt from the table of characters drawn up for an analysis of the hyphomycete genus *Menispora*, will illustrate this principle.

No.	Type	Description of Character	Symbol	Weighting
1	Primary	curved phialoconidia — present absent	A	9
			B	
2	Secondary	phialoconidia — aseptate 1-septate up to 3-septate	A	1
			B	
			C	
3	Secondary	phialoconidia — length < 10 μ 10-15 μ 15-20 μ 20-25 μ > 25 μ	A	1
			B	
			C	
			D	
			E	

No. Type	Description of Character	Symbol	Weighting	
4 Secondary	phialoconidia — width	< 2 μ	A	1
		2-3 μ	B	
		3-4 μ	C	
		> 4 μ	D	
5 Secondary	proximal end of phialoconidia —	rounded	A	1
		tapered	B	
		pointed	C	
6 Secondary	distal end of phialoconidia —	rounded	A	1
		tapered	B	
		pointed	C	
7 Secondary	setulae —	present	A	3
		absent	B	
8 Tertiary	setulae — length	< 2 μ	A	1
		2-4 μ	B	
		4-6 μ	C	
		6-8 μ	D	
		8-10 μ	E	
		> 10 μ	F	
9 Tertiary	setulae — insertion	both terminal	A	1
		1 terminal, 1 subterminal	B	
		both subterminal	C	

Three hypothetical species of *Menispora*, α , β , and γ , are similar in all respects except that α and β possess conidial setulae while γ lacks them, and that α and β differ in both subsidiary characters of the setulae.

The conidial features of these organisms might be scored as follows:

Character Number	Character Type	Organism		
		α	β	γ
1	Primary	AAAAAAAAA	AAAAAAAAA	AAAAAAAAA
2	Secondary	B	B	B
3	Secondary	B	B	B
4	Secondary	C	C	C
5	Secondary	A	A	A
6	Secondary	A	A	A
7	Secondary	AAA	AAA	BBB
8	Tertiary	A	F	—
9	Tertiary	B	C	—

$$Ma\beta = \frac{17}{19} = 0.895$$

$$Ma\gamma = \frac{14}{17} = 0.823$$

It is apparent that the differences in the tertiary characters of a secondary character cannot outweigh the similarity generated by the presence of that secondary character.

SCORING OF QUANTITATIVE CHARACTERS

The existence of quantitative characters creates further difficulties in the realistic presentation of data to the computer. A quantitative character may be defined as follows: any character which varies significantly from one organism to another in a way which can be counted or measured and expressed numerically. It is reasonable to suggest that all quantitative characters depend on qualitative ones. A spore must be present before it can have dimensions; petals must be present before they can be counted; fruits cannot be weighed if they are absent. In addition, quantitative characters do not themselves possess subsidiary characters — they are the end of a chain of dependence.

In the past most investigators have divided the total range of expression of a quantitative character into more or less equal steps or states, but in a reexamination of scoring techniques, Kendrick (1964) suggested that when organisms of greatly disparate sizes are being compared, it may be necessary to introduce some kind of arithmetic or geometric progression to determine the size of successive steps, if the number of such steps is not to proliferate beyond all reason. The most important reason for restricting the number of states recognized within the range of expression of a single character is the environmentally induced variation observable in all living organisms. This variation might cause the same organism to express a quantitative character in different states at different times, or cause individuals of the same taxon from different environments to fall into different states. This is most unsatisfactory, because under the scoring system generally employed at present, a complete character difference is scored, both when a pair of organisms have their respective expressions of a character in contiguous states and when these expressions lie at opposite ends of the total range of states. This arbitrary decision is required by the Adansonian approach, which insists on absolute uniformity in scoring, and is indifferent to the resultant inequity in expression. Only if the Adansonian approach is discarded does it become possible to express, to some extent, the magnitude of differences in quantitative characters, and to reduce the difficulties caused when organisms from the same taxon fall into adjacent states.

Sneath (1962) proposed an integer additive scoring technique which would make allowance for the magnitude of quantitative differences, but he was quite aware that the scheme had two major defects which would preclude its adoption. He suggested that a five state quantitative character could be scored as follows:

	States				
	1	2	3	4	5
Organism with character expressed in state 1 scores:	+	—	—	—	—
Organism with character expressed in state 2 scores:	+	+	—	—	—
Organism with character expressed in state 3 scores:	+	+	+	—	—
Organism with character expressed in state 4 scores:	+	+	+	+	—
Organism with character expressed in state 5 scores:	+	+	+	+	+

It is clear from this table that a quantitative character could, if expressed in its highest state, have a score several times greater than that of any unweighted qualitative character, whose score is still restricted to unity. This would be quite unrealistic. A further disadvantage of integer additive scoring is that when two large organisms (organisms expressing quantitative characters in their higher states) are compared, the similarity coefficient derived will be higher than that between two smaller but otherwise equally similar organisms.

Kendrick and Proctor (1964) discussed the use of fractional values to express differences in quantitative characters, allowing a maximum score of unity for the highest expression of any particular character, but because of programming difficulties attendant on the introduction of fractional scoring they were unable to test the scheme.

Kendrick (1964) has proposed a scheme which would circumvent the use of fractions but have the same effect as their introduction. If any unweighted character previously given a score of unity is now given a score of 60, fractional values can be stated in whole numbers, as from 1/2 to 1/6 of 60. Quantitative characters can score a maximum of 60 if expressed in the highest state. If their expression falls in other, lower, states, they score less, depending on how many states have originally be recognized. 60 can be divided by all numbers from 2 to 6, and units of 30, 20, 15, 12 and 10 can be used to represent the individual states of characters with 2, 3, 4, 5, and 6 states, respectively. The fractions of 60 are scored additively. The following examples will serve to illustrate the idea.

	State		
	1	2	3
Organism A with character present in state 1 scores:	20		
Organism B with character present in state 2 scores:	20	20	
Organism C with character present in state 3 scores:	20	20	20

$$S_{AB} = \frac{20}{40}$$

$$S_{AC} = \frac{20}{60}$$

$$S_{BC} = \frac{40}{60}$$

	State				
	1	2	3	4	5
Organism A with character present in state 1 scores:	12				
Organism B with character present in state 2 scores:	12	12			
Organism C with character present in state 3 scores:	12	12	12		
Organism D with character present in state 4 scores:	12	12	12	12	
Organism E with character present in state 5 scores:	12	12	12	12	12

$$S_{AC} = \frac{12}{36}$$

$$S_{AE} = \frac{12}{60}, \text{ etc.}$$

In a discussion of alternative ways of applying additive fractional scoring, Kendrick favoured the derivation of matching coefficients, in which negative matches (i.e. mutual absences) are counted as similarities.

	State					
	1	2	3	4	5	6
Organism A with character present in state 1 scores:	10	—	—	—	—	—
Organism B with character present in state 2 scores:	10	10	—	—	—	—
Organism C with character present in state 5 scores:	10	10	10	10	10	—

$$M_{AB} = \frac{50}{60}$$

$$M_{AC} = \frac{20}{60}$$

$$M_{BC} = \frac{30}{60}$$

Note that 'states' of quantitative characters are no longer mutually exclusive. It is more appropriate to term them 'subcharacters', and they may be treated by the machine in the same manner as qualitative characters. However, there are two main differences between subcharacters and characters. (1) Subcharacters frequently have 'fractional' scores. (2) Each character is scrutinized to ensure that its variation is not too closely correlated with that of any other character included in the analysis. Subcharacters, on the other hand, are obviously entirely correlated from right to left — the fourth subcharacter of a given quantitative character cannot be present unless the third subcharacter is already present, and the third cannot be present unless the first and second are.

*

In conclusion, it is hoped that this paper will perform two functions.

(1) To bring the problems raised by 'dependence', 'complexity', and 'weighting' to the notice of the Adansonian school of taximetrists and, I hope, generate a favourable climate among them for the discussion of such problems wherever they arise.

(2) To help convince orthodox taxonomists who have previously rejected the taximetric approach altogether because of its apparent rigidity, that it can be a flexible technique, and that it may yet approach its professed goal, the derivation of classifications based on the maximum amount of information.

ACKNOWLEDGMENTS

I wish to thank Dr. L. K. Weresub and Miss J. Proctor for their constructive criticisms.

References

- BEERS, R. J. and LOCKHART, W. R. 1962. — Experimental methods in computer taxonomy. *J. Gen. Microbiol.* 28: 633-640.
- KENDRICK, W. B. 1962. — The *Leptographium* complex. *Verticicladiella* Hughes. *Can. Journ. Botany* 40: 771-797.
- KENDRICK, W. B. 1964. — Quantitative characters in computer taxonomy. *In: Phenetic and phylogenetic classification*. Ed. Heywood and McNeill. (Syst. Assoc. Publ. 6) 105-114.
- KENDRICK, W. B. and PROCTOR, J. R. 1964. — Computer taxonomy in the fungi imperfecti. *Can. Journ. Botany* 42: 65-88.
- LOCKHART, W. R. 1964. — Scoring of data and group-formation in quantitative taxonomy. *In: Developments in Industrial Microbiology*, Vol. 5: 162-168. Washington, D.C.
- LOCKHART, W. R. and HARTMAN, P. A. 1963. — Formation of monothetic groups in quantitative bacterial taxonomy. *Journ. Bacteriol.*, 85: 68-77.
- LUBISCHEW, A. A. 1963. — On some contradictions in general taxonomy and evolution. *Evolution* 17: 414-430.

- PROCTOR, J. R. and KENDRICK, W. B. 1963. — Unequal weighting in numerical taxonomy. *Nature* 197: 716-717.
- SNEATH, P. H. A. 1957. — The application of computers to taxonomy. *Journ. Gen. Microbiol.* 17: 201-226.
- SNEATH, P. H. A. 1962. — The construction of taxonomic groups. *Microbial Classification* (12th Symp. Soc. Gen. Microbiol.): 289-332.
-