A METHOD OF SELECTION OF CHARACTERS IN NUMERICAL TAXONOMY

WALTER J. LE QUESNE

Abstract

A method of selecting from a data matrix the characters most likely to lead to valid phylogenetic conclusions is put forward, based on the concept of a *uniquely derived character* and its logical consequences. A *coefficient of character-state randomness* is defined which gives a measure of how far the distribution of character-states is random between the different characters. These principles are illustrated by application to published data on the seven species of *Argodrepana*.

The basis of conventional taxonomy has been the selection of characters by a subjective process in such a way as to produce the most self-consistent scheme of relationships between species. In numerical taxonomy, the normal approach is to use as wide a range of characters as possible and to try to choose these in an objective way with the purpose of removing the subjective element as far as practicable. The method proposed below was evolved with the plan of eliminating some out of the range of characters used, in a purely objective manner, in the hope of leading to relationships which were somewhat more likely to be phylogenetically valid. It rests, however, on somewhat different assumptions from the method proposed by Camin and Sokal (1965). Application to observed data can also lead to some interesting deductions concerning the speciation process.

The 'uniquely derived character' and logical consequences.—If one is studying the taxonomy of a group, a character that has evolved only in one direction on a single occasion in its history is more likely to give an unambiguous indication of its phylogeny: this concept will be defined as that of the *uniquely derived character*, using the word 'uniquely' in the same sense of occurring only once. Let us first suppose that all the characters have been reduced to a purely dichotomous form (with alternatives A and B). Now, if character 1 had the ancestral character-state 1_A which once during the development of the group altered to 1_B and if the independent character 2 originally represented by character-state 2_A also once altered to 2_B, not more than three out of the four possible combinations of these two character-states (*i.e.*, 1_A2_A, 1_A2_B, 1_B2_A and 1_B2_B) will be found. If, therefore, for any pair of dichotomous characters, all four combinations are found within the group, either character 1 or character 2 is not a uniquely derived character, or alternatively neither of them is. In practice, this will apply whether A or B is the ancestral character in each case.

It must be pointed out, however, that if three or less of the possible combinations are found, it does not necessarily prove that characters 1 and 2 are both uniquely derived characters, but only that they may possibly be. If, for example, character 2 changed from 2_A to 2_B on two or more occasions on each of which character 1 was in the same state, only three combinations will be found. Moreover, if the four combinations have been evolved during the history of the group, one of these may have died out again or alternatively not be represented in the material studied.

It may also be noted that if we consider two characters, both of which are expressed as both A and B character-states in the group of species considered, and find that only two out of the above-mentioned four possible combinations are found, these two characters are, in the group concerned, completely correlated and can be applied together in the formation of any key. All such relationships are thus well worth noting.

If we allow characters also to be ex-

pressed in a form where three possibilities are in a series (*e.g.*, a measurement which could be above, within or below the median range, represented by A, X and B respectively), it too can be seen that if all four possible combinations 1_A2_A, 1_A2_B, 1_B2_A and 1_B2_B of the extreme form occur, character 1 and character 2 cannot both be uniquely derived characters.

Application to a data matrix.—To apply this principle to a specific group, a data matrix is first prepared in the usual way, giving the character-state (expressed as A, B or X, as defined above) for each species and character (*i.e.*, 'test' in the sense of Sokal and Sneath) considered. Then each of the possible pairs of characters is considered to see how many of the possible combinations (excluding any terms including X) AA, AB, BA and BB occur among all the species. A matrix is drawn up with a space for the relationship between each pair of characters and a mark put on this in each case where all four possibilities have been found: this will be called the *character-pair matrix*.

At this stage, the number of marks corresponding to each character (in its combination with each of the other characters) is counted. Although it is not possible to say that any particular character cannot be a uniquely derived character, that with the greatest number of marks is the least likely to be so. Vertical and horizontal lines are drawn to eliminate the marks concerned and the number of remaining marks again counted in the case of all the other characters and the process repeated until all the marks have been eliminated.

Ideally, the characters left can be used to work out the phylogeny, at least in part. In cases where the selection of characters for elimination can be done in a number of ways that do not differ very widely in their probability, any phylogenetic deductions will be uncertain, but at least a figure can be given for the maximum number of characters which could be uniquely derived.

The coefficient of character-state randomness.—If none of the characters were uniquely derived, one would not expect to find all four combinations of character-states for every pair of characters. In fact, utilizing the number of species found in each character-state for each character, it is possible to calculate the probability of finding the four character-state combinations for each pair of characters, assuming purely random distribution of the character-states among the species. If the number of pairs of characters for which the four combinations of character-states are actually observed (in part or the whole of the character-pair matrix) is divided by the sum of the calculated probabilities on the above assumptions, one gets a ratio which can be called the *coefficient of character-state randomness*. This can most conveniently be expressed as a percentage, 0% representing entirely apparently uniquely derived characters and 100% representing no apparent correlation in the distribution of the character-states.

Application to data on Argodrepana.—The above principles can be simply illustrated by application to the data on *Argodrepana* (Lepidoptera, Drepanidae) quoted by Wilkinson (1967). This is a simple case, applied only to seven species, and more complex examples, such as that of *Teldenia*, will be discussed in separate publications. In the latter case, at least, the coefficient of character-state randomness is much higher than in *Argodrepana*, and consequently it is not easy to deduce the most probable phylogeny.

Only characters where both character-states A and B were represented by at least two species were used, since otherwise all four combinations cannot possibly be obtained. In order to simplify some of the subsequent calculations, it is convenient to define the less frequent of the character-states used in the analysis for each character as A and the more (or equally) so as B. To comply with this, the following substitutions were made in the symbolic representation in Wilkinson's paper:

Character 1: 1 becomes A and 2 becomes B

TABLE 1. DATA MATRIX FOR *Argodrepana* SPP., MODIFIED FROM DATA OF WILKINSON (1967).

				Species						
Character	n_A	n_B	n_X	*1*	*2*	*3*	*4*	*5*	*6*	*7*
9	2	5	0	B	B	B	B	B	A	A
12				B	B	B	B	B	A	A
13				A	A	B	B	B	B	B
17				B	B	B	B	B	A	A
30				A	A	B	B	B	B	B
34				B	B	B	B	B	A	A
36				A	A	B	B	B	B	B
44				B	B	B	A	A	B	B
45				B	B	B	A	A	B	B
46				B	B	B	B	B	A	A
58				B	B	B	B	A	B	A
71				A	A	B	B	B	B	B
80				A	A	B	B	B	B	B
83				B	B	B	A	A	B	B
11	2	4	1	B	B	X	B	B	A	A
68				B	B	B	A	B	A	X
76				A	A	X	B	B	B	B
2	2	3	2	A	A	X	X	B	B	B
3				A	A	X	X	B	B	B
25	2	2	3	B	X	A	X	A	B	X
1	3	4	0	B	A	A	B	A	B	B
47				B	B	A	B	A	A	B
75				B	B	A	A	A	B	B

Characters 2, 3; C becomes B
Characters 9, 12, 17, 46; C becomes A and A becomes B
Character 11; C becomes A and D becomes B
Characters 13, 47, 75, 80; 1 becomes A and O becomes B
Character 25; 150 becomes X (A = below, B = above)
Characters 30, 36, 44; B becomes A and O becomes B
Character 34; E becomes A
Character 45; C becomes A and O becomes B
Character 58; B becomes A and E becomes B
Character 68; 3, 2, 1 become A, X, B respectively
Character 71; 2 becomes A and 1 becomes B
Character 76; 4, 1, 2 become A, X, B respectively
Character 83; 2 becomes A and O or 1 becomes B

If n_A, n_B, n_X represent respectively the number of species for which the character-state is A, B or X (the latter including those for which it is undetermined), the characters are grouped according to values as shown in Table 1.

On comparing each pair of characters and marking with a cross where all four combinations of character-states occur, we obtain the character-pair matrix shown in Table 2. The number of marks for each character in combination with each of the others is given at the foot of the column, designated as N_x.

Characters 1, 47, 58 and 68 are successively deleted from this matrix by crossing out the vertical and horizontal lines corresponding to these four characters. This eliminates all marks. Thus, 19 out of the 23 characters suitable for this analysis behave as uniquely derived characters.

The following patterns of character-states are found among the remaining 19 characters, allowing an X to be replaced by either an A or a B to fit into the scheme.

Species							*Characters*
1	*2*	*3*	*4*	*5*	*6*	*7*	
A	A	B	B	B	B	B	2, 3, 13, 30, 35, 71, 76, 80
B	B	B	B	B	A	A	9, 11, 12, 17, 34, 46
B	B	B	A	A	B	B	44, 45, 83
B	B	A	A	A	B	B	25, 75

In the first three of these groups, we have examples of completely correlated characters. In the last case we have correlated characters 25 and 75 by regarding one species with wing-span 150mm as A and two as B, so that this relationship is of doubtful validity.

In every case of a uniquely derived character either all species with character-state A or all species with character-state B must be directly derived from a common ancestor. If all the above characters are actually uniquely derived, one can put forward several possible cladograms. The most probable of these can be derived on the assumption that all species with character-state A (a smaller group in each case than all with character-state B) are directly derived from

TABLE 2. CHARACTER-PAIR MATRIX BASED ON DATA FOR *Argodrepana*.

Character	1	2	3	9	11	12	13	17	25	30	34	36	44	45	46	47	58	68	71	75	76	80	83
1	–																						
2	x	–																					
3	x		–																				
9				–																			
11					–																		
12						–																	
13	x						–																
17								–															
25									–														
30	x									–													
34											–												
36	x											–											
44	x												–										
45	x													–									
46															–								
47	x			x	x	x		x			x		x	x	x	–							
58	x			x	x	x		x			x		x	x	x	x	–						
68													x	x		x		–					
71	x																		–				
75	x															x	x			–			
76	x																	x			–		
80	x															–						–	–
83	x					–							–			x	x	x					–
N_x	14	1	1	2	2	2	1	2	0	1	2	1	4	4	2	13	12	5	1	3	2	1	4

a common ancestor, from which one can deduce the following:

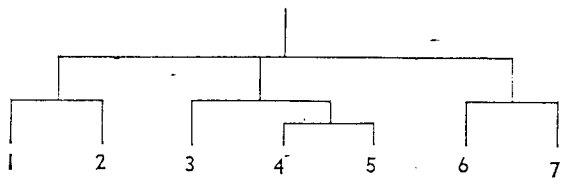
species 1 and 2 are directly derived from a common ancestor;

species 6 and 7 are directly derived from a common ancestor;

species 4 and 5 are directly derived from a common ancestor;

species 3 and the common ancestor of 4 and 5 are directly derived from a common ancestor.

We thus derive the following cladogram to show the phylogeny:



This diagram is not intended to be a phenogram or a phylogram.

Alternatively, the data left in Table 2 after elimination of the apparently unsuitable characters can be used to work out similarity coefficients between the species in the normal way and these converted into a "phenogram." In this case, I believe that the latter would necessarily display one of the cladistic relationships derived on logical grounds.

Coefficient of character-state randomness in Argodrepana.—For the calculation of the coefficient of character-state randomness (as defined above), we have to find the probability of all four character-state combinations being present in each of the fourteen possible combinations from pairs of values of n_A, n_B and n_x. The symbol N_1 represents the number of characters with the first pattern of values of n_A n_B and n_x and N_2 that with the second, N_{12} is the number of combinations of pairs consisting of one of each of these patterns. By multiplying the probability in each case by N_{12} and then adding the products, the expected value corresponding to a completely random

TABLE 3. CALCULATION OF PROBABILITIES OF ALL FOUR CHARACTER-STATE COMBINATIONS BEING PRESENT IN THE DIFFERENT CASES.

1) n_A	n_B	n_X	N_1	2) n_A	n_B	n_X	N_2	Prob.	N_{12}	Prob. × N_{12}
2	5	0	14	2	5	0	14	0.4762	91	43.33
				2	4	1	3	0.3810	42	16.00
				2	3	2	2	0.2857	28	8.00
				2	2	3	1	0.1905	14	2.67
				3	4	0	3	0.5714	42	24.00
2	4	1	3	2	4	1	3	0.3048	3	0.91
				2	3	2	2	0.2286	6	1.37
				2	2	3	1	0.1524	3	0.46
				3	4	0	3	0.5714	9	5.14
2	3	2	2	2	3	2	2	0.1429	1	0.14
				2	2	3	1	0.0558	2	0.11
				3	4	0	3	0.5143	6	3.09
2	2	3	1	3	4	0	3	0.3429	3	1.03
3	4	0	3	3	4	0	3	0.8571	3	2.57
							Totals		253	108.8

distribution of character-states is obtained, as shown in Table 3.

Using the whole of the above data in conjunction with the fact that the number of marked positions on the character-pair matrix is 40, the coefficient of character-state randomness is $40/108.8 \times 100\% = 36.8\%$.

As pointed out earlier in this paper, this figure indicates the extent to which the pattern of character-states is random rather than that to be expected were the characters uniquely derived. Here its relatively low value suggests that the pattern is considerably ordered.

Conclusions.—This method will enable an objective assessment of the probable phylogeny of certain groups and in all cases will give a numerical measure of the extent to which speciation can be regarded as being due to uniquely derived characters (or at least to characters which appear to be unique). Before we can make any generalizations about the results obtained, the method will have to be applied to a number of cases, preferably not all in closely related groups.

If a data matrix has been assembled for purposes of numerical taxonomy, this approach should not be difficult to apply. In cases where the numbers of species or characters is not too large (*e.g.*, 32 species and 30 to 40 characters or 13 species and 80 characters), the character-pair matrix can be worked out in a few hours without mechanical aids, but in large groups the task could be computerized.

ACKNOWLEDGMENTS

I wish to thank Dr. R. E. Blackith, Mr. J. C. Gower and Dr. C. Wilkinson for their valuable comments (*in litt.*) on the above-mentioned ideas.

REFERENCES

CAMIN, J. H. AND R. R. SOKAL. 1965. A method for deducing branching sequences in phylogeny. Evolution, 19:311–326.

WILKINSON, C. 1967. A taxonomic revision of the genus *Teldenia* Moore (Lepidoptera: Drepanidae, Drepaninae). Trans. R. Ent. Soc. Lond., 119:303–362.

70 Lye Green Road, Chesham, Bucks., England.