ARTICLES

A new method for increasing the robustness of cladistic analyses

David Morrison
Institute of Banksia Studies

Introduction

A cladogram is a set of explicit hypotheses concerning the phylogenetic relationships among a group of taxa. As such, it is amenable to testing and re-assessment in the light of new character data, or even by a new examination of the old character data. The robustness of a cladogram (or part of a cladogram) is then a measure of how resistant the cladogram is to change, either as new data are introduced into the data matrix or as the old data are modified. Robustness is usually a product of the congruence between the characters in the data matrix, and is therefore a result of the number of characters that support any particular branch of the cladogram.

Many different methods have been proposed for measuring the robustness, or degree of confirmation, of a cladogram or parts of a cladogram. These include consistency indices (Farris, 1989; Archie, 1989b), information statistics (Brooks *et al.*, 1986), normal deviate tests (Le Quesne, 1989), randomization tests (Archie, 1989a; Faith and Cranston, 1991; Faith, 1992), and bootstrapping (Felsenstein, 1985; Sanderson, 1989). However, all of these methods are sensitive to the same basic characteristic of the data matrix:- the maximum number of congruent characters (i.e. those characters that agree on the same tree topology).

Therefore, in order to increase the robustness of a cladogram we need to increase the number of congruent characters in our data matrix. That is, we don't just need more data — rather, we need more data that agree with the data that we already have. There are two ways of doing this:- 1) we can employ the empirical method of selectively collecting more data from the specimens at our disposal; or 2) we can employ the inductive logical method of data enrichment.

The latter method is preferable because it allows us to increase the size of our data matrices without recourse to the expense and trouble of actually doing any further work. It also gives us far greater control over our data, because there is no guarantee with the former method that we will be able to find any further congruent characters — the data enrichment method does, however, offer this guarantee. The only requirements for use of the data enrichment method are that you already know what the answer should be, and that you already have at least some data. The method does not, therefore, offer you complete freedom from the inconvenience and embarrassment of uncontrollable empirical results. Nevertheless, most systematists can meet both of these minimum requirements for the use of this method.

The data enrichment method was first proposed by Lewis (1957), and it has since been applied in many fields of scientific endeavour, including hazardology (Cholmondeley and Mayer, 1978), sociology (Hickey, 1983), and psychology (McKinley *et al.*, 1986). However, I believe that this is the first time its application to cladistic analysis has been proposed. Therefore, I will first illustrate the method by presenting the classic example of Lewis (1957), and then I will explore the possibilities of this method for cladistics by reference to an example from the plant family Epacridaceae.

The Data Enrichment Method

By way of example, Lewis (1957) presents an experiment performed to test the ability of a specific sound receiver to detect an audio signal. The experiment is performed in such a way that in each of a series of trials the experimenter learns either that detection was accomplished or that it was not. Trials are made with the source intensity set at six different levels, and at each of these six intensity levels a number of tests are made and the result (detection or no detection) recorded. The data from the initial experiment of 213 tests are summarized in Table 1.

We now wish to increase the amount of data available at each source level by the method of data enrichment. It is reasonable to assume that sound detectability is a function of source level and that, if all other parameters are held constant, a loud sound

Table 1. Raw data.

Source level (db)	Number of detections	Number of failures to detect	Probability of detection
62	5	40	0.11
65	10	30	0.25
68	15	20	0.43
71	20	10	0.67
74	25	5	0.83
77	30	3	0.91

Table 2. Enriched data.

Source level (db)	Number of virtual detections	Number of virtual failures	Probability of detection
62	5	108	0.04
65	15	68	0.18
68	30	38	0.44
71	50	18	0.74
74	75	8	0.90
77	105	3	0.97

is easier to detect than one of smaller intensity. Thus, it is safe to assume that if a signal was detected at a given level then it would have been detected at all higher source intensity levels. Moreover, if a signal was not detected at a given level then it would not have been detected at any lower level of source intensity. Using these simple assumptions, the data collected at one source level can be used to add to the data available for other levels, since we know how these experiments would have come out had we actually performed them. So, the number of detections at any one source intensity is added to the data for all of the higher source intensities, and the number of failures to detect is added to the data for all of the lower source intensities. Treating all of the data in this fashion, we can compile the results of 213 actual tests and a further 310 "virtual" tests, as shown in Table 2.

Two things are apparent at once. Firstly, the probabilities of detection given in Table 2 are quite different from those deduced crudely and empirically in Table 1. Secondly, the number of "virtual" trials at each level of source intensity is much larger than the actual number of trials. Hence, we may be more confident of the results in Table 2 than we are of any of the results in Table 1.

Application to Cladistics

It is clear that application of the data enrichment method to cladistic analysis is straightforward, and that we can therefore dramatically increase the size of a data matrix without the necessity of actually collecting any more data. The application follows from the observation that a synapomorphy on one branch of a cladogram is also logically a synapomorphy on all branches of the tree that are closer to the terminal taxa. Consequently, any branch on a cladogram not only has those synapomorphies that are ascribed to it by the computer analysis program but also has "virtual" synapomorphies from all branches that are closer to the root of the tree.

This can be made obvious by an example. Figure 1 is a modified version of the preferred cladogram presented by Morrison & Powell (1990) from their empirically rigorous cladistic analysis of the plant family Epacridaceae and its relatives. The analysis was of 39 genera from the Epacridaceae, Ericaceae and Clethraceae, for which there were only 21 empirically-derived binary characters. The cladogram shown has 31 steps, but unfortunately it's one of 26 equally parsimonious trees found by the microcomputer program "Hennig86" using an exact algorithm. The consistency index is 0.67, and the retention index is 0.79. The permutation tail probability (PTP) is 0.01 (this being the probability of obtaining a tree with this amount of cladistic structure by chance alone), but the shortest tree is only two steps shorter than the trees produced by the data randomization. It is obvious from these statistics and the large number of alternative trees that there is only a moderate degree of robustness in the cladogram shown, and consequently it is problematic to justify any claim that it represents the true phylogeny.

However, we can now apply the logic of the data-enrichment method to this data set. For example, the branch near the root of the tree is indicated in Figure 1 as being supported only by character 2, but it is actually also supported by character 1, giving us one new "virtual" character on that branch. The next branch from the root is indicated as only being supported by character 3, but it is actually supported by characters 1 and 2 as well,

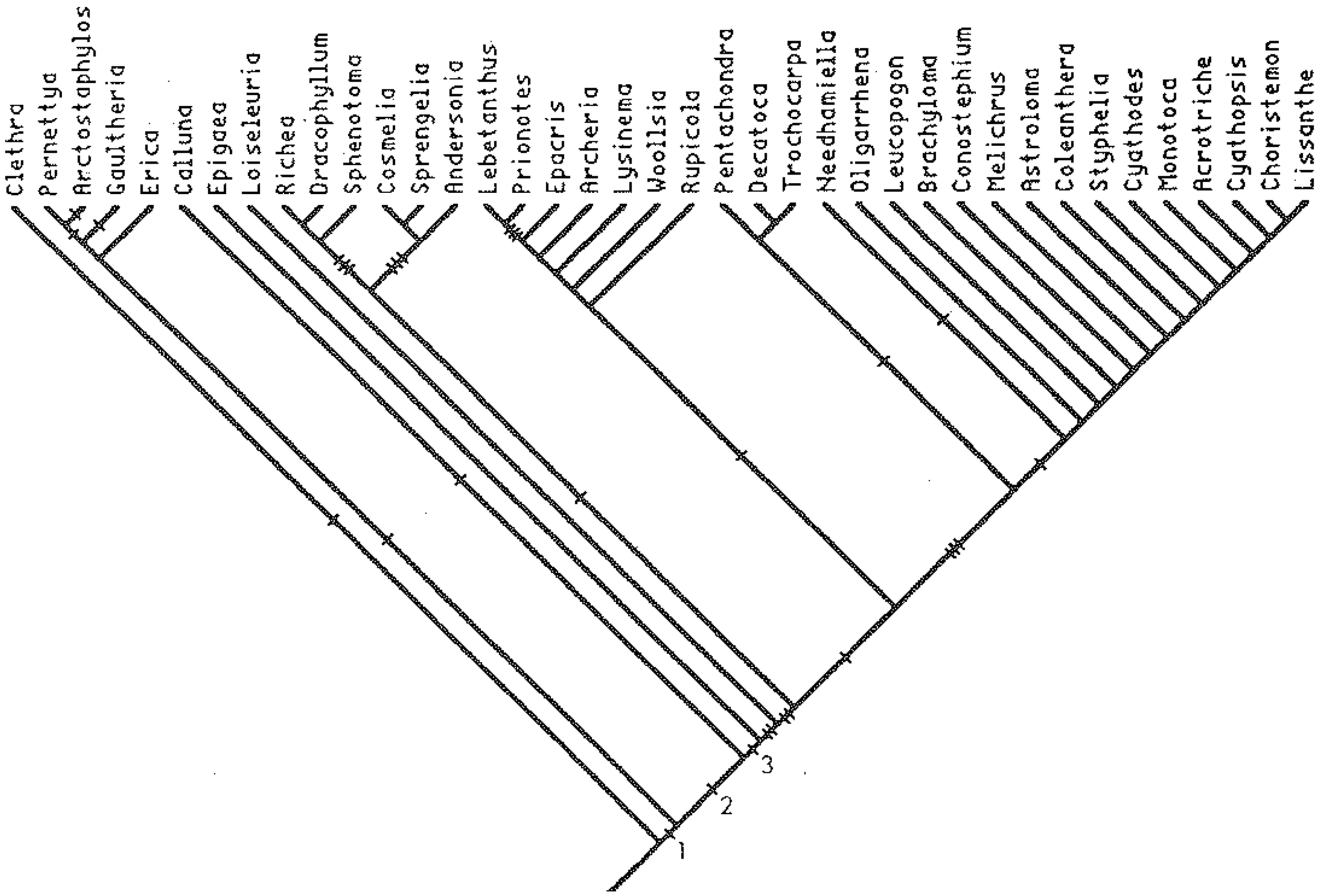

Figure 1. Cladogram of the Epacridaceae and its relatives. All of the genera are from the Epacridaceae except:- *Clethra* (Clethraceae); *Arctostaphylos*, *Gaultheria* and *Pernettya* (Vaccinioideae, Ericaceae); *Calluna* and *Erica* (Ericoideae, Ericaceae); and *Epigaea* and *Loiseleuria* (Rhododendroideae, Ericaceae). All of the character changes are marked, but only those character changes referred to in the text are numbered.

giving us two more "virtual" characters for that branch. In this fashion, we can increase the number of characters in the data set to 112 binary characters, with 21 of the old empirical characters and 91 of the new "virtual" characters.

The cladogram derived from the analysis of this enriched data set has exactly the same topology as the old one, but it now has 122 steps, and it's the only tree of that length. The consistency index has been increased to 0.91, and the retention index is now 0.95. Furthermore, the tree is now nearly 100 steps shorter than the trees produced by the PTP test. This is therefore a far more robust (and therefore more desirable) cladogram.

So, we have substantially increased the size of the data set, and all of the new characters are completely congruent with our chosen cladogram. We now have only one most parsimonious cladogram, and we have also increased the degree of confirmation of each branch on the tree. Consequently, we can now be more confident that our cladogram represents the true phylogeny.

Conclusion

It is clear from this example that the data enrichment method has three desirable contributions to make to cladistics:- 1) we can increase the size of our data matrices without recourse to the difficulties inherent in the empirical method; 2) the new data will always be congruent with our preferred cladogram; and 3) the resulting cladogram will always be more robust. What more could you ask for?

Acknowledgements

Thanks to John Trueman and Dan Faith for doing the PTP tests.

References

Archie, J.W. (1989a) A randomization test for phylogenetic information in systematic data. *Syst. Zool.* **38**: 239-252.

Archie, J.W. (1989b) Homoplasy excess ratios: new indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the consistency index. *Syst. Zool.* **38**: 253-269.

Brooks, D.R., O'Grady, R.T., and Wiley, E.O. (1986) A measure of the information content of phylogenetic trees, and its use as an optimality criterion. *Syst. Zool.* **35**: 571-581.

Cholmondeley, L., and Mayer, H. (1978) On a new method in hazardology. *J. Irrep. Results* **24(1)**: 17-18.

Faith, D.P. (1992) Cladistic permutation tests for monophyly and polyphyly. *Syst. Zool.* in press.

Faith, D.P., and Cranston, P.S. (1991) Could a cladogram this short have arisen by chance alone?: On permutation tests for cladistic structure. *Cladistics* **7**: 1-28.

Farris, J.S. (1989) The retention index and the rescaled consistency index. *Cladistics* **5**: 417-419.

Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783-791.

Hickey, J.L.S. (1983) Reducing automobile accidents. In: *The Best of the Journal of Irreproducible Results* (G.H. Scherr, ed.), p. 88. Workman Publishing, New York.

Le Quesne, W.J. (1989) The normal deviate test of phylogenetic value of a data matrix. *Syst. Zool.* **38**: 51-54.

Lewis, H.R. (1957) The data enrichment method. *Operations Research* **5:** 551-554.

McKinley, C.K., Reid, W.B., and Anonymous. (1986) The re-assessment of criticism and defenses of depth psychology with supporting data. In: *The Journal of Irreproducible Results: Selected Papers* (G.H. Scherr, ed.), pp. 56-57. Dorset Press, New York.

Morrison, D.A., and Powell, J.M. (1990) Systematics and evolution of the Epacridaceae. *IX Meeting of the Willi Hennig Society Abstracts*, p. 39.

Sanderson, M.J. (1989) Confidence limits on phylogenies: the bootstrap revisited. *Cladistics* **5**: 113-129.