

A Method for Estimating the Relative Importance of Characters in Cladistic Analyses

DAVID DEGUSTA

*Department of Anthropological Sciences, Building 360, Stanford University, Stanford, CA 94305-2117, USA;
E-mail: degusta@stanford.edu*

Abstract.—The method of character importance ranking (CIR) is proposed here as a means for estimating the relative “importance” of characters in cladistic analyses, especially those based on morphological features. CIR uses the weighting variable to incrementally remove one character at a time from the analysis, and then evaluates the impact of the removal on the shape of the cladogram. The greater the impact, the more important the character. The CIR method for determining which characters drive the shape of a particular cladogram has several applications. It identifies the characters with the strongest (though not necessarily most accurate) signal in a cladistic analysis; it permits the informed prioritization of characters for further investigation via genetic, developmental, and functional approaches; and it highlights characters whose definition, scoring, independence, and variation should be reviewed with particular care. The application of CIR reveals that at least some cladograms depend entirely on a single character. [Character analysis; cladogram shape; importance; parsimony.]

The cladograms generated by a particular cladistic analysis obviously depend, in part, on the characters used in the analysis. It is unlikely, though, that all characters have equal influence on the shape of the resulting cladogram, even in an unweighted analysis. A variety of measures exist that examine the “quality” of individual characters relative to a given cladogram (i.e., their congruence, consistency, level of homoplasy, etc.). However, there appears to be a paucity of methods that examine the influence of individual characters on the shape of a given cladogram, apart from their “quality.” For example, the removal of one particular character from an analysis may not change the most parsimonious cladogram at all, and thus this character has little influence on cladogram shape. In contrast, the removal of another character from the analysis may result in changes at several nodes, rendering it quite “important” in this sense (separate from “quality” measures). A new method, termed character importance ranking (CIR), is proposed here for quantifying the relative importance of the characters used in a cladistic analysis.

CIR is useful because it identifies the degree to which a cladogram depends on individual characters, as well as identifies those characters, thus facilitating further study of the most influential characters. Problems with important characters (e.g., functional correlations, non-heritability, etc.) will, by definition, have a greater confounding effect on a cladogram than will such problems with relatively unimportant characters. Although the results of additional study of the important characters may eventually lead to modifications in character lists and definitions, CIR is *not* proposed as a means for selecting or weighting characters, nor for discriminating between “good” characters and “bad” characters, nor is it a character weighting method. CIR simply identifies those characters with the greatest impact on the shape of a given cladogram—it evaluates the strength of a character’s signal, but not the accuracy of that signal. As such, it is conceptually distinct from indices of character “quality” (e.g., the consistency index).

Although CIR was developed for morphological parsimony analyses, and is presented here mainly in that

regard, it is theoretically applicable to all types of data (though the computation of CIR is probably impractical for large DNA sequence databases). With further work, it could be productively extended for application to likelihood analyses.

METHODS

The CIR method calculates an “importance value” for each character in a given cladistic analysis. CIR does so by using the weighting variable to incrementally remove a single character at a time from an analysis (e.g., the character weight is set at 0.8, then 0.6, 0.4, 0.2, and finally 0). For each new weight of the character, the most parsimonious tree is generated. Then, using the symmetric difference metric (which measures the number of splits present in one tree but not the other; Steel and Penny, 1993), the topographic distance is measured between the original most parsimonious tree and each of the trees produced by the incremental downweighting. The mean of these distances is taken to be the “importance value” of that character, as it represents the effects on tree shape of the downweighting and incremental removal of that character from the analysis. A character that can be entirely removed without altering the cladogram will have an importance value of 0. If setting the weight of a character to slightly less than 1 induces major changes in the cladogram, that character will have a very high importance value. For the purposes of this paper, then, “important” characters are defined as those with the greatest impact on the shape of the resulting cladogram (i.e., those with the largest importance values as calculated by CIR). It is important to keep in mind that CIR only considers the impact of a character’s downweighting on the shape of the cladogram, which is here termed “importance” as a shorthand.

The importance value of a character represents the mean topographic distance between the original cladogram and the cladograms produced by variably weighting that character. However, absolute importance values are generally not comparable across analyses, due to differences in the parameters of the analyses (i.e., number

of most parsimonious cladograms) as well as the dependence of most distance measures on the number of operational taxonomic units (Steel and Penny, 1993). In other words, a character with an importance value of 2 in one analysis is not necessarily twice as important as a character in a different analysis with a value of 1; though this would be the case if both characters were from the same analysis. The rank ordering of characters by importance value is, of course, always comparable across analyses, and this should be the focus of interpretation.

Because the goal of CIR is to establish the *relative* importance of the characters in a given analysis, the existence of multiple "best" (equally parsimonious) cladograms generally does not preclude its application, though more than about 10 equally parsimonious trees makes for tedious comparisons. Similarly, whatever choices of settings (i.e., search algorithm, assumptions, etc.) were made in a given analysis can be maintained in the process of character importance ranking.

The incremental downweighting of an individual character in CIR allows for a more refined estimate of the character's importance than the simple inclusion/exclusion of the character. For example, imagine that the complete removal of character 1 from an analysis results in a particular change "X" to the resulting most parsimonious tree, and that the complete removal of character 2 from the analysis also produces the same change "X." In terms of inclusion/exclusion, then, characters 1 and 2 would have identical importance values. But perhaps just downweighting character 1 to slightly less than its original weight of 1.0 (say a weight of 0.8) also produces change "X" to the tree, whereas a similar downweighting of character 2 produces no change. In terms of importance, as defined here, character 1 is more important than character 2, a difference detectable only by incremental downweighting rather than simple inclusion versus exclusion.

For some cladistic analyses, CIR will demonstrate that there are no individual "important" characters (i.e., all characters have equal, and zero, importance values). Although this result obviously precludes the prioritization of characters for further investigation, it does document that the cladogram in question is independent of any single character, which is useful information. A more general limitation is that CIR does not identify characters that may only be important in combination with others. In other words, it is theoretically possible that two characters might fail to alter the best cladogram when removed individually, but their removal in tandem may have a major effect on the cladogram.

To apply CIR to a cladistic analysis requires the data set for the analysis and the particular settings used. The CIR method is implemented as two Nexus files for PAUP* 4.0 (Swofford, 1998). By using the CIR PAUP files in conjunction with a spreadsheet program, character importance values can be obtained almost automatically. The generation of importance values typically takes between 30 minutes and 2 hours, depending on data set size and processor speed. The general algorithm for CIR is given in the Appendix. The CIR

PAUP* 4.0 files, along with detailed instructions and the data sets analyzed here, can be downloaded from <http://www.stanford.edu/~degusta>.

Subsequent to the development of CIR, a related method was located in the literature—the sequential character removal (SCR) method (Davis, 1993; Davis et al., 1993). The purpose of SCR, however, is completely different than CIR: SCR provides an index of stability for individual clades. In SCR, separate cladistic analyses are conducted of all possible data sets derived by the removal of individual characters and character combinations of successively increasing number. The resulting clade stability index (CSI) is the ratio of the minimum number of characters whose removal causes the collapse of that clade to the total number of informative characters (Davis, 1993).

SCR is not aimed at identifying important characters. It also employs only the complete removal of multiple characters, rather than CIR's incremental downweighting of single characters. SCR is computationally intensive, rendering the full implementation of the method impractical for even small data sets (Davis, 1993), whereas CIR is computationally practical for all save the very largest data sets.

MATERIALS

CIR is applied here to three cladistic analyses to illustrate the utility of this method: Graham et al.'s (1991) analysis of charophycean green algae (9 OTUs, 21 morphological characters), Chamberlain and Wood's (1987) analysis of fossil hominids (9 OTUs, 90 morphological characters), and Anderberg's (1986) analysis of *Pegolettia* plants (9 OTUs, 19 morphological characters). The details of the cladistic analyses used, including the data sets, can be found in the original publications and will not be repeated here. These three analyses are representative of CIR results based on the circa 20 cladistic analyses to which CIR has been applied to so far.

RESULTS

Applying CIR to the cladistic analysis of charophycean green algae by Graham et al. (1991) demonstrates that 4 characters are more important (values of 0.3 and 0.4) than the remaining 17 characters (values of 0), as detailed in Table 1. Three of these four important characters are characteristics of the zygote: "eggs retained" (12), "zygote retained" (13), and "substantial enlargement of the zygote" (14). Graham et al. (1991) note that these three characters might be collapsed into a single multistate ordered character, but justify their division on the grounds of clarity and caution. The results of CIR demonstrate that the Graham et al. (1991) cladogram depends on this decision, whereas this is not the case for similar decisions about other characters in their analysis. Indeed, excluding two of these characters (12 and 13) results in eight equally parsimonious trees, many with notably different topology than the one original most parsimonious cladogram. Although CIR does not bear on the question

TABLE 1. Importance scores, tree steps, and consistency index for characters in Graham et al. (1991).

Character no.	CIR score	Tree steps	CI
6	0.43	1	1
13	0.43	1	1
12	0.33	1	1
14	0.33	1	1
1	0	1	1
2	0	1	0
3	0	2	1
4	0	1	1
5	0	3	0.33
7	0	1	1
8	0	1	1
9	0	1	1
10	0	1	1
11	0	1	1
15	0	1	1
16	0	1	1
17	0	3	0.33
18	0	1	1
19	0	1	1
20	0	1	1
21	0	1	1

of whether these characters should be collapsed or not, it has spotlighted the critical importance of this decision.

Applying CIR to the cladistic analysis of *Pegolettia* by Anderberg (1986) revealed five characters as being more important (values of 1.5 to 3.7) than the remaining 14 characters (values of 0), as shown in Table 2. Of these five characters (7 to 11), two relate to the presence or absence of crystal clumps (either on anthers or style-branches), suggesting that this feature is of substantial phylogenetic importance for this group. As such, further investigation of the comparative morphology, ontogeny, and formation of crystal clumps may directly impact estimations of *Pegolettia* phylogeny.

Applying CIR to the cladistic analysis of fossil hominids by Chamberlain and Wood (1987) demonstrates

TABLE 2. Importance scores, tree steps, and consistency index for characters in Anderberg et al. (1986).

Character no.	CIR score	Tree steps	CI
11	3.67	1	1
7	3.63	1	1
9	3.29	2	0.5
10	1.9	2	0.5
8	1.5	2	0.5
1	0	Constant	Constant
2	0	Constant	Constant
3	0	Constant	Constant
4	0	1	1
5	0	1	1
6	0	1	1
12	0	3	0.33
13	0	1	1
14	0	1	1
15	0	1	1
16	0	1	1
17	0	1	1
18	0	1	1
19	0	1	1

TABLE 3. Importance scores, tree steps, and consistency index for selected characters from Chamberlain and Wood (1987). All "important" characters are shown paired, where possible, with similar "unimportant" characters.

Character no.	CIR score	Tree steps	CI
52	0.86	4	0.5
2	0	4	0.5
53	0.86	3	0.667
6	0	3	0.667
57	0.67	2	0.5
8	0	2	0.5
61	0.67	4	0.75
22	0	4	0.75
62	0.86	3	1
81	0.67	2	1
19	0	2	1

that six characters are more important (values of 0.7 to 0.9) than the remaining 84 characters (values of 0), as outlined in Table 3. The two most important characters are mental foramen height and the height of the foramen supraspinosum, for which Chamberlain and Wood (1987) used the measurements reported in Chamberlain (1987). Given the dependency of the cladogram on these characters, the raw measurements for these two variables were reviewed. This examination revealed that Chamberlain (1987) lists measurements of supraspinosum height for a number of fossils that do not preserve any trace whatsoever of this feature (e.g., AL 128-23, AL 288-1i, CKT Jaw K) and includes highly problematic measurements of mental foramen height as well. If either of those two characters, or both, are removed from the cladistic analysis, the resulting cladograms are all markedly different than that of the original analysis (as predicted by CIR). These clear errors in scoring could have been located without CIR, but the use of CIR identified the characters for which a careful review was most crucial.

DISCUSSION

The preceding examples demonstrate how CIR can identify issues regarding character definition (as for Graham et al., 1991) and character scoring (as for Chamberlain and Wood, 1987), as well as suggesting characters whose further study may be particularly valuable for systematics (as for Anderberg, 1986). Based on those cases, as well as the application of CIR to an additional circa 20 analyses (data not shown), it is possible to draw some general conclusions regarding the nature of the CIR method. There is not a correlation between the "importance" of a character (in terms of CIR) and the consistency of that character with the particular cladogram (Tables 1 to 3), confirming theoretical expectations. The number of characters identified as "important" (i.e., importance value greater than 0) does not seem to scale significantly with the ratio of number of characters to number of OTUs. In any case, the number of "important" characters identified for an analysis is of much less interest than which characters they are (as illustrated above). The approach of averaging the topographic distances

obtained for the incremental downweighting of a particular character was found to be a reasonable representation of the tree shape change, as the rank order of the "important" characters was not altered by using a threshold approach (see Appendix) or several other alternatives. Comparison of these results with those obtained by simple exclusion of characters shows that a more refined estimate of importance (i.e., fewer characters with identical importance values) is obtained with incremental downweighting.

Even so, it is important to keep in mind that CIR only considers the impact of a character on the shape of a cladogram. It is clear that characters identified as important may be those that provide widespread support (in terms of supporting many different nodes), or they may be "extreme" characters that strongly influence the cladogram shape in a particular fashion, or they may be characters that maintain a weakly supported node, or they may be characters that maintain a balance between competing cladograms. By examining the distribution of "important" characters on particular trees, it is apparent that although they sometimes involve deep nodes, they sometimes involve terminal nodes. Such characters may or may not have anything in common beyond altering the cladogram (which they may do in a wide variety of ways, both within an analysis and across analyses). The crucial point is that "important" characters directly influence the result of the analysis, and therefore merit more immediate scrutiny than those characters which do not. Such further scrutiny should include consideration of the specific ways in which the tree is altered (local versus global changes), and the reasons for such alterations (data conflict versus lack of data). At this point, it does not seem possible to predict, a priori, which characters will be important, which supports the use of CIR as a nonredundant index.

CONCLUSIONS

By systematically identifying those characters that directly influence the resulting cladogram, CIR spotlights those characters most in need of careful review. As demonstrated here, such review often results in significant alterations to the preferred cladogram. CIR also identifies the extent to which cladistic analyses depend on single characters, documenting that (for most cladistic analyses) not all characters contribute equally to the shape of the cladogram. By identifying the characters with the strongest (though not necessarily most accurate) signal in a cladistic analysis, CIR sets up a feedback loop between characters and trees that may prove useful in refining estimates of phylogeny.

ACKNOWLEDGMENTS

For helpful discussions and assistance, I thank W. Henry Gilbert, Ken Angielczyk, F. Clark Howell, John Hutchinson, Diogo Meyer, Kevin Padian, James Parham, Alan Shabel, and Tim White, all at the University of California, Berkeley. I also thank an anonymous reviewer and associate editor Dan Faith for many helpful comments. This work was

supported in part by the Laboratory for Human Evolutionary Studies of the Museum of Vertebrate Zoology, U. C. Berkeley.

REFERENCES

- Anderberg, A. 1986. The genus *Pegolettia* (Compositae, Inuleae). *Cladistics* 2:158–186.
- Chamberlain, A. T. 1987. A taxonomic review and phylogenetic analysis of *Homo habilis*. Ph.D. dissertation, University of Liverpool.
- Chamberlain, A. T., and B. A. Wood. 1987. Early hominid phylogeny. *J. Hum. Evol.* 16:119–133.
- Davis, J. I. 1993. Character removal as a means for assessing stability of clades. *Cladistics* 9:201–210.
- Davis, J. I., M. W. Frohlich, and R. J. Soreng. 1993. Cladistic characters and cladogram stability. *Syst. Bot.* 18:188–196.
- Graham, L. E., C. F. Delwiche, and B. D. Mishler. 1991. Phylogenetic connections between the 'green algae' and the 'bryophytes.' *Adv. Bryol.* 4:213–244.
- Steel, M. A., and D. Penny. 1993. Distributions of tree comparison metrics: Some new results. *Syst. Biol.* 42:126–141.
- Swofford, D. L. 1998. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4.0b4. Sinauer Associates, Sunderland, Massachusetts.

First submitted 4 December 2002; reviews returned 15 August 2003;

final acceptance 7 March 2004

Associate Editor: Dan Faith

APPENDIX

CHARACTER IMPORTANCE RANKING ALGORITHM

- Variables are defined as follows.
 - Start with a cladistic analysis that uses characters C_1 through C_n .
 - Let each character have a corresponding weighting variable W_1 through W_n .
 - Let T_0 be the one best cladogram that results from an analysis of characters C_1 through C_n (if there are multiple equally parsimonious cladograms, use each one sequentially and compare the results).
 - Let IMP_x be the importance value of character C_x .
- Let W_x assume a range of values between 1 and 0 (e.g., 0.9, 0.8, ..., 0).
 - If the original analysis used character weighting, then either multiply each of the values between 1 and 0 by the original weight, or discard the original weighting to establish the importance of unmodified characters.
- Let all other $W = 1$.
- For each value of W_x , determine the most parsimonious cladogram(s).
- For each of those cladograms, calculate the distance (D) between it and T_0 . Different distance measures can be employed, but the current implementation uses the PAUP* 4.0 symmetric difference metric (Steel and Penny, 1993).
- $IMP_x = \text{mean of all } D \text{ calculated in step 5.}$
 - In the current implementation, the mean is taken for all the distances. It may be more precise to first calculate the mean of each set of distances (if there is more than one) for each weighting value, and then take the mean of those averages. Preliminary empirical findings suggest, however, that these two approaches yield very similar results in practice.
 - Alternately, a threshold approach could be used whereby IMP_x equals the maximum value of W_x such that D_{W_x} is greater than some constant.
- Repeat steps 2 through 6 for each W (i.e., $x = 1$ to n). In other words, the weighting variable for each character is manipulated in turn.
- Rank all characters by IMP value.

N.B. The CIR PAUP* 4.0 files, along with detailed instructions and data sets, can be downloaded from <http://www.stanford.edu/~degusta>.