DOCUMENT RESUME

ED 407 227 SE 059 945

AUTHOR	Brewer, Steven D.
TITLE	Constructing Student Problems in Phylogenetic Tree Construction.
PUB DATE	Mar 97
NOTE	33p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Biology; *Concept Teaching; *Evolution; Fundamental Concepts; Higher Education; Instructional Development; Knowledge Base for Teaching; Models; *Problem Solving; Schematic Studies; Science Education; Teaching Methods
IDENTIFIERS	*Phylogenetics

ABSTRACT

Evolution is often equated with natural selection and is taught from a primarily functional perspective while comparative and historical approaches, which are critical for developing an appreciation of the power of evolutionary theory, are often neglected. This report describes a study of expert problem-solving in phylogenetic tree construction. Results from that study are then used to describe problems in this domain and factors that govern problem difficulty. A problem-based approach to the teaching and learning of evolution was considered. Three series of research problems were constructed that varied the numbers of solutions, taxa, and characters. Each problem consisted of a matrix of coded and polarized phylogenetic data organized by taxa and characters. Nine expert phylogenetic systematists participated in the research project by thinking aloud while constructing phylogenetic trees to account for the problem data matrices. All of the experts agreed that the problems were a realistic characterization of the concepts and processes central to their discipline. Simple tree construction problems such as these allow students to become familiar with the processes used by scientists to explain evolutionary history. The appendix contains a primer of phylogenetic assumptions, diagrammatic elements, and terms. Contains 15 references.(PVD)

Running Head: CONSTRUCTING PHYLOGENETIC PROBLEMS

Constructing Student Problems in Phylogenetic Tree Construction

Steven D. Brewer

Western Michigan University

Constructing Student Problems
in Phylogenetic Tree Construction

This paper describes a research program from the problem solving research tradition in science education to improve the teaching of evolution. Evolution is undoubtedly the most important theoretical framework in biology. Unfortunately evolution is rarely accorded a place in the biology curriculum commensurate with its importance within the discipline. Evolution is often equated with natural selection and is taught from a primarily functional perspective while comparative and historical approaches, that are critical for developing an appreciation of the power of evolutionary theory, are often neglected. This contributes to evolution being poorly understood and widely disparaged among both teachers and American society at large.

A problem-based approach to the teaching and learning of evolution may offer a number of benefits to students. Stewart (1988) has outlined four classes of potential learning outcomes from the use of problem-solving in genetics: (a) the conceptual structure (laws, theories, and their organization) of a particular discipline; (b) problem-solving heuristics that are not specific to a particular discipline; (c) content-specific problem-solving procedures (domain-specific instantiations of general heuristics and problem-solving algorithms specific to the domain); and (d) insight into the nature of science as an intellectual activity. Similar potential learning outcomes are likely from a problem-based approach to the teaching of evolution.

An approach to teaching science developed by the BioQUEST Curriculum Consortium offers greater potential learning outcomes for students than other more traditional approaches (Jungck & Calley, 1985). This approach has been called the "3 P's": problem posing, problem solving, and peer persuasion. To implement this approach successfully, however, a more extensive knowledge base is required than for traditional instruction (Reif, 1983). Teachers must be familiar not only with the conceptual knowledge of a domain, but also the strategic knowledge necessary to engage in effective problem-solving. In addition to solving problems, however, teachers must also have this knowledge organized in a form that will facilitate instruction. To be successful, instruction in solving problems requires a knowledge base composed of at least three bodies of information: (1) conceptual structure that relates tasks to conceptual knowledge, (2) relevant problems that encompass the range of phenomena to be addressed, (3) explicit procedures that include: (a) models of problem solving that can lead to success and (b) strategies and heuristics that can guide how to implement those model across the full range of situations that students may encounter.

Although genetics problem solving and instruction has been relatively well studied from this perspective (See Stewart and Hafner 1995 for a review), other areas of biology have not. This is particularly true of areas that have not traditionally been conceptualized from a problem-solving perspective. This report describes a study of expert problem-solving in phylogenetic tree construction and uses results from that study to describe problems in this domain and factors that govern problem difficulty.

Methods

An initial literature review provided insight into basic phylogenetic problems and methods. Among others, Ridley (1986) and Brooks & McLennan (1991) provided an overview; Eldredge and Cracraft (1980) and Wiley (1981) provided insight into nature of phylogenetic problems and solutions; and Wiley, Siegal-Causey, Brooks, & Funk, (1991) provided a primer of methods. The literature review resulted in: (a) a statement that illustrates the situations that phylogenetic inference is useful for addressing (Table 1), (b) a statement that relates tasks to conceptual knowledge (Table 2), and (c) Phylogenetic Investigator, a software problem-solving environment that was used to present problems to experts (Brewer and Hafner, 1996).

A problem-solving research methodology was developed based on Larkin and Rainard (1984), Ericsson and Simon (1993), and Ericsson and Smith (1995). Three series of research problems (Table 1) were constructed that varied the numbers of solutions,

Table 1

Summary of Research Problems

Series 1 (Ambiguity axis): 5 taxa, 5 characters held constant

Problem	Variables	
1.1	1 solution	1 optimization
1.2	1 solution	2 optimizations
1.3	2 solutions	2 optimizations
1.4	3 solutions	2 optimizations
1.5	4 solutions	2 optimizations

Series 2 (Revision problems): 5 taxa, 5 characters held constant

Problem	Variables
2.1	1 new taxon that "fits" with previous taxa
2.2	2 new characters result in restructured tree
2.3	1 new taxon results in more valid solutions
2.4	1 new taxon results less valid solutions

Series 4 (Character axis): 5 taxa, 1 solution held constant

Problem	Variables
4.1	10 characters
4.2	15 characters
4.3	20 characters

Table 1—Continued

Series 5 (Taxon axis): 10 characters, 1 solution held constant

Problem	Variables
5.1	6 taxa
5.2	8 taxa
5.3	10 taxa

taxa, and characters. Each problem consisted of a matrix of coded and polarized phylogenetic data organized by taxa and characters. In addition, a fourth series of problems contained revision components that required: additions to prior solutions, restructuring of prior solutions, or increased or decreased numbers of solutions. Nine expert phylogenetic systematists participated in the research project by thinking aloud while constructing phylogenetic trees to account for the problem data matrices. The data consisted of the transcripts of the think aloud protocols, segmented according to the actions taking in the drawing environment, and coupled with the finished phylogenetic trees. The data from participant S2 was described separately from the analysis below. This participant accepted solutions in a fundamentally different form that made direct comparisons with the results from other experts difficult..

The strategic knowledge, content knowledge, and knowledge organization that experts use to construct phylogenetic trees were identified from these data. A synthesis of these components was used to create a procedural model of expert performance for phylogenetic tree construction. Understanding how experts go about solving these problems led to new conceptualizations of the nature of phylogenetic problems and a clearer understanding of the factors that give rise to difficulty in problems of phylogenetic tree construction.

Results

Success

No participant found all of the most parsimonious solutions for every problem and some problems contained solutions which were found by no-one. Table 2 lists each topology of each problem on a different line. Each line contains the number of the topology, the number of optimizations for that topology. The number for each participant shows how many optimizations were found for that topology. A zero indicates a topology that was not described. A blank space indicates problems that were not attempted by that participant. The average is calculated by summing the number of optimizations across participants, dividing by number of participants that attempted that problem and dividing that result by the number of possible optimizations. This provides a measure of overall problem difficulty, although it is not entirely valid because some participants were admittedly less concerned with finding all optimizations than others.

Variability Across Participants

Success in problem solving by participant is summarized in Table 3 as the percentage of problems in which a participant found at least 1 most parsimonious topology and the percentage of the total number of topologies that were found by each participant. The best performance in finding at least 1 most parsimonious topology (93%) indicates finding a most parsimonious solution for 14 of 15 problems.

The best performance in finding the total number of topologies (83%) represents finding 19 of 23 possible topologies. All of the subjects were able to find at least one most parsimonious solution to a problem in 60% or more of the problems. Performance was universally lower across topologies than across problems. The largest different between performance across problems and performance across topologies was S7. This may be a reflection of this subject's unique strategy, termed "duplicated taxa", which often provides an efficient path toward a single solution, but does not provide a

Table 2

Number of Optimizations per Topology by Problem and Subject

Problem	Topologies	Optimizations	S1	S3	S4	S5	S6	S7	S8	S9	Average
1.1	1	1	1	1	1	1	1	1	1	0	86
1.2	1	2	2	2	2	2	2	2	1	2	93
1.3	1	2	2	2	2	2	2	2	1	1	86
	2	2	2	2	2	2	2	0	1	1	71
1.4	1	2	0	1	2	2	2	0	0	1	43
	2	2	2	0	0	0	0	0	0	0	14
	3	3	3	3	3	3	2	2	0	1	71
1.5	1	3	2	3	3	3	2	2	0	1	67
	2	2	0	0	0	0	0	0	0	0	0
	3	3	2	0	3	3	0	2	1	2	62
	4	2	2	0	0	0	0	0	0	0	14

Table 2—Continued

Problem	Topologies	Optimizations	S1	S3	S4	S5	S6	S7	S8	S9	Average
4.3	1	6	6	0	1	5	0	0	0	0	29
5.3	1	1	1	1	0	1	0	1	0	1	71
2.1	1	1	1	1	1	1		1	0	1	86
2.1a	1	1	1	1	1	1		1	0	1	86
2.2	1	2	2	1	2	2		2	0	0	64
2.2a	1	2	2	1	0	2		1	1	0	50
2.3	1	1	1	1	1	1		1	1	1	100
2.3a	1	1	0	0	0	0		0	0	0	0
	2	1	1	1	1	0		1	1	0	71
2.4	1	1	1	0	0	0		1	1	0	43
	2	1	1	1	0	1		0	0	1	57
2.4a	1	1	0	1	0	0		1	1	0	43

mechanism for evaluating global hypotheses. The range of performance is slightly larger across topologies than across problems, but the difference is small and may be explained by the fact that S8 rarely looked for alternate topologies.

Table 3

Average Success in Percent by Participant Across Problems and Topologies

	S1	S3	S4	S5	S6	S7[a]	S8[b]	S9
Average across Problems	93	93	73	87	71	93	60	60
Average across Topologies	83	70	61	70	54	65	43	52

[a]S7 used the duplicated taxa strategy. [b]S8 used the order of divergence strategy.

Variability Across Problem Types

The research problems were initially constructed with the goal of studying the variability across the factors which give rise to complexity in phylogenetic tree construction. These factors were seen as increasing numbers of solutions (series 1), characters (series 4), and taxa (series 5). In addition, a series of revision problems (series 2) was posed to assess differences between model-using and model-revising problem solving. Problem 4.3, which had increased numbers of characters proved to be the most difficult problem and was correctly solved by only 3 participants. Finding additional most parsimonious solutions proved to be the most difficult aspect of problem solving and two topologies were found by no-one (topology 2 of 1.5 and topology 1 of 2.3a).

All participants (1, 3, 4, 5, 6, 7, 8, 9) used essentially the same pattern of strategies across problem types. Some heuristics were employed for solving complex problems. The heuristic by participants (4, 5) of listing character distributions and then organizing characters by inclusion/exclusion seemed particularly useful for solving problem 4.3. Several participants (1, 4, 5, 6) used considerations of order in the matrix to enhance the ability to recognize inclusion/exclusion hypotheses more clearly when solving problems that were perceived to be more complex (4.3 and 5.3)

One result of this study is a new conceptualization of problem difficulty as the ratio between actual and potential signal-bearing characters and whether a solution is or is not constructed from the largest inclusion/exclusion character group. It quickly became apparent that the degree of homoplasy in a problem contributed to its difficulty. Initial measures of homoplasy used to evaluate problem difficulty, like the consistency index (CI), proved inadequate. Difficulty is not simply an attribute of problems: finding each topology in a problem represents a unique subproblem and it is at this level that the evaluation of difficulty must take place. The consistency index, or the ratio of tree length and numbers of characters, is the same for all equally parsimonious topologies. Statements by participants (1, 5) suggested that difficulty was a function of the signal to noise ratio. I first eliminated the characters which could contribute no signal to a topology: autapomorphies and whole-ingroup synapomorphies. I then used the remaining characters to calculate the ratio of the number of non-homoplasious characters and the total number of remaining characters. This value, or Signal Index (SI) is inversely proportional to the number of homoplasious characters in a problem.

In problem 1.5, all of the characters (5) can potentially contribute to the signal. Two of the topologies have 3 non-homoplasious characters (2, 3, and 5) yielding an SI of .6. The other two topologies have only two non-homoplasious characters (3 and 5), resulting in a lower SI of .4. Across all of the problems constructed for this research, the SI ranged from 0.26 to 1.00.

Whether or not a topology was based on the largest inclusion/exclusion character group also appeared strongly related to difficulty. This results from the use of the inclusion/phenomenon as a means of approximating the group of most parsimonious solutions. Topologies based on groups of characters other than the largest group were rarely discovered. There were three problems where this occurred: 1.4, 1.5, and 4.3. This idea is described below in more detail as a conceptual model of problem difficulty.

The two factors, Signal Index and whether solutions were based on the largest inclusion/exclusion group were analyzed as factors related to percentage of topologies found by participants using multiple regression with Minitab (Minitab, 1992). Both factors, Signal Index (p=.017) and set membership (p=.031) were statistically significant. Figure 1 illustrates the percentage of experts that found topologies plotted against the signal index for topologies. Regression lines are provided for the condition where a topology is constructed from the largest or some smaller inclusion/exclusion group.

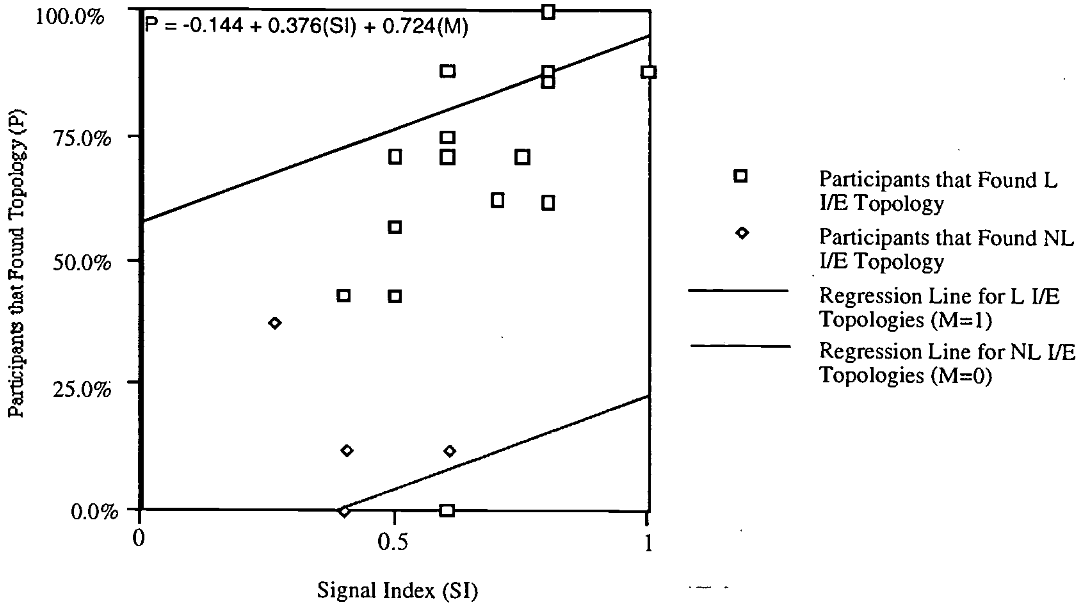


Figure 1. The percentage of participants that found topologies (P) is plotted against signal index (SI). Square symbols are used to represent topologies constructed from the largest inclusion/exclusion (L I/E) group and diamonds are used when a smaller group (NL I/E) was used. Regression lines are plotted that predict performance when the topology is constructed from the largest inclusion/exclusion group (M=1) or not (M=0).

When both factors were considered, the regression function was highly significant (p<.001) and accounted for 59% of the variability (adjusted $R^2 = 54.7$). The regression function was P = -0.144 + 0.376 SI + 0.724 M, where P is the percentage of participants that found a topology, SI is the signal index, and M is the set membership of the topology operationalized as 0 for topologies derived from one of the largest inclusion/exclusion groups, or 1 for topologies not based on a largest inclusion/exclusion groups. One observation, identified as an outlier, resulted from the fact that one topology for problem 2.3a was found by no-one. This case is described in more detail below.

Factors that contribute to variability in performance include: the numbers of characters, taxa, and solutions; the order in which solutions were considered by a participant; practice effects; and fatigue. Most participants were able to find at least 1 most parsimonious topology for most of the problems regardless of these other factors.

The number of characters by itself is not be a good predictor of difficulty. When characters are perfectly consistent, finding the solution is fairly trivial regardless of the number of characters. Increasing numbers of characters in situations of homoplasy, however, magnifies the difficulty of finding a solution.

Increasing the numbers of taxa in a problem did not substantially increase its difficulty in the range used here. As the numbers of taxa increase, the problem becomes more complex and time consuming, but in the case of a set of characters that was perfectly consistent across the taxa, finding a solution would not be difficult.

Increasing the numbers of solutions in a problem by itself does not seem to affect the difficulty in finding a single most parsimonious solution, although finding all of the alternates is extremely problematic. The difficulty of finding any particular solution is described above as a function of whether that topology is within or outside the set of trees defined by the largest inclusion/exclusion hypotheses. However, it is likely that whether a problem has multiple topologies or not and the order in which topologies are found, are factors that interact and contribute to difficulty in complex ways. The minimum length topology for Problem 4.3 was found by three of the eight (38%) participants who attempted this problem. This is substantially higher than the regression function would predict (4%). The fact that this problem has only a single solution, in spite of the fact that there is no solution within the set defined by the largest inclusion/exclusion character groups, may be a factor that contributed to participants finding the solution. Experts are often capable of determining whether a given solution represents a minimum length tree, but are typically unable to tell whether other solutions exist. This means that experts can work purposefully until they find a tree that looks like a minimum length tree. Up to that point, they are usually capable of finding ways to improve the tree. Having found a best tree, however, no expert possessed a systematic means of moving from that best tree to other best trees that were not within the largest inclusion/exclusion group.

A Conceptual Model of Problem Difficulty

All of the participants in this study used “parsimony” as the criterion to optimize trees. The central theme of phylogenetic tree construction procedures, as described here, is that parsimony can not be easily applied to data in any form other than a tree. All of the methods used by experts begin with a method of approximation to generate a first tree and then use parsimony to optimize that tree by evaluating character arrangements. The predominant method used has been termed "inclusion/exclusion."

Inclusion/exclusion hypotheses are structured by taking one character (usually the most inclusive) or a pair of identical characters and then dividing the rest of the characters into three groups based on their relationship to the first character or characters: inclusive, exclusive, or conflicting. Inclusion/exclusion hypotheses are evaluated by comparing the numbers of inclusive/exclusive characters and incompatible characters. Parsimony hypotheses reflect branching arrangements of taxa supported by characters and are evaluated by considering the number of steps, or character state transitions, required to represent an arrangement in tree form. The relationship between the inclusion/exclusion hypotheses and the set of most parsimonious trees determines how effective this strategy of approximation will be at finding some, most, or all of the most parsimonious trees.

For any group of taxa there is a set of possible topologies A. The number of topologies in this set increases exponentially with the number of taxa. All of the participants in the study were interested in finding subset P composed of the topologies of trees that were most parsimonious (i.e. which minimize the number of character state transitions). Participants searched for this subset by finding a related series of subsets defined by all of the character compatibility groups. For simplicity, only two subsets are described here: Subset I is the set of trees defined by all groups of inclusive/exclusive characters and subset L is the set of trees described by the largest groups of inclusive/exclusive characters. These subsets are related to one another by the rule that subset P and L are always subsets of I, but P and L do not necessary intersect.

The intersection or non-intersection between sets of solutions is important for problem solving. There are 5 ways in which P and L can be related: P and L can contain the same set of trees, P can be a subset of L, L can be a subset of P, L and P can intersect incompletely, or L and P can be disjunct.

In the case that all characters are inclusive/exclusive, P and L collapse into each other (Figure 2). Research Problem 1.1 illustrates this condition. Problem 1.1 (Figure 3)

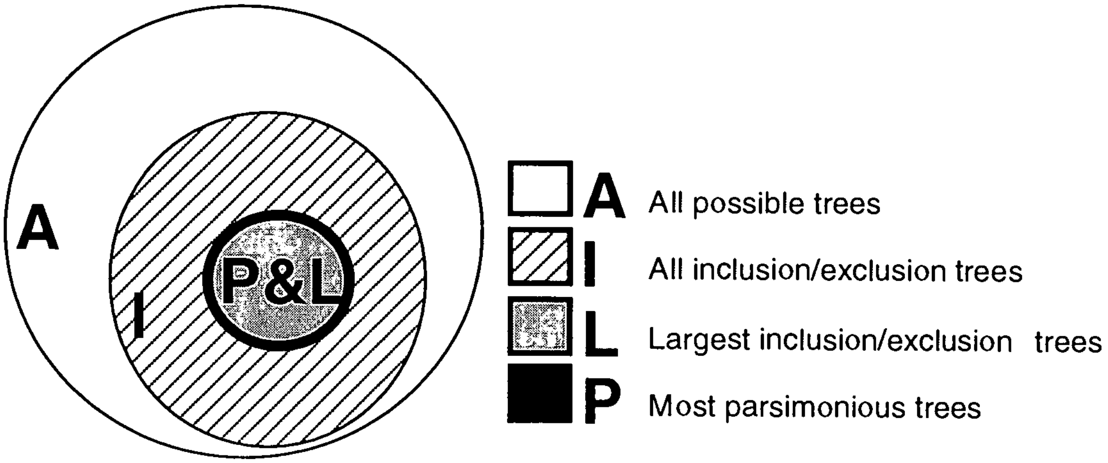

A
I
P&L
□ A All possible trees
▨ I All inclusion/exclusion trees
▩ L Largest inclusion/exclusion trees
■ P Most parsimonious trees

Figure 2. The case in which the set of most parsimonious trees is congruent with the set of largest inclusion/exclusion trees.

has a whole-group synapomorphy defined by character 1. Character 5 is nested within two identical characters, 3 and 4, which are exclusive from character 2. This results in a single tree with no homoplasy. The single most parsimonious solution (Figure 4) requires 5 steps.

SPC	1	2	3	4	5
R80	I	I	0	0	0
R89	I	0	I	I	0
R81	I	0	I	I	I
R82	I	I	0	0	0
R86	I	0	I	I	I
F98	0	0	0	0	0

Figure 3. The data matrix for problem 1.1.

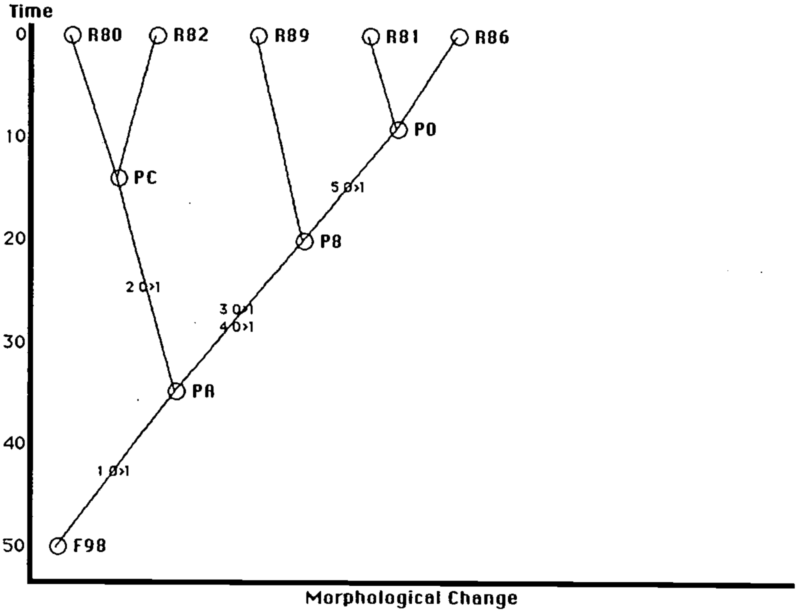


Figure 4. The most parsimonious solution for problem 1.1.

Set L may be a subset of P (Figure 5) or that P is a subset of L (not pictured). In these cases finding L is a useful approximation of P. Finding the P that is distinct from L is still problematic. This condition is illustrated by research problem 1.4 Problem 1.4

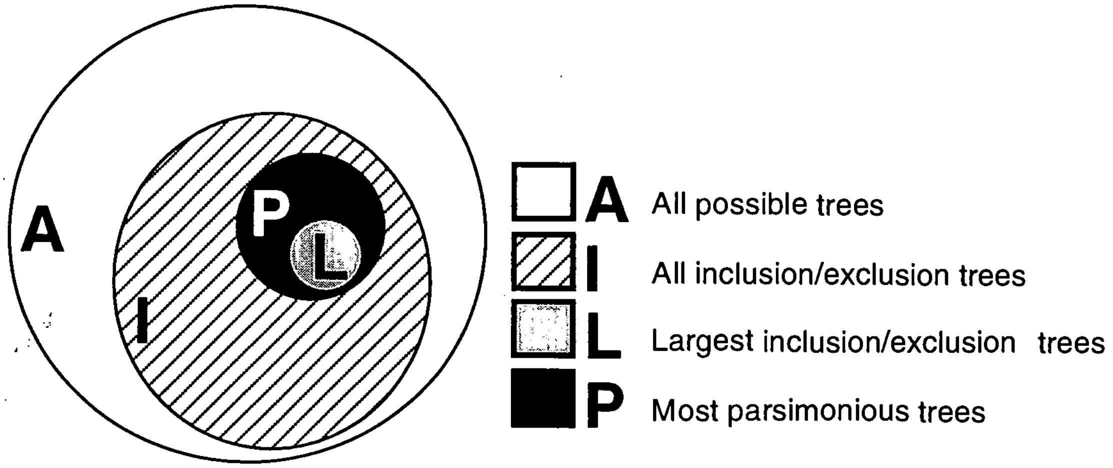

A
P
L
I
A All possible trees
I All inclusion/exclusion trees
L Largest inclusion/exclusion trees
P Most parsimonious trees

Figure 5. The case in which the set of largest inclusion/exclusion trees is a subset of the set of most parsimonious trees.

(Figure 6) has no evident whole-ingroup synapomorphy. Characters 3 and 5 are identical and exclusive from character 2. Character 1 conflicts with character 4. Character 4 conflicts with all other characters. Three topologies are possible: Solution 1 (Figure 7) involves homoplasy in 4 and 1 (4 gained in R83 and the common ancestor of R89 and R81, and 1 either gained in R86 and R81 or gained in the common ancestor of R86, R89 and R81 and lost in R89); solution 2 (Figure 8) involves homoplasy in characters 2 and 4 (character 2 gained in R80 and R83, character gained in the common ancestor of R83, R89, R81, and R86 and then lost in R86); and solution 3 (Figure 9) involves homoplasy only in character 4, which requires 3 gains (in R83, R89, and R81) or two gains and a loss (gained prior to the common ancestor of R89, R86, and R81 and lost in R86). The 3 most

SPC	1	2	3	4	5
R83	0	1	0	1	0
R80	0	1	0	0	0
R86	1	0	1	0	1
R89	0	0	1	1	1
R81	1	0	1	1	1
F98	0	0	0	0	0

Figure 6. The data matrix for problem 1.4.

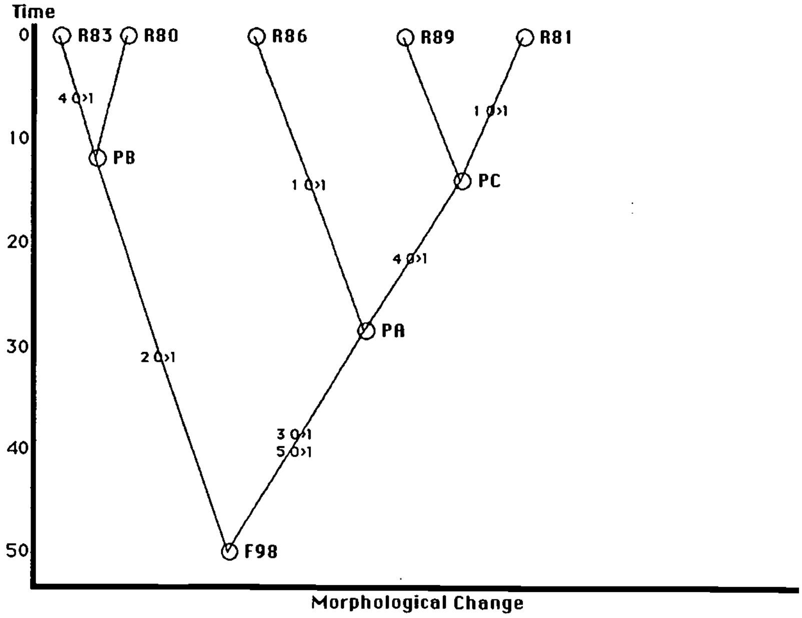


Figure 7. Solution 1 of three equally parsimonious solutions for problem 1.4.

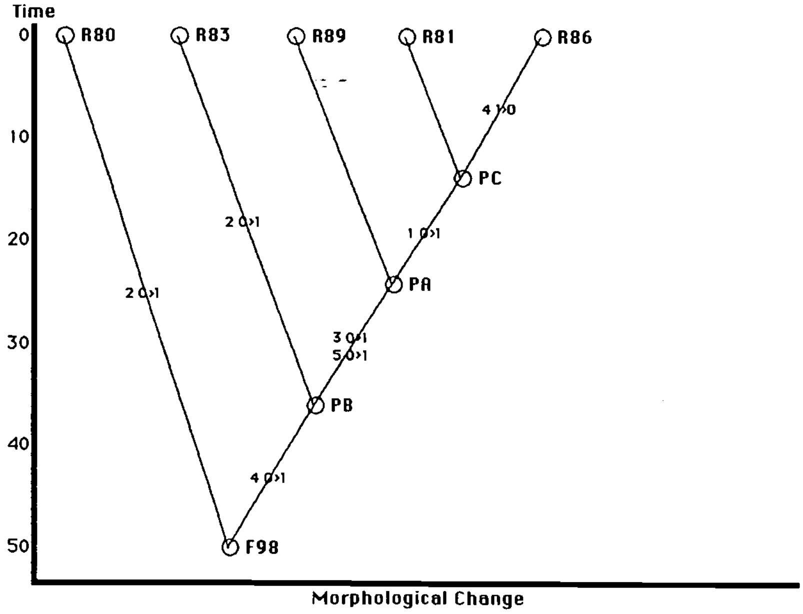


Figure 8. Solution 2 of three equally parsimonious solutions for problem 1.4.

parsimonious solutions require 7 steps. In research problem 1.4 , seven of eight experts found the topology that is derived from the largest inclusion/exclusion group. Of the other two topologies, formed of smaller inclusion/exclusion groups, one was found by five of eight experts and the other found only by one.

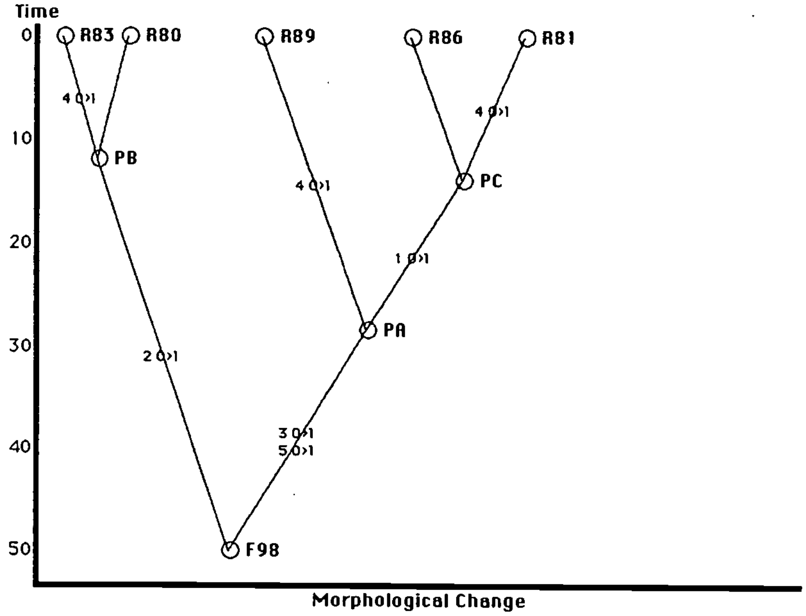


Figure 9. Solution 3 of three equally parsimonious solutions for problem 1.4.

When L is a subset of P, finding P that is separate from L is difficult. The reverse is not true. By its nature, L tends to be easy to explore in its entirety. For example S5, while solving problem 2.4 said:

> So here it's hard to pick ... a set of characters that think you ought to just go with because there are actually sort of two sets and its basically 1 and 2 vs. 4 and 5 and so what I might actually do is draw trees based on those two sets initially and see what they say about each other.

It is exactly this quality that makes finding the largest inclusion/exclusion groups so useful as a form of approximation. If P is a subset of L, it is easy to examine all of L and simply determine which parts are part of P.

Sets P and L may only incompletely intersect (Figure 10). In these cases, inclusion/exclusion analysis reduces the amount of the solution space which needs to be searched to find at least one most parsimonious solution to a problem. As above, it will be difficult to find most topologies that derive from smaller inclusion/exclusion groups while topologies that derive from largest inclusion/exclusion groups that are not most parsimonious can be easily discarded.

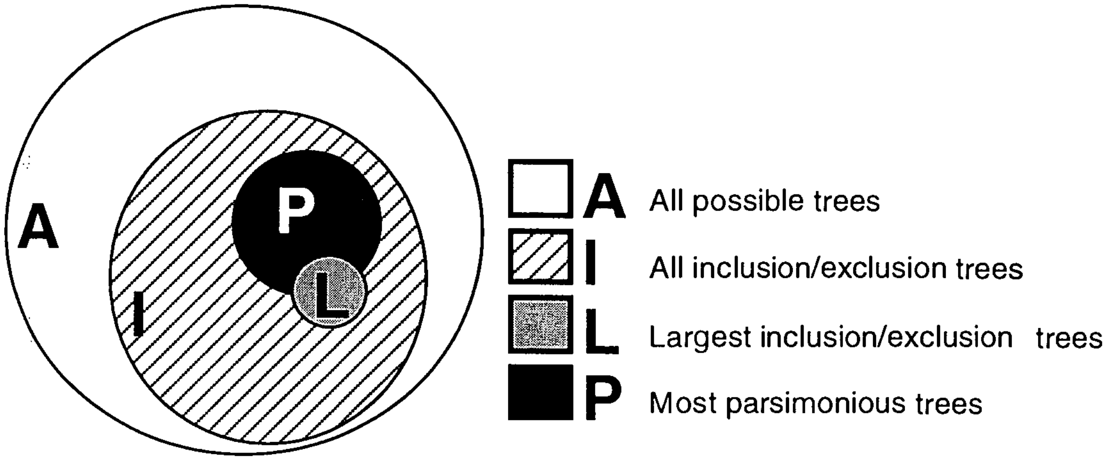

A
P
L
I
A All possible trees
I All inclusion/exclusion trees
L Largest inclusion/exclusion trees
P Most parsimonious trees

Figure 10. The case in which there is incomplete intersection between the set of most parsimonious trees and the set of largest inclusion/exclusion trees.

In the last case, which occurs only under conditions of extreme inconsistency among the data, P and L are completely disjunct (Figure 11). In this case, seeking L will lead to incorrect solutions unless it can be counterbalanced by some other strategy to allow the problem solver to move from this set to the set of most parsimonious trees.

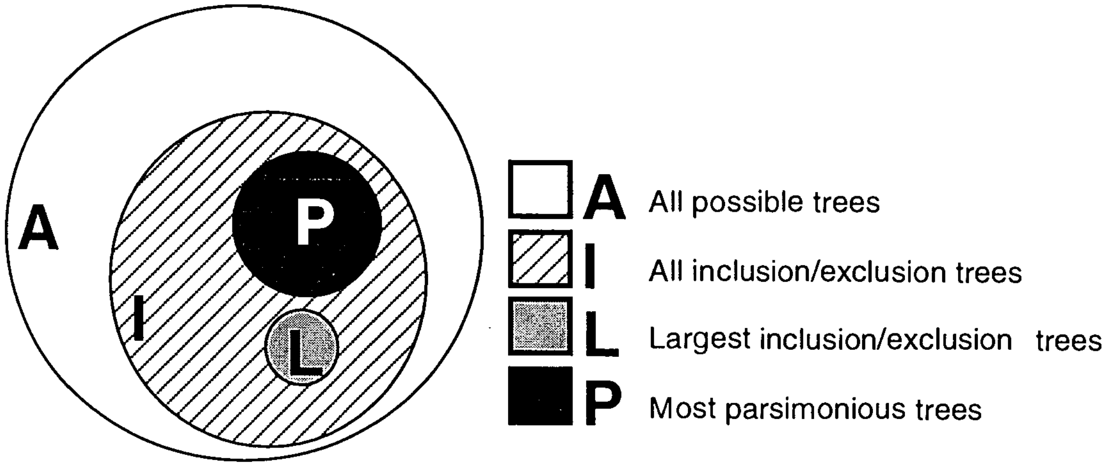

A
P
I
L
A All possible trees
I All inclusion/exclusion trees
L Largest inclusion/exclusion trees
P Most parsimonious trees

Figure 11. The case in which the set of most parsimonious trees is disjunct from the set of largest inclusion/exclusion trees.

SPC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
R89	0	1	0	1	1	1	1	1	0	0	0	1	1	1	1	1	0	0	0	0
R80	1	0	0	1	0	1	0	1	1	1	0	0	1	0	0	1	0	1	0	1
R81	1	1	1	0	0	0	0	1	0	0	1	1	0	0	0	0	1	1	0	0
R82	1	0	0	0	0	1	1	1	1	1	1	0	1	1	0	0	0	0	1	0
R85	0	1	1	0	1	0	0	1	1	0	0	0	0	0	1	0	1	0	1	1
F98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 12. The data matrix for problem 4.3.

Only one of the research problems (4.3) was of this type. Problem 4.3 (Figure 12) has a whole ingroup synapomorphy (character 8). Characters 6 and 13 are identical and exclusive from 3 and 17. Characters 6 and 13 define R89, R80 and R82 with characters 1 (with Homoplasy in R81), 9 (with homoplasy in R85) and 10, defining R80 and R82 as sister taxa. All other characters are homoplasious (Character 2 can be either gained in the common ancestor of R81 and R85 and also in R89 or gained in the common ancestor of the whole ingroup and lost in the common ancestor of R82 and R89. Characters 4 and 16 can be gained in R89 and R80 or gained in common ancestor of R82 and R89 and lost in R82. Characters 5 and 15 must be gained in R85 and R89. Character 7 can be gained in

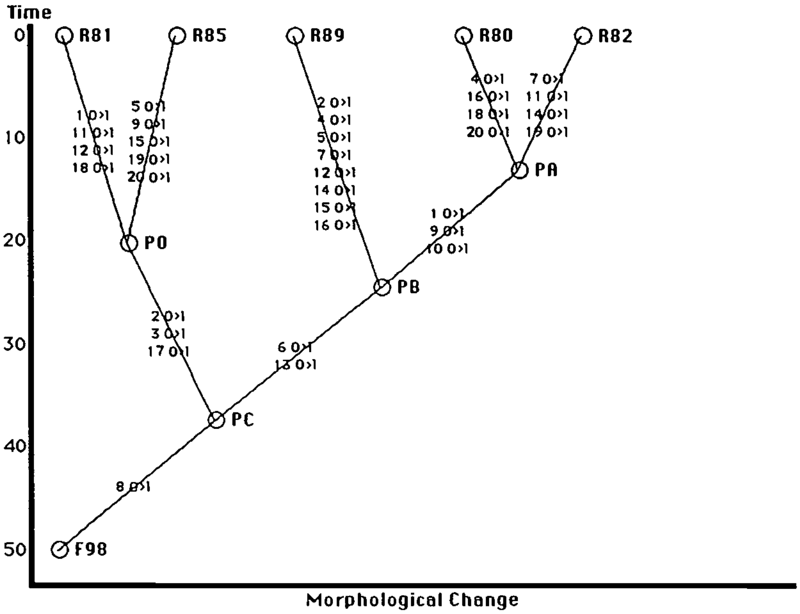


Figure 13. The most parsimonious solution for problem 4.3.

R89 and R82 or gained in common ancestor of R82 and R89 and lost in R80. In problem 4.3 the two largest inclusion/exclusion groups are: {3, 4, 6, 8, 13, 16, 17} and {3, 6, 7, 8, 13, 14, 17}. In the most parsimonious tree (Fig. 13), the characters 4 and 16 from the first group are false and characters 7 and 14 from the second group are false. The most parsimonious tree is based on a smaller character compatibility group composed of {3, 6, 8, 10, 13, 17}.

This model can be used explain differences in performance by experts across topologies with this set of problems (See regression analysis above). The set of problems used in this study was not designed to elicit differences using these categories. Future studies that use this conceptualization of problem difficulty may be able to produce a more fine-grained analysis of the differences in difficulty across problem categories.

Conclusions

This research was conducted with the goal of developing instructional materials for a course that taught domain-specific problem solving using a cognitive apprenticeship approach. Using this approach, the instructor demonstrates problem solving (modeling), helps students solve problems (coaching), and encourages students to solve problems autonomously (fading), until students have developed competence at solving the problems independently. This model allows an instructor to understand what gives rise to difficulty in phylogenetic problems. The instructor can then generate a series of problems that introduce difficulty and problematic phenomena gradually.

Simple tree construction problems such as these allow students to become familiar with the processes used by scientists to explain evolutionary history. All of the experts in the study agreed that the problems were a realistic characterization of the concepts and processes central to their discipline. At the same time, it should be recognized that the processes as presented in this study have been decontextualized and that students, especially the introductory students who might benefit most from solving these problems, should also work with problems constructed from rich data sets including real or realistic imaginary organisms, such as the Caminalcules (Sokal, 1983) or Dendrogrammaceae (Duncan, Philips, & Wagner, 1980). Tree construction is less an end in itself than a means to understanding evolutionary relationships among organisms.

Students with a tree-based conception of phylogenetic biology should be better prepared to understand evolutionary biology and its central role in the rest of biological theory. The ability to see evolution as a branching and historical structure, rather than a ladder or straight line, lies at the heart of much of modern biology. Discarding the ladder-based approach to conceptualizing evolutionary progress may also help students free themselves from the mythos that some organisms are higher or lower than others. This concept, central to understanding the revolutionary power of Darwin's work, is still elusive to many students. Developing a solid foundation of phylogenetic concepts may transform the way many students experience these ideas and help foster a less anthropocentric view of the history of life.

Literature Cited

Brooks, D. R., & McLennan, D. A. (1991). Phylogeny, Ecology, and Behavior. Chicago: University of Chicago Press.

Duncan, T., Phillips, R. B., & W. H. Wagner, J. (1980). A comparison of branching diagrams derived by various phenetic and cladistic methods. Systematic Botany, 5(3), 264-293.

Eldredge, N., & Cracraft, J. (1980). Phylogenetic Patterns and the Evolutionary Process: Method and theory in comparative biology. New-York: Columbia University Press.

Ericsson, K. A., & Simon, H. A. (1993). Protocol Analysis: Verbal reports as data. Cambridge: MIT Press.

Ericsson, K. A., & Smith, J. (1991). Empirical studies of expertise: Prospects and limits. In K. A. Ericsson & J. Smith (Eds.), Toward a general theory of expertise: Prospects and limits (pp. 1-38). Cambridge: Cambridge University Press.

Jungck, J. & Calley, J. (1985). Strategic simulations and post-Socratic pedagogy: Construction computer software to develop long-term inference through experimental inquiry. American Biology Teacher. 14(2), 137-146.

Larkin, J., & Rainard, B. (1984). A research methodology for studying how people think. Journal of Research in Science Teaching, 21(3), 235-254.

Reif, F. (1983). Understanding and teaching problem solving in physics. In Lectures at the International Summer School on Physics Education, La Londe Les Maures, France:

Ridley, M. (1986). Evolution and Classification: The reformation of cladism. New York: Longman Group Limited.

Sokal, R. R. (1983a). A phylogenetic analysis of the Caminalcules: I. The data base. Systematic Zoology, 32(2), 159-184.

Sokal, R. R. (1983b). A phylogenetic analysis of the Caminalcules: II. Estimating the true cladogram. Systematic Zoology, 32(2), 185-201.

Stewart, J. (1988). Potential learning outcomes from solving genetics problems: A typology of problems. Science Education, 72(2), 237-254.

Stewart, J., & Hafner, R. S. (1994). Research on problem solving: Genetics. In D. L. Gable (Eds.), Handbook of research on science teaching and learning (pp. 284-200). New York: MacMillan Publishing Co.

Wiley, E. O. (1981). Phylogenetics: The principles and practice of phylogenetic systematics. New York: John Wiley & Sons.

Wiley, E. O., Siegal-Causey, D., Brooks, D. R., & Funk, V. A. (1991). The Compleat Cladist (Special Publication No. 19). University of Kansas Museum of Natural History.

Appendix A

A Primer of Phylogenetic Assumptions, Diagrammatic Elements, and Terms

A Primer of Phylogenetic Assumptions, Diagrammatic Elements, and Terms

Assumptions of Phylogenetic Inference:

1. There is only one true phylogeny.
2. Shared characters are the result of homology.
3. The polarity of character states is knowable.

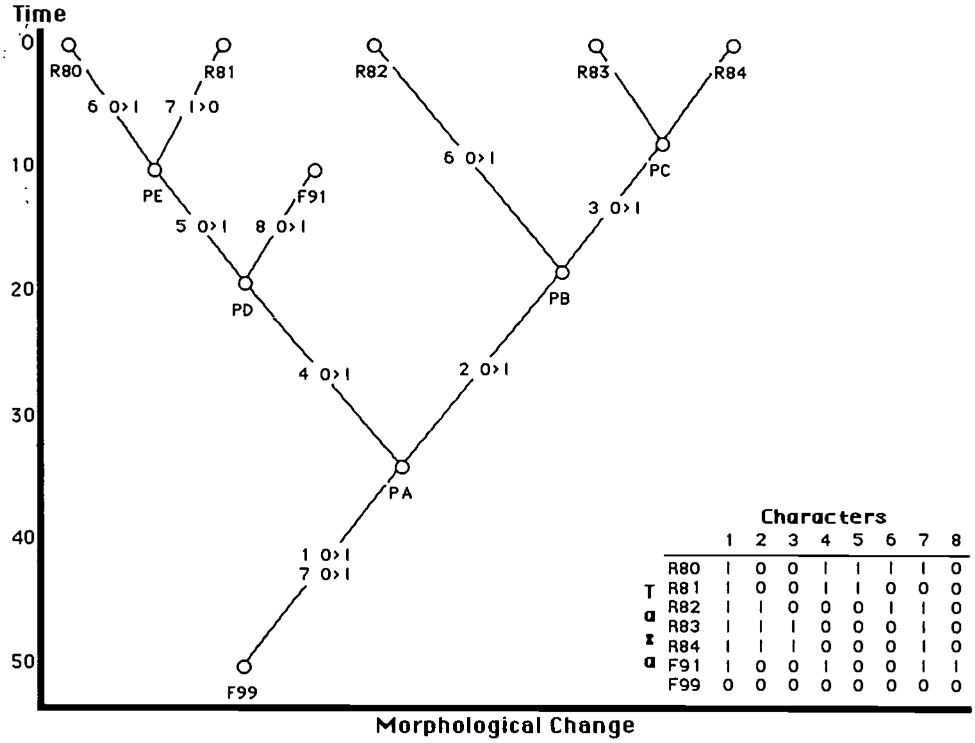
Elements of Phylogenetic Diagrams

Figure 1 illustrates an example phylogenetic tree created using Phylogenetic Investigator. This section describes the phylogenetic tree and its elements. Terms are organized alphabetically at the end with definitions and examples that also reference this tree where possible.

The data matrix from which this diagram is generated appears in the lower right hand corner showing characters in columns and taxa in rows. The intersection between each row and column has a symbol that indicates where that taxon has the apomorphic (1) or plesiomorphic (0) form of the character.

The phylogenetic tree is constructed along two axes. The ordinate represents time divided into 50 units and the abscissa represents morphological change as a continuous, unitless variable. The small circles are nodes. Each node has a designation associated with it. Nodes that begin with "R" represent recent taxa. Nodes that begin with "F" represent fossil taxa. Nodes that begin with "P" are postulated taxa. Lines that link nodes together indicate lines of ancestor/descendant relationship. Some links contain one or more transitions. Each transition (e.g., "1 0>1" or "1 1>0") indicates that the referenced character (1) changed in state either from plesiomorphic to apomorphic (0>1) or reversed from apomorphic to plesiomorphic (1>0) at some point in time along the link on which it appears.

In Figure 1, characters 1-5 are represented as being homologous. Characters 6 and 7 are homoplasious in this diagram. Character 8 is an autapomorphy and is irrelevant to the decision-making process of tree construction. An autapomorphic character is always constructed as a transition immediately prior to the taxon that possesses it.



Characters

Taxa	1	2	3	4	5	6	7	8
R80	1	0	0	1	1	1	1	0
R81	1	0	0	1	1	0	0	0
R82	1	1	0	0	0	1	1	0
R83	1	1	1	0	0	0	1	0
R84	1	1	1	0	0	0	1	0
F91	1	0	0	1	0	0	1	1
F99	0	0	0	0	0	0	0	0

Figure 1. An Example Phylogenetic Tree

Character 1 is a whole-group synapomorphy that supports the existence of postulated ancestor PA. Character 1 is inclusive of all other characters. Character 2, which groups R82, R83 and R84, supports node PB. Character 2 is inclusive of character 3 and exclusive of character 4. Character 3, which groups R83 and R84, supports node PC. Character 4, which groups R80, R81, and F91, supports node PD. Character 4 is inclusive of character 5 and exclusive of character 2. Character 5, which groups R80 and R1, supports node PE.

Character 6 claims that R80 and R82 are a group. For character 6 to be true, characters 2, 5, and 4 would have to be false. In other words, in order to save one step in character 6, at least three other steps would required. Character 6 is most parsimoniously gained convergently in R80 and R82. Character 7 claims that all of the taxa except for R81 are a group. For character 7 to be true, characters 4 and 5 would have to be false. Saving a step in character 7 would result in at least two added steps Character 7 is most parsimoniously optimized as a reversal in R81.

Terms of Phylogenetic Inference

Ancestor

A taxon, previous in time to a second taxon, from which the second taxon is descended. For example, Figure 1 proposes that a postulated taxon PC is the common ancestor of R83 and R84.

Apomorphy

An evolutionary character, usually coded as "1", that represents an evolutionarily novel state. Character 1 is an apomorphy in all of the taxa of the ingroup (Fig. 1).

Autapomorphy

The transition of a character that is uniquely evolutionarily novel (apomorphic) for a taxon. Character 8 an autapomorphy because it is possessed in the apomorphic state only by taxon F91 (Fig. 1).

Character

A recognizable feature that varies among taxa. For example, among ladybugs, the characters might include the presence or absence of spots. Characters are numbered, polarized, coded, and presented in columns in the data matrix (Fig. 1).

Clade

A monophyletic taxon.

Cladogram

A form of phylogenetic tree that can only show sister-group relationships. Figure 1 illustrates sister-group relationships between all of the taxa, except F99, which is claimed to be a true ancestor of all of the other taxa.

Conflict

A quality of characters that contain incompletely overlapping distributions of apomorphies. Characters 5 and 6 conflict because both are apomorphic for 80, but 5 is apomorphic for 81 and 6 is apomorphic for 82 (Fig. 1).

Convergence

A form of homoplasy whereby two taxa share a character that has appeared independently in separate lineages. Character 6 arises convergently in taxa R80 and R81 (Fig. 1).

Data Matrix

A summary table of states with taxa in rows and characters in columns. The data matrix appears in the lower right-hand corner (Fig. 1).

Descendant

A taxon which is the genealogical product of an earlier taxon. Taxon R84 is a descendant of PC (Fig. 1).

Exclusive
Characters whose distributions of apomorphies do not overlap. Characters 2 and 4 are exclusive of one another (Fig. 1).

Homology
The quality of characters that are shared as the result of common ancestry. See assumption 2. Characters 1, 2, 3, 4, 5 are assumed to be homologous (Fig. 1).

Homoplasy
Characters that are shared due to causes other than homology (evolutionary convergence or reversal). Character 6 is homoplasious and explained using convergence and character 7 is homoplasious and explained using reversal (Fig. 1).

Inclusive
When one character's distribution of apomorphies is a superset of another character's distribution of apomorphies. Character 2 is inclusive of character 3 (Fig. 1).

Ingroup
The group of taxa currently being studied using phylogenetic inference. Taxa R80, R81, R82, R83, R84 and F91 are members of the ingroup (Fig. 1).

Link
A line in between nodes in Phylogenetic Investigator that represents lines of ancestor/descendant relationships. The link between R83 and PC represents a hypothetical ancestor/descendant relationship between R83 and PC (Fig. 1).

Monophyletic
A taxon that includes only the complete set of descendant taxa of an ancestral species. The group of R83 and R84 (and PC) is a monophyletic taxon (Fig. 1).

Node
A circle in Phylogenetic Investigator used to represent a taxon. R80 is a node that represents a taxon (Fig. 1).

Optimization
The process or product of distributing a homoplasious character on a phylogenetic tree. Characters 6 and 7 are optimized in Figure 1.

Outgroup
A group of taxa used to polarize the character states.

Parallelism
A convergence.

Paraphyletic
A grouping of taxa that does not reflect the underlying evolutionary relationships by removing taxa from a monophyletic taxon. A grouping of R82 and R84 is paraphyletic (Fig. 1).

Parsimony
A principle used to justify selecting the hypothesis that requires the fewest transitions and a corollary to assumption 2: By assuming homology, one also selects the hypothesis that minimizes the number of assumptions of homoplasy. The phylogenetic tree in Figure 1 is the most parsimonious explanation of the data.

Phylogenetic tree
A branching diagram that can illustrate both sister group and ancestor/descendant relationships among a set of taxa. Figure 1 is a phylogenetic tree.

Phylogeny

The set of ancestor/descendant relationships that form the genealogy of a set of taxa. A phylogenetic tree (Fig. 1) is a hypothetical representation of these relationships.

Plesiomorphy

A form of a character (state) which is evolutionarily preexisting for the group of taxa under study (the ingroup). Character 2 is retained in the plesiomorphic state by R80, R81, and F91 (Fig. 1). Character 7 occurs in the plesiomorphic state in taxon R81 and this is explained using a hypothesis of reversal (Fig. 1).

Polarity

Whether a form of a character (a state) is considered apomorphic (evolutionary novel) or plesiomorphic (evolutionarily preexisting). This is usually done through comparison with an outgroup.

Polyphyletic

A grouping of taxa that does not reflect the underlying evolutionary relationships by adding unrelated taxa to a monophyletic taxon. A grouping of R81, R83, and R84 would be polyphyletic (Fig. 1)

Reversal

The transition of a character that is apomorphic in some ancestor, changes polarity back to the plesiomorphic state resulting in descendant taxa which are plesiomorphic for that character. Character 7 is optimized as a reversal in taxon R81 (Fig. 1).

Sister group

The most closely related taxon to another taxon. R82 is the sister group to the taxon of R83 and R84 (Fig. 1)

State

A form of a character that is polarized as either apomorphic or plesiomorphic and coded as "1" or "0". For example, among ladybugs, the absence of spots might represent the plesiomorphic state and the presence of spots might represent the apomorphic state.

Steps

The number of transitions required to explain a character or characters. Character 6 is explained in two steps (Fig. 1).

Synapomorphy

The transition of a character that is homologously shared in the evolutionary novel (apomorphic) condition. Character 1 is a synapomorphy for the whole ingroup (Fig. 1).

Taxon

A group of organisms that is given a name. The complete set of taxa descended from a common ancestor is a monophyletic taxon. Incomplete sets are paraphyletic and sets with extra unrelated taxa are polyphyletic. R80, R81, and F91 are a monophyletic taxon because they all are hypothesized to have descended from PD (Fig. 1).

Topology

An arrangement of sister-group or ancestor/descendant relationships among a group of taxa. Figure 1 has only one most parsimonious topology—any rearrangement of the relationships among the taxa would require more steps than the current tree to explain all of the characters.

Transition — A point in time in a lineage at which a character is hypothesized to have changed in state. At some point between 20 and 35 units of time before the present, character 4 is hypothesized to have changed in state in taxon PD (Fig. 1).

Treelength — The steps, or number of transitions, required to explain the data matrix using a phylogenetic tree. Figure 1 requires a treelength of 10 steps to most parsimoniously explain the data in the matrix.

AERA CONFERENCE 1997

SE059945

ERIC

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Constructing Student problems in phylogenetic tree construction

Author(s): Steven D. Brewer

Corporate Source:

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☒

Check here
For Level 1 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

The sample sticker shown below will be affixed to all Level 2 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

☐

Check here
For Level 2 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature:	Printed Name/Position/Title: Steven D. Brewer; Director BCRC	
Organization/Address: Biology Department UMASS Amherst, MA 01003	Telephone: 413 545 2272	FAX: 413 545 3243
	E-Mail Address: sbrewer@bio.umass.ed	Date: 3/25/97

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on Assessment and Evaluation
210 O'Boyle Hall
The Catholic University of America
Washington, DC 20064

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2d Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com