Zootaxa 5071 (1): 131–142
https://www.mapress.com/j/zt/

Article

ISSN 1175-5326 (print edition)
ZOOTAXA
ISSN 1175-5334 (online edition)

https://doi.org/10.11646/zootaxa.5071.1.7
http://zoobank.org/urn:lsid:zoobank.org:pub:BF5EB6D0-F92D-4F68-A4FF-DCEA65DF810E

Character correlation and its use for identification

JAMES E. HAYDEN
Florida Department of Agriculture and Consumer Services—Division of Plant Industry, 1911 SW 34th Street, Gainesville, Florida, 32608 USA. *james.hayden@fdacs.gov;* *https://orcid.org/0000-0003-1802-1064*

Abstract

A method is presented for correlating phylogenetic characters through cladistic analysis. It extends the use of phylogenetic datasets for diagnostic purposes. It improves matrix-based identification tools by predicting novel character-state combinations that were not observed when the key was constructed. By interpreting homoplasy as analytical error, hypothetical character-state combinations are tested for the homoplasy that they would add to the shortest tree(s). The correlation is equal to the homoplasy summed across all state combinations, divided by a maximum possible value. The results depend on uncertainty about the sequence of state transitions and their overlap among characters. A correlation index *r* is proposed for sets of non-additive characters; it is a kind of multiple-regression value, and its ensemble value *R* is a statistic of a whole matrix. This approach can be used to select sets of the best "proxy" characters to substitute for unobservable characters of interest. The concept can be extended to continuous characters. Worked examples are given with datasets of various insect orders.

Key words: continuous character, correlation, diagnostics, homoplasy, matrix key, multistate character, regression

Introduction

Identification keys in entomology face a twofold problem. The first problem is that, when identifying a specimen, very often it belongs to a species that was not considered when the key was constructed. The new species may exhibit a combination of character states that is not in the key, in which case it may lead to more than one couplet. New species should be anticipated and accommodated somehow. The second problem, which is hardly limited to entomology, is how to arrive at a suitably precise answer without having to examine excessively many characters.

The solution lies in taking information from phylogenies, because phylogenetic hypotheses provide generalizations about patterns in nature that artificial keys cannot provide. However, the solution requires a shift in focus: from narrowing down the choice to one species, to narrowing down the choice to one state of some character. The purposes of identification are various, but often the goal is not to arrive at a name but rather to infer some important characters that are unavailable or too difficult to observe. Instead, it would be desirable to have a set of ″proxy characters″that roughly corresponds to the unobserved character. Besides, in a world with so many undescribed species, narrowing down to one is impossible. Thus, the second problem is to find out how many characters, and which ones, are needed to infer another character's state to some arbitrary degree of precision.

Existing matrix-based keys have ways to reduce the choices to one species efficiently (Hall 1970, Dallwitz 1974). One of the most important ways is to choose characters whose states are distributed as evenly as possible among the remaining species or that divide the remaining ones in half (for example, the "Best" function in Lucid Player [Lucidcentral 2021]). The probability then is higher of reducing the remaining choices to a particular species by checking fewer characters on average. This approach implies that sequential characters should be minimally correlated: that is, taxa that share one state of the first character should have multiple states of the second character. This yields the greatest number of character-state *combinations*, by which species may be differentiated more readily. Characters that are strongly correlated in the usual statistical sense are not very useful, because they present fewer combinations of states.

In contrast, the new method reduces the choices to one state in an unobserved target character. To do so, it takes

Accepted by J. Morrone: 2 Nov. 2021; published: 22 Nov. 2021

the opposite approach by choosing characters that are most strongly correlated. The number of terminals that exhibit each state is irrelevant: only the number of observed and possible combinations are of interest. Given a few observed state-combinations, any additional ones that would be likely to come to light are the intermediary, heterobathmic combinations (Hennig 1965), and these have no homoplasy. One should expect to observe them in nature, even if as fossils. The question is not how "informative" a character is for a phylogeny, but how informative a character is of another character *through* a phylogeny.

The plausibility that a hypothetical character-state combination exists is quantified by how many extra steps it would add to the shortest cladogram(s). A combination that adds zero extra steps is most plausible, and it is counted the same as a combination that has been observed. For characters *x* and *y* with respectively *m* and *n* states, all possible *m*n* combinations of states are considered. If all of these combinations exist in the matrix, *x* and *y* have zero correlation a priori, and considering *x* and *y* together is useless. If only a few of the *m*n* state combinations exist in the matrix, then some of the unobserved ones could add more steps.

The practical way to implement this is, for any state combination, to search for a place on the shortest cladogram(s) where it could fit in with no extra steps (Fig. 1). A state combination may force multiple extra steps if the characters are ordered or if more than two characters are regressed: for example, with two three-state characters, $x = 0$ and $y = 2$ may force two steps if both are ordered 0-1-2 and five other combinations exist in the dataset. If one cannot do better than one extra step, that is at least better (more plausible) than another combination that requires two steps.

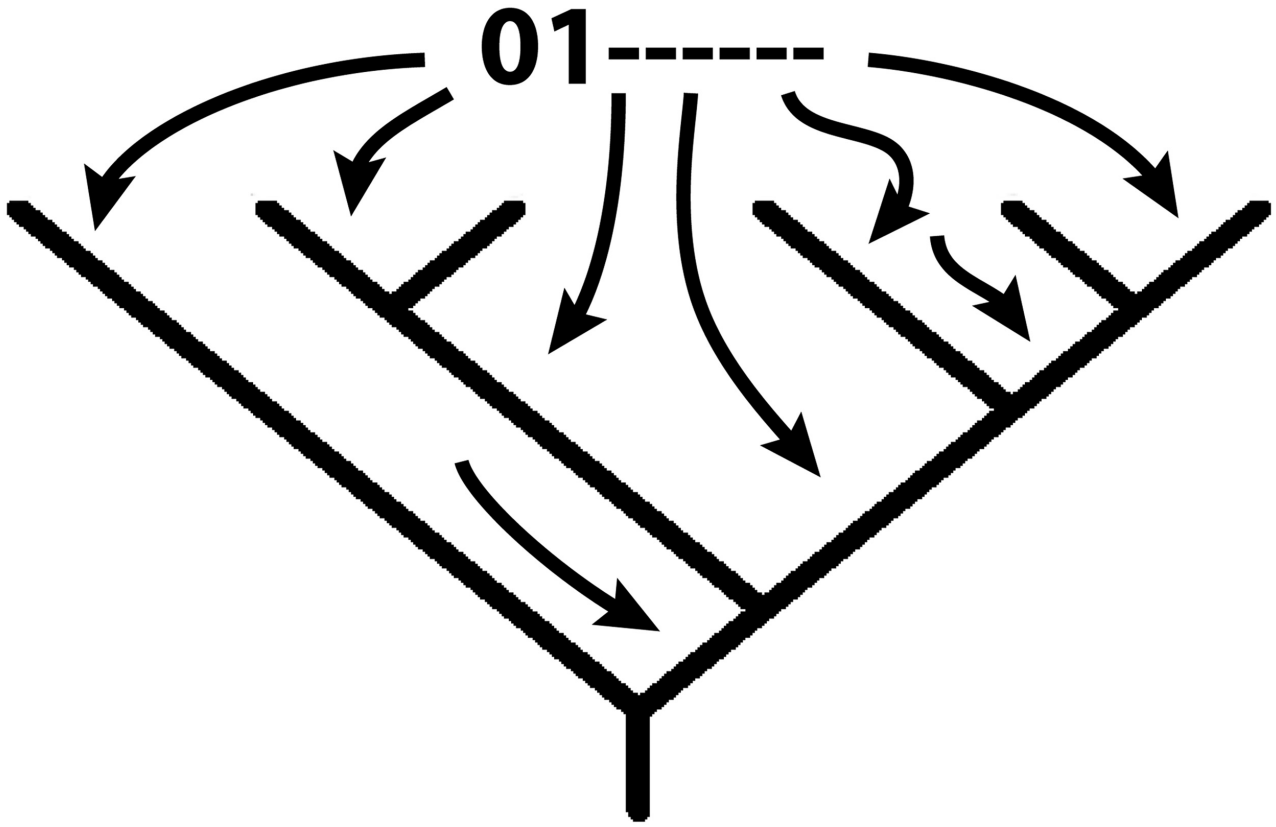


FIGURE 1. Cladogram being probed with a wildcard scored for the first two characters.

The algorithmic steps are as follows:

1) Search for shortest trees and save the trees and their length.
2) Activate a wildcard taxon *W* in the matrix that is unscored except for the characters in question (*x*, *y*, etc.).
3) Score *W* for a state combination, such as x = 0, y = 1.
4) Analyze again, constraining to the original tree.
5) If more than one shortest tree is found, loop through all trees, constraining to each in turn. Save only the shortest length among the trees.
6) Find the difference in length between the new trees (with the scored *W*) and the original tree(s).

7) Repeat steps 3–6, recoding *W* for every possible state combination (0,1; 0,2; 1,0; ...).
8) Store the tree-length differences for all state combinations in an array of *m* by *n* states (or more dimensions if more characters).

Thus, all of the shortest cladograms are probed with a wildcard terminal to find where it can fit in. A new heuristic search is made for every state combination. The process can be executed with TNT 1.5 (Goloboff and Catalano 2016) with the scripts *picktwo.run* through *pickfive.run* and *picktwocont.run* (available at <https://github.com/jeh8i8/TNT_scripts_character_correlation>), which execute the analysis and calculate the correlation statistic *r* that is explained below.

To be clear, "correlation" here has nothing to do with phylogenetically independent contrasts (Felsenstein, 1985), and it does not require estimated branch lengths. It does not address how frequently a set of characters co-evolves; it only addresses the plausibility of observing a character-state combination. The sense of "correlation" used here also differs from the "hierarchic correlation" of Farris (1969), which is a measure of a character's fit to the hierarchy, quantified by the consistency index.

A distinction must be made between the *general* correlation of characters, whereby *any* state of character *x* can be used to infer any other state of the correlated *y*, versus the approach of focusing specifically on one particular state of a target character. The results may yield different sets of best characters. Although the latter sense may find more popular practical use, the former general sense must be explained first because it is a kind of average that governs the results of the latter.

A simple example shows the difference that a phylogeny can make in a correlation. Consider two pairs of unordered characters: 0, 1 and 2, 3 (Table 1). They belong to a larger matrix with numerous characters. Both pairs would seem to be perfectly correlated on the face of it. Each pair exhibits the smallest possible number of combinations (five). If character 0 is state 0, then character 1 must also be 0; if 0 is 1, then 1 is 1; if character 2 is state 3, then 3 must also be 3. If they were quantitative variables and were correlated with standard statistical analysis, both pairs would have $r^2 = 1$. However, considering phylogenetic information might change that. Next, a wildcard is coded iteratively for all possible state-combinations for characters 0 and 1.

```
W   01 ?? ??? [other characters unscored]
W   02 ?? ???...
...
W   43 ?? ???...
```

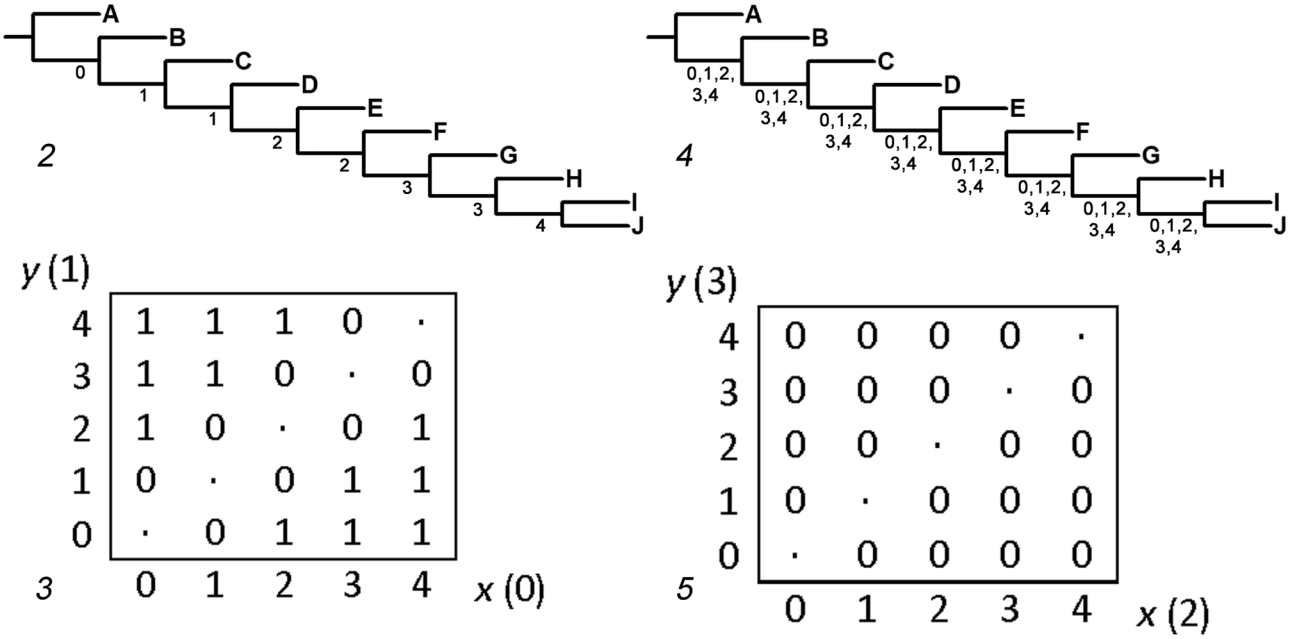
TABLE 1. Four characters of a larger matrix of many characters.

Char.	01	23	456
A	00	00	...
B	00	11	
C	11	22	
D	11	33	
E	22	44	
F	22	00	
G	33	11	
H	33	22	
I	44	33	
J	44	44	

The analysis of the whole matrix returns one pectinate tree (Fig. 2). Coding the wildcard for all the alternative state-combinations for 0 and 1 results in a correlation graph (Fig. 3), where "1" in the rows and columns indicates that the state-combination cannot fit anywhere without adding a step. In those cases, the wildcard must attach to the tree on at least two branches or sectors: one based on the evidence of character 0, and the other on the evidence of character 1.

On the other hand, characters 2 and 3 are optimized ambiguously, following standard Fitch optimization (Fitch 1971) (Fig. 4). Now any novel state combination can be accommodated with 0 steps (Fig. 5). Thus, the distribution of the states across the phylogeny can drastically change the plausibility of alternative combinations, even though the number of observed combinations is equally minimal. If character 2 is state 1, then character 3 could be anything from 0 to 4.

Alternatively, a reanalysis of the matrix could return a dichotomous tree (Fig. 6), as may result if other characters are reweighted or ordered. This results in totally ambiguous optimization of characters 0 and 1. It yields a correlation like Fig. 5 instead of Fig. 3, with all unobserved combinations having 0 steps.



FIGURES 2–5. Pectinate cladogram and unordered correlations. 2, pectinate tree, optimized for characters 0 and 1; 3, correlation of characters 0 (x) and 1 (y) on pectinate tree ("·": observed state combination; 0: zero extra steps needed; 1: one extra step needed); 4, pectinate tree with optimization for characters 2 and 3, showing ambiguity; 5, correlation of characters 2 (x) and 3 (y) on pectinate tree.

In these examples, poor correlation results from ambiguous optimization. It should be noted that correlation depends not on the number of state transformations in general but specifically on their inferred locations on the tree. Under some circumstances, it is possible to get good correlation in the presence of ambiguous optimization. Homoplasy also affects correlation, resulting in high or low correlation depending on the circumstances.

Indices for general correlation

An index is needed to find sets of highly correlated characters. This allows sets with different numbers of characters to be made comparable, as well as characters with different numbers of states and different assumptions about transformations. A simple index of correlation is defined:

(1) $r = s / m$

where *s* is the sum of the steps across all combinations, and *m* is the maximum possible number of steps. For two characters, the general equation for the maximum *m* is:

(2) $$m = \sum_{i=1}^{d} (G-i)*(L-i)$$

where G = the greater the number of states, L = the lesser number of states, and d = the lesser diameter of G or L. Eq. 2 is applicable to ordered characters. "Diameter" means the greatest number of steps between any two states (synonymous with "range," Farris, 1969). For two unordered characters, d = 1 and *m* simplifies to the difference between the maximum and minimum number of state combinations:

(3) $r = s / ((G-1)(L-1))$

Equation 2 can be illustrated by a correlation where two ordered characters are maximally correlated (Fig. 7). The observed combinations "·" are distributed as if *x* evolved through its states before *y* started changing. The particular distribution of homoplasy across the graph space does not matter for the total, but it is easier to illustrate this way.

Figure 8 shows a case where the characters are partially but not completely ordered (i.e. d < L-1). The maximum is the sum of (G-i)*(L-i) up to i = d = 2, so for Fig. 8, 5*4 + 4*3 = 23. The least diameter among the correlated characters is preferred, because any wildcard terminal will attach to the cladogram through that character state.

The general equation for multiple unordered characters (Equation 4) was found by induction from analyzed matrices.

(4) $m = (K-1)\prod_{i}^{K} n_i - \sum_{j}^{K}(\prod_{i}^{K} n_i/n_j) + 1$

K is the total number of characters, and n_i and n_j are the number of states in the i^{th} and j^{th} characters. For example, for four characters with 2, 3, 3, and 4 states, m = (4-1)*2*3*3*4 - 2*3*3 - 2*3*4 - 2*3*4 - 3*3*4 + 1 = 115.

For any set of characters, the observed number of steps is divided by the maximum possible correlation. Equation 4 is substituted for *m* in Eq. 1:

(5) $r = s / ((K-1)\prod_{i}^{K} n_i - \sum_{j}^{K}(\prod_{i}^{K} n_i/n_j) + 1)$

Equations 4 and 5 apply only to non-additive characters, but generalization to more than two ordered characters should be possible.

When more than two characters are correlated, it is necessary to constrain the results to the original trees. The hypothetical combinations coded in the wildcard should not be allowed to alter topologies based on real data.

Values of *r* converge as the number of characters per combination increases. The central value is the one for all characters of a matrix analyzed at the same time, called the *ensemble correlation*, *R*. This is a useful statistic to report for small datasets because it summarizes a general property of the dataset. *R* would be hard to calculate for larger datasets because of the large product sums. For a ten-character dataset of binary characters, *m* = 4097, and there are 1024 combinations to be tested for *s* steps in the numerator. For twenty binary characters, $m > 9.4 \times 10^6$, and more than a million combinations would have to be tested.

Continuous characters

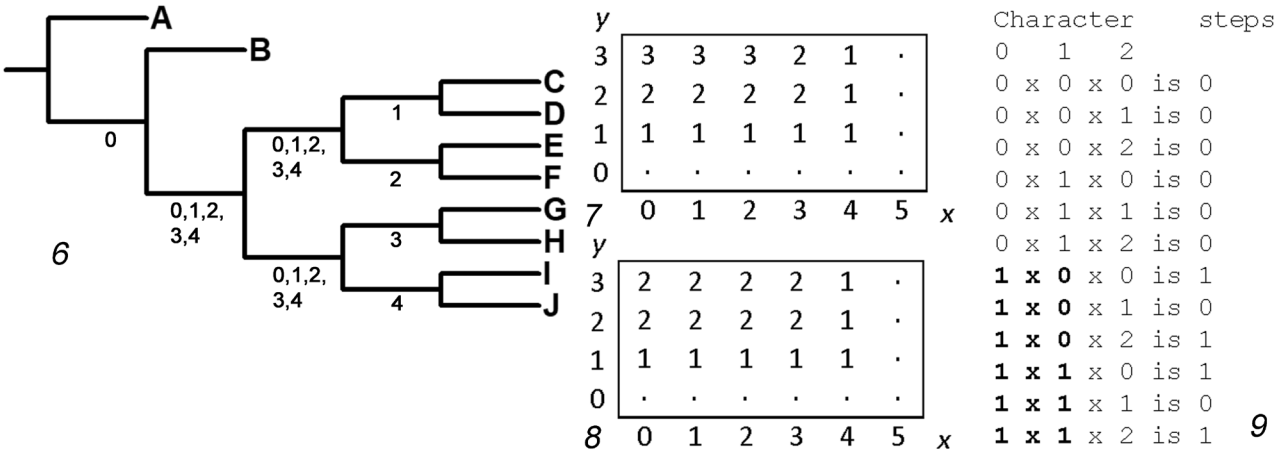
Ordering characters may change the correlation by altering the number of possible state-transformations. It can be extended to continuous characters (quantitative variables), but this requires three adjustments. The first is that the number of points to test in the Cartesian space is infinite. TNT 1.5 allows coding variables to three decimal points. A rough search may be done to save computing time, but non-zero homoplasy may appear at some points that are observed if, in fact, the observed value is very close. The second issue concerns the calculation of *m* and *r* when the increments are less than 1. Without an appropriate modification of Eq. 2, continuous characters can only be correlated for illustration, not for comparison. The third issue concerns subset polymorphisms. The problem is not restricted to continuous data, but it is most evident when a terminal has ranges of values. Nixon and Davis (1991) showed that polymorphic terminals should be split to obtain the correct length. Otherwise, some observed state combinations may show homoplasy at their points on the correlation graph. Ideally, terminals should be split to code the edges of the ranges of characters, and then constrained for monophyly.

Proxy characters for one state

In this situation, the goal is to find a set of characters that best predicts one particular state of a "target" character to the exclusion of other states. Calculating *r* with Eq. 5 is not necessary, but the same scripts are used to calculate the number of steps for each combination. The log output of *picktwo.run*, etc. is scanned for combinations that have 0 steps with that state and >0 steps with other states. Typical output is shown in Fig. 9, whereby state 1 of character 2 is the target. There are four combinations for characters 0 and 1, and two of them select 2:1. Thus, choosing to observe characters 0 and 1 would have a 50% chance of yielding a combination that could only occur with character 2, state 1. If the other combinations are discovered instead, they could just as plausibly occur with 2:0 or 2:2. Indeed, observing only character 0 is sufficient in this simple case, because 0:1 by itself would predict 2:1.

Worked examples

The first two examples give the *R* value for some (necessarily) small datasets. The next two show how to correlate a set of proxy characters with a target character, increasing from one to four characters until some degree of precision is achieved. The third example shows the graphic correlation of continuous characters under two weighting schemes.



FIGURES 6–9. Dichotomous cladogram, ordered correlations, and output. 6, dichotomous cladogram that results in zero correlation between characters 0 and 1; 7, one correlation graph showing the maximal possible correlation between two additive characters; 8, correlation with incomplete ordering; 9, output to determine which state-combinations are compatible only with state 1 of the third character (combinations that predict only 2:1 in bold).

1. Lees and Smith (1991) presented a 7x7 matrix of sunset moths (Lepidoptera: Uraniidae). Four characters are multistate, and 1536 combinations are possible. Analyzed with equal weights (script *rhoseven.run*), the single resulting cladogram was tested with all combinations. The total sum of steps $s = 2724$, the maximum number of steps $m = 5249$, and $R = 0.519$.

2. Carpenter (1988) presented a matrix with nine characters of social wasps (Hymenoptera: Vespidae). Two characters have three states and the rest are binary, so 1152 combinations are possible. Analyzed with equal weights, it results in two cladograms. Testing the combinations on both trees (script *rhonine.run*), $s = 2896$, $m = 4417$, and $R = 0.656$.

3. *Cryptophlebia-Ecdytolopha* group. Adamski and Brown (2001, their table 2) analyzed a dataset of genera in the *Ecdytolopha* group (Lepidoptera: Tortricidae), comprising eight terminals and 33 characters. Four of their six characters of larval morphology (#1, 2, 4, 5, or 2, 3, 5, 6 in their numbering) are chosen to correlate with character 30, the texture of the corpus bursae of the female genitalia. That character has three states, of which state 1 is a synapomorphy of *Thaumatotibia* Zacher and *Gymnandrosoma* Dyar, two economically important taxa of interest for regulatory diagnostics. The other characters are binary. The scripts *picktwo.run* through *pickfive.run* were used.

First, the correlation values of various combinations together with character 30 are shown in Table 2, second column. These combinations most successfully predict *any* state of the target character 30. Second, the combinations were determined that have 0 steps with 30:1 and >0 steps with 30:0 and 30:2. If any of these character-state combinations are found (see *combinations*), one can confidently predict that only 30:1 will be observed. The best combinations of proxy characters are those that have a *greater* percentage of combinations that select only 30:1.

TABLE 2. Sets of larval characters and their performance when correlated with character 30 in the *Ecdytolopha* dataset. Numbers under *"r"* are correlation values, with the most successful in bold. The fraction shows the number of combinations that predict only 30:1 over the total number of combinations. The particular combinations that correspond with 30:1 are in the fourth column. In italics are state combinations that are *not* observed with 30:1 in the matrix of Adamski and Brown (2001).

Character sets	*r*	Fraction with 30:1	Combinations
1	**1.00**	0/2	-
2	0.50	0/2	-
4	0.00	0/2	-
5	0.50	0/2	-
1,2	**0.67**	1/4	1:1, 2:0
1,4	0.56	0/4	-
1,5	**0.67**	1/4	1:1, 5:0
2,4	0.44	1/4	2:0, 4:1
2,5	0.44	1/4	*2:1, 5:0*
4,5	0.44	1/4	4:1, 5:0
1,2,4	**0.55**	1/8	1:1, 2:0, 4:1
1,2,5	**0.55**	**3/8**	1:1, 2:0, 5:0; *1:1, 2:0, 5:1;* 1:1, 2:1, 5:0
2,4,5	0.45	**3/8**	2:0, 4:1, 5:0; *2:0, 4:1, 5:1;* 2:1, 4:1, 5:0
1,2,4,5	0.53	3/16	1:1, 2:0, 4:1, 5:0; *1:1, 2:0, 4:1, 5:1; 1:1, 2:1, 4:1, 5:0*

To state it concretely: to predict the texture of the corpus bursae in general using a larval specimen, one should examine the abdominal D1 pinaculum (1), the spiracle of the eighth abdominal segment (2), the anal fork (5), and optionally the ventral setae of the ninth abdominal segment (4). If one observes the D1 pinaculum with a notch and the spiracle in the middle of the A8 segment, or possibly a notched D1 pinaculum, the A8 spiracle moved to the posterior, plus an anal fork, one can assume that the corpus bursae in an adult moth is punctate, not smooth nor spiculate. Characters 2, 4, and 5 present alternative possibilities.

4. *Xylomoia*. Mikkola (1998) revised the moth genus *Xylomoia* Staudinger (Lepidoptera: Noctuidae) with a cladistic analysis of thirteen terminals and 33 characters, including external morphology and internal genitalia. The result was two similar cladograms of 86 steps. The five apomorphies of *Xylomoia* are of the male genitalia, so dissection is necessary to positively identify a specimen as belonging to that genus. Therefore, a set of external characters is sought that correlates with one of these internal characters. Character 14 is a suitable target character because it has three states, and state 1 is a unique, unreversed synapomorphy of *Xylomoia*.

Of six characters of external morphology, a set of up to four proxy characters is sought to correlate with character 14. Characters 0 through 5, which are of external morphology, are tested. The *r* values for Character 14 and combinations of the external characters are shown in Table 3, together with the fraction of combinations that predict only 14:1. The results show that to predict *any* state of the costal sclerite (character 14), one should observe characters 0, 1, 2, and 4: the shape of the frons, the male antenna, the development of the forewing maculation, and the presence or absence of a dash on the forewing. On the other hand, choosing characters 0, 3, 4, and 5 would give the best chance of seeing a combination that can only occur with a truncated costal sclerite (14:1), which is an apomorphy of *Xylomoia*.

TABLE 3. Characters in combination with character 14, their correlation, and the fraction of successful combinations that predict 14:1. The greatest *r* values and fractions for a number of characters are in bold. The best particular state combinations are not shown.

Combination	*r*	Fraction with 14:1
0	**1.00**	**1/2**
1	0.75	1/3
2	0.25	0/3
3	0.00	0/2
4	0.00	0/2
5	0.00	0/2
0,1	**0.81**	**2/6**
0,2	0.50	**2/6**
0,3	0.44	2/4
0,4	0.44	2/4
0,5	0.44	2/4
1,2	0.46	2/9
1,3	0.38	**2/6**
1,4	0.38	**2/6**
1,5	0.38	**2/6**
2,3	0.19	1/6
2,4	0.31	1/6
2,5	0.19	1/6
3,4	0.00	0/4
3,5	0.00	0/4
4,5	0.00	0/4
0,1,2	**0.59**	4/18
0,1,3	0.53	4/12
0,1,4	0.53	3/12
0,1,5	0.53	4/12
0,2,3	0.35	5/12
0,2,4	0.39	5/12
0,2,5	0.35	5/12
0,3,4	0.28	**4/8**
0,3,5	0.28	**4/8**
0,4,5	0.28	**4/8**
1,2,3	0.33	4/18
1,2,4	0.38	5/18
1,2,5	0.33	5/18
1,3,4	0.24	4/12
1,3,5	0.24	4/12
1,4,5	0.24	4/12
2,3,4	0.24	4/12
2,3,5	0.14	3/12
2,4,5	0.24	4/12
3,4,5	0.00	0/9
0,1,2,3	0.45	8/36
0,1,2,4	**0.47**	9/36

.....continued on the next page

TABLE 3. (Continued)

Combination	*r*	Fraction with 14:1
0,1,2,5	0.45	9/36
0,1,3,4	0.39	8/24
0,1,3,5	0.39	8/24
0,1,4,5	0.39	8/24
0,2,3,4	0.32	11/24
0,2,3,5	0.26	11/24
0,2,4,5	0.32	11/24
0,3,4,5	0.20	**8/16**
1,2,3,4	0.29	12/36
1,2,3,5	0.25	11/36
1,2,4,5	0.29	11/36
1,3,4,5	0.18	8/24
2,3,4,5	0.20	10/24

As an extension of this exercise, one may observe and score one character at a time, followed with new analysis. The script *picktwo.run* is run, keeping the wildcard terminal activated and using it for scoring. On the basis of the prior results, first character 0 is chosen, observed, and scored as state 0. The remaining five characters are run in combination with 14.

Character	*r*	Combinations predicting only 14:1
1	0.75	1/3
2	0.25	0/3
3	0.00	0/2
4	0.00	0/2
5	0.00	0/2

Character 1 is chosen because it is the best, and it is found to have state 2.

2	2.25	2/3
3	2.00	2/2
4	2.00	2/2
5	2.00	2/2

Character 3 is arbitrarily chosen and found to have state 1.

2	2.25	2/3
4	2.00	2/2
5	2.00	2/2

Character 4 is scored as 1.

2	2.25	2/3
5	2.00	2/2

Correlation values greater than 1 may result when not all combinations have been tested because one or more of the characters have been scored.

This is based on a real case. After Mikkola's revision, Lafontaine and Schmidt (2010) transferred *X. indirecta* (Grote) from *Oligia* Hübner. That species may be scored in Mikkola's dataset (1998) as follows: 02011000001113 1011011110000101201. Character 2:0 adds one step to the cladogram (for a length of 87), so this combination had not been predicted. The observed states of the other external characters add no extra steps. Adding the completely scored terminal *X. indirecta* results in two cladograms with length 94 and topology slightly but not greatly altered from the original two.

5. *Sternolophus*. Nasserzadeh *et al*. (2017) performed a cladistic analysis of water beetles of the genus *Sternolophus* Solier, including eight continuous characters. The terminals have single (average) values for the continuous characters, not ranges, which for this analysis obviates having to split terminals. They obtained one cladogram with equal weights (Fig. 10; length = 146.130). Continuous characters 0 and 7 were correlated (Fig. 11), which respectively represent the mean body length in millimeters and the width of the aedeagus (ratio of length to width).

Each cell has the fraction of extra steps needed to accommodate that combination, shaded by the amount of homoplasy. X's represent combinations attested in the dataset, and they are placed on the cell with the nearest rounded values. (The structures were originally measured to two decimal places, but the present analysis only went by single-decimal increments. X's that occur on shaded cells actually occur very close by.) It can be seen that the attested combinations define the edges of the correlation. It is important to search thoroughly for each combination to find the minimum number of steps. If condensing trees after the search (TNT command *collapse [*), it is important to filter out the suboptimal ones (with *best*). It also helps to test at sufficiently fine increments to pick up the nuances of the graph space. The increments are 0.1 in this example, but two decimal spaces would have been preferable. If that were so, the combination 0:12.05, 7:2.45 possessed by *S. immarginatus* Orchymont would have been connected to an area of zeroes.

Reanalysis with implied weights, $k = 5$, results in a different cladogram (Fig. 12; length = 147.800, fit = 9.10054). The correlation of characters 0 and 7 is again shown in Fig. 13. Clearly, a different weighting method can change the character correlation by altering the topology. Some areas gain homoplasy, and other areas lose it, such as around combination 0:12.00 and 7:3.40. A peculiarity of Fig. 13 is a "peninsula" of homoplasy on the right-hand side. That arises because the species that represent the two combinations in the 0:15.50 column are not closely related in the result from implied weighting. A weaker weighting value, k = 10, gives an intermediate situation (not figured) with a narrower peninsula and lower values in the corners. Such peninsulas can also occur in correlations from unweighted analyses.

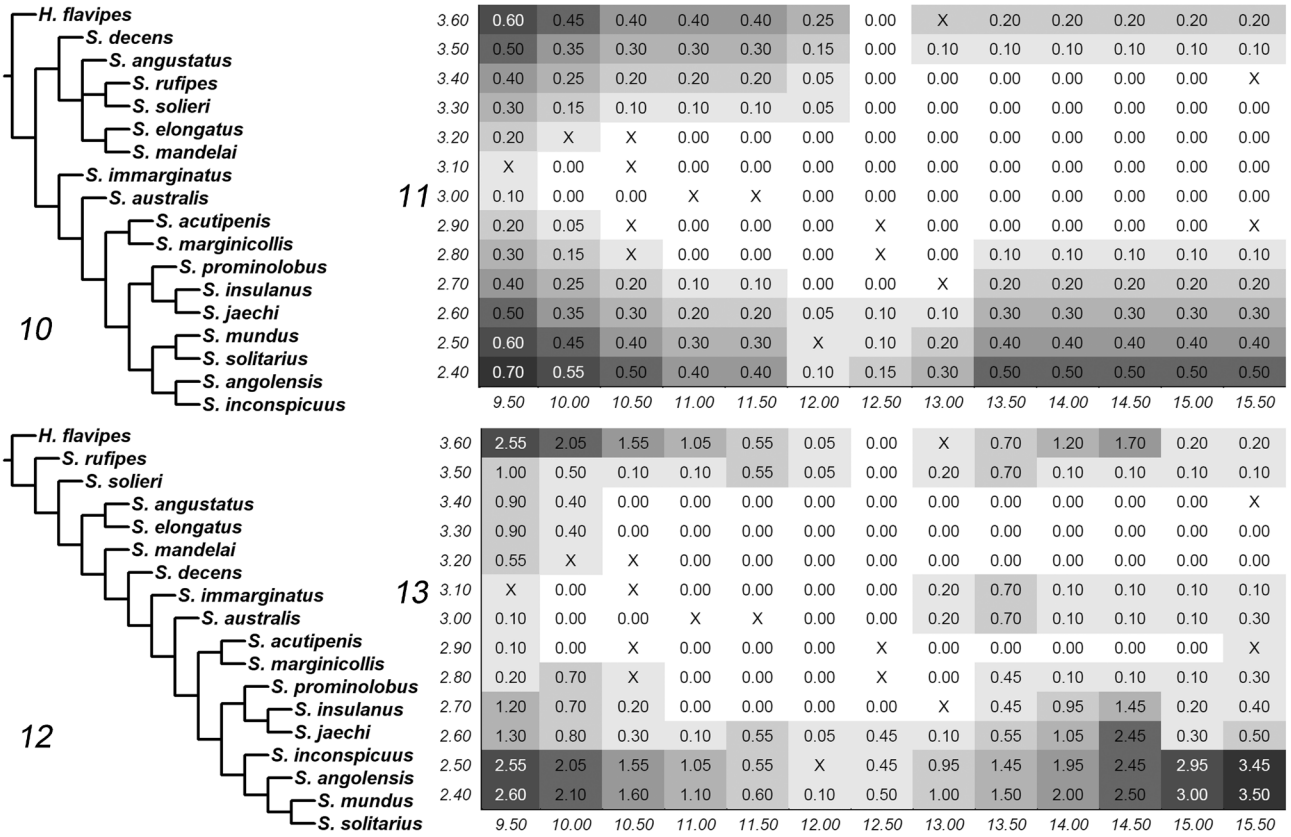
Discussion

The approach outlined here significantly improves taxonomic identification technology by assessing not only what has been observed, but also what is possible. It enables predictions about life stages without rearing or culturing. It is uniquely suited for situations where many undescribed or poorly known species remain in the pool of choices.

This approach of counting the possible steps added to the shortest cladograms builds on the interpretation of homoplasy as a kind of phylogenetic error (Nixon and Carpenter, 2012). It uses homoplasy to quantify the plausibility of observing any given character combination. The connection between cladistic analysis to correlation goes back to Farris (1969, 1979), who argued that hypotheses should be chosen that minimize error. "Correlation" in the standard statistical sense has a specific meaning, and the function of prediction might better correspond to statistical "regression." The difference is beside the point: both are functions of error distributed in Cartesian space, as is the present method. Every point in that space is either observed or else assigned a quantity of error based on the analysis. Strong correlation means that new data points should be observed only in a small part of the space, and weak correlation means that they may reasonably occur in most of the space.

The method is nonparametric because it does not assume that error follows a normal distribution around a central line. With multiple characters, the correlation graphs occupy a hyperdimensional Cartesian space in ways that can defy visualization. "Heat maps" like Fig. 11 and 13 can transform in unusual ways as the topology changes: one must remember that, however strange the distributions of homoplasy may seem, a cladogram underlies them.

There is a trade-off between using fewer characters that have higher correlation and more that have lower. In some cases, observing fewer characters may allow more states unevenly, whereas more characters are more likely to rule out combinations that affect all the states of all characters. When two characters have $r = 1.0$, it is worthwhile to look at their correlation graph to see if the distribution of error is unsatisfactory, and to add more characters to even the distribution out, even if the extra characters reduce r.



FIGURES 10–13. Cladograms of Nasserzadeh et al. (2017) and correlations of continuous characters. 10, single cladogram from equal weights for the *Sternolophus* dataset (length = 146.130); 11, correlation graph based on cladogram with equal weights (x-axis: char. 0; y-axis: char. 7); 12, single cladogram from implied weighting; 13, correlation of characters 0 and 7 based on implied-weighting cladogram.

In the worked examples, proxy selection took a lot of time due to character exploration. A shortcut method would test only the combinations that include characters with the greatest *r* values in smaller sets: e.g. character 0 in *Ecdytolopha* and characters 1, 2 and/or 5 in *Xylomoia*. Not all of the numerous possible combinations are tested; it is more efficient to test only those that performed well in smaller sets. A caveat is that this is not proven to be a general behavior. Among larger sets, some combinations could have higher *r* values despite not including characters that have high *r* values in smaller sets.

Some improvements could be made to this approach, building on the principles outlined here. First, equations for *m* (the maximum number of steps) are needed for standard weights, multiple ordered characters, and continuous characters. Second, the extension from continuous characters to landmark characters should be possible. Third, a computational method would be useful to assess the sets of states directly on each node, instead of indirectly probing with a wildcard. The latter method is limited by the heuristic search, whereas the former could be calculated more efficiently. It would still need a way to count the number of extra steps for non-zero values. Finally, by whichever method, the user should have the option of constraining to particular character-state reconstructions along the cladogram, as an alternative to ambiguous reconstructions. As noted above, ambiguity has a partial but not total influence on the correlation values, so choosing particular reconstructions might increase those values generally but in ways that are not entirely predictable.

Acknowledgments

This idea was first developed and explored with the support of the Cornell University Graduate Program and the Rea Postdoctoral Fellowship of the Carnegie Museum of Natural History. Thanks to Eduardo Almeida, Jason Dombroskie, Jessica Awad, Paul Skelley, and two anonymous reviewers for critical suggestions and advice. The Willi

Hennig Society sponsored TNT 1.5. This publication was partly supported by the Florida Department of Agriculture and Consumer Services, Division of Plant Industry.

References

Adamski, D. & Brown, J.W. (2002 [2001]) Systematic revision of the *Ecdytolopha* group of genera (Lepidoptera: Tortricidae: Grapholitini) in the New World. *Entomologica Scandinavica*, Supplement 58, 1–86.

Carpenter, J.M. (1988) Choosing among multiple equally parsimonious cladograms. *Cladistics*, 4, 291–296. https://doi.org/10.1111/j.1096-0031.1988.tb00476.x

Dallwitz, M.J. (1974) A flexible computer program for generating identification keys. *Systematic Zoology*, 23, 50–57. https://doi.org/10.1093/sysbio/23.1.50

Farris, J.S. (1969) A successive approximations approach to character weighting. *Systematic Zoology*, 18 (4), 374–385. https://doi.org/10.2307/2412182

Farris, J.S. (1979) The information content of the phylogenetic system. *Systematic Zoology*, 28 (4), 483–519. https://doi.org/10.2307/sysbio/28.4.483

Felsenstein, J. (1985) Phylogenies and the comparative method. *American Naturalist*, 125 (1), 1–15. https://doi.org/10.1086/284325

Fitch, W.M. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, 20 (4), 406–416. https://doi.org/10.1093/sysbio/20.4.406

Goloboff, P. & Catalano, S. (2016) TNT, version 1.5, with a full implementation of phylogenetic morphometrics. *Cladistics*, 32 (3), 221–238. https://doi.org/10.1111/cla.12160

Hall, A.V. (1970) A computer-based system for forming identification keys. *Taxon*, 19 (1), 12–18. https://doi.org/10.2307/1217908

Hennig, W. (1965) Phylogenetic systematics. *Annual Review of Entomology*, 10, 97–116. https://doi.org/10.1146/annurev.en.10.010165.000525

Lafontaine, J.D. & Schmidt, B.C. (2010) Annotated check list of the Noctuoidea (Insecta, Lepidoptera) of North America north of Mexico. *ZooKeys*, 40, 1–239. https://doi.org/10.3897/zookeys.40.414

Lees, D.C. & Smith, N.G. (1991) Foodplant associations of the Uraniinae (Uraniidae) and their systematic, evolutionary, and ecological significance. *Journal of the Lepidopterists' Society*, 45 (4), 296–347.

Lucidcentral (2021) Lucid Player. Version 4. Available from: https://www.lucidcentral.org/index.php/lucid-player/ (accessed 4 April 2021).

Mikkola, K. (1998) Revision of the genus *Xylomoia* Staudinger (Lepidoptera: Noctuidae), with descriptions of two new species. *Systematic Entomology*, 23, 173–186. https://doi.org/10.1046/j.1365-3113.1998.00055.x

Nasserzadeh, H., Alipanah, H. & Gilasian, E. (2017) Phylogenetic study of the genus *Sternolophus* Solier (Coleoptera, Hydrophilidae) based on adult morphology. *ZooKeys*, 712, 69–85. https://doi.org/10.3897/zookeys.712.14085

Nixon, K.C. & Carpenter, J.M. (2012) On homology. *Cladistics*, 28, 160–169. https://doi.org/10.1111/j.1096-0031.2011.00371.x

Nixon, K.C. & Davis, J.I. (1991) Polymorphic taxa, missing values and cladistic analysis. *Cladistics*, 7, 233–241. https://doi.org/10.1111/j.1096-0031.1991.tb00036.x