II

Parsimony, character analysis, and optimization of sequence characters

CHAPTER 4

The logic of the data matrix in phylogenetic analysis

Brent D. Mishler

4.1 Introduction

The process of phylogenetic analysis inherently consists of two phases. First a data matrix is assembled, then a phylogenetic tree is inferred from that matrix. There is obviously some feedback between these two phases, yet they remain logically distinct parts of the overall process. One could easily argue that the first phase of phylogenetic analysis is the most important; the tree is basically just a re-representation of the data matrix with no value added. This is especially true from a parsimony viewpoint, the point of which is to maintain an isomorphism between a data matrix and a cladogram. We should be very suspicious of any attempt to add something beyond the data in translating a matrix into a tree!

Paradoxically, despite the logical preeminence of data matrix construction in phylogenetic analysis, by far the greatest effort in phylogenetic theory has been directed at the second phase of analysis, the question of how to turn a data matrix into a tree. Extensive series of publications have been elaborated to attempt to justify such tree building approaches as neighbor-joining, maximum likelihood, and Bayesian inference, while ignoring entirely the nature of the data matrix that must underlie any analysis. The reasons for this asymmetry in research on phylogenetic theory are not entirely clear, but it probably has to do with the fact that the problem of tree building may appear simpler, more clear-cut. Perhaps it is just a matter of research fashions. For whatever reason, relatively little attention has been paid to the assembly of the data matrix, and it is high time to examine this all-important part of systematic research. At stake are each of the logical elements of the data matrix: the rows (what are the terminals?), the columns (what are the characters?), and the individual entries (what are the character states?).

The tree of life is inherently fractal-like in its complexity, which complicates the search for answers to these questions. Look closely at one *lineage* of a phylogeny (defined as a diachronic connection between an ancestor and a descendent) and it dissolves into many smaller lineages, and so on, down to a very fine scale. Thus the nature of both the *terminal units* (TUs; the twigs of the tree in any particular analysis) and the characters (hypotheses of homology, markers that serve as evidence for the past existence of a lineage) change as one goes up and down this 'fractal' scale. Furthermore, there is a tight interrelationship between TUs and character states, since they are reciprocally recognized during the character analysis process.

This chapter will deal with logical issues involving the elements of the data matrix in light of the nested and interrelated nature of TUs and characters. I will argue at the end that if care is taken to construct an appropriate data matrix to address a particular question of relationships at a given level, then simple parsimony analysis is all that is needed to transform the matrix into a tree. Debates over more-complicated models for tree building can then be seen for what they are: attempts to compensate for marginal data.