Palaeontologia Electronica
http://palaeo-electronica.org

StrataPhy: A NEW COMPUTER PROGRAM FOR STRATOCLADISTIC ANALYSIS

Jonathan D. Marcot and David L. Fox

Jonathan D. Marcot. Department of Geology and Geophysics, University of Minnesota, 310 Pillsbury Drive SE, Minneapolis, MN, 55455-0219
Current address: Department of Animal Biology, University of Illinois at Urbana-Champaign, 505 S. Goodwin Ave., Urbana, IL, 61801. jmarcot@life.uiuc.edu
David L. Fox. Department of Geology and Geophysics, University of Minnesota, 310 Pillsbury Drive SE, Minneapolis, MN, 55455-0219. dlfox@umn.edu

ABSTRACT

StrataPhy is a computer program designed to perform stratocladistic analysis. Stratocladistic analysis minimizes *ad hoc* hypotheses of both character homoplasy and non-preservation in the fossil record allowing for the simultaneous analysis of morphologic and stratigraphic data. Prior to StrataPhy, stratocladistic analyses required multiple computer programs and manual branch arrangement. StrataPhy employs full TBR branch swapping coupled with an integrated search for the optimal assignment of taxa as ancestors. The algorithms involved in a StrataPhy stratocladistic search are discussed in detail. StrataPhy reads standard NEXUS formatted files, and additional formatting required for a StrataPhy analysis is described. We also describe a reanalysis of a previously published data set that emphasizes the potential utility of StrataPhy over previous approaches to stratocladistic analysis.

KEY WORDS: computer program, software, phylogenetic inference, phylogeny, stratocladistics

INTRODUCTION

Phylogenetic analysis is fundamental to many modern evolutionary studies of fossil taxa. Not only does an understanding of the evolutionary relationships among taxa contribute to our understanding of biotic diversity in general, but phylogenetic trees themselves also have become a useful component to the analysis of many other aspects of biotic diversity and evolution (Harvey and Pagel 1991; Huelsenbeck and Rannala 1997) including morphologic evolution (e.g., Wagner 1996; Stockmeyer Lofgren et al. 2003), taxonomic diversification (e.g., Slowinski and Guyer 1993; Sanderson and Donoghue 1996), and biogeography (e.g., Lieberman 2005; Ree et al. 2005). Needless to say, the accu-

PE Article Number: 11.1.5A
Copyright: Paleontological Society March 2008
Submission: 24 September 2007. Acceptance: 4 February 2008

Marcot, Jonathan D. and Fox, David L. 2008. StrataPhy: A New Computer Program for Stratocladistic Analysis. *Palaeontologia Electronica* Vol. 11, Issue 1; 5A:16p;
http://palaeo-electronica.org/2008_1/142/index.html

racy of phylogenetic inferences is critical to the success of these analyses (Wagner 1998, 2000).

Traditional maximum-parsimony cladistic analysis (henceforth "cladistics") seeks the set of relationships (summarized using a branching diagram or cladogram) that minimizes the number of *ad hoc* hypotheses of character evolution (e.g., Farris 1983). These *ad hoc* explanations are in the form of homoplasy (i.e., convergence, parallelism, or reversal), in which characters exhibit more than the minimum number of state changes required given the distribution of character states among taxa. The total number of these *ad hoc* hypotheses becomes the currency by which cladograms may be evaluated, and the number of hypotheses of homoplasy beyond the hypothetical minimum (i.e., zero, in the case of no homoplasy) has been referred to as the total parsimony debt (Fisher 1982, 1992). In practice, cladistics employs simple algorithms to calculate the minimum parsimony debt for a given data set of characters on a particular cladogram. This minimum can be compared among competing cladograms for the same set of taxa, and the cladogram (or cladograms) exhibiting the minimum debt is chosen as optimal.

Stratocladistics (Fisher 1992, 1994; Clyde and Fisher 1997; Fox et al. 1999; Bodenbender and Fisher 2001) was developed as an extension of cladistics to make use of the temporal order of taxa in the fossil record as data in phylogenetic analysis. It simultaneously considers both character data and stratigraphic data in the form of the intervals of first and last appearance of taxa to be analyzed. Stratocladistics subscribes to the identical philosophy as cladistics; both seek the phylogenetic hypothesis requiring the minimum number of *ad hoc* hypotheses. As in cladistics, *ad hoc* hypotheses include homoplasy, but stratocladistics also considers instances of nonpreservation of particular taxa during intervals in which other taxa are preserved (henceforth "gaps") to be *ad hoc* explanations. In this way, the total parsimony debt incurred by a tree is the sum of its character debt and stratigraphic debt.

Phylogenetic hypotheses (i.e., trees) make predictions about the distribution of taxa in the fossil record (e.g., Novacek and Norell 1982; Smith 1988; Norell 1992, 1993). For example, taxa that originate early in a clade's history should be found relatively early in the fossil record. If not, then a hypothesis of nonpreservation must be stated or implied to account for a taxon's absence from the early part of the record. Stratocladistics considers stratigraphic intervals in which a lineage is expected to be present based on the topology of the phylogenetic tree but is, in fact, not observed to be evidence against a phylogenetic hypothesis. In such a case, the implied gap is a result of the (perhaps provisional) acceptance of the tree over other trees that may not imply the same gap. The tree (or trees) that simultaneously minimizes *ad hoc* hypotheses of nonpreservation, as well as those of character homoplasy, is considered optimal (Fisher 1992). See Bodenbender and Fisher (2001) for a more extensive description of stratocladistic analysis.

Stratocladistics differs from traditional cladistics in another important way in that stratocladistics allows for taxa to be designated as ancestral to other taxa if that arrangement reduces the instances of nonpreservation without invoking more instances of homoplasy. In doing so, the resulting trees are *phylogenetic trees* that describe the ancestor-descendant relationships among taxa, whereas *cladograms* from cladistic analysis are only diagrams that depict recency of common ancestry and sister group relationships (see Hull 1979), or "hierarchies founded on homology hypotheses" (Brochu et al. 2001, p. 174). These evolutionary trees offer more specific hypotheses of the evolutionary relationships among taxa and are therefore more easily refuted with additional data (Fox et al. 1999). Foote (1996) suggests that the incidence of ancestors in the fossil record is not negligible; under reasonable models of evolution and preservation, at least 1-10% of known fossil taxa are likely to be direct ancestors. Therefore, this distinction between cladograms and evolutionary trees that explicitly include ancestral taxa has important implications for those studies that use either type of trees in the analysis of biologic diversity or character evolution. For example, Wagner (2000) demonstrated that failure to recognize ancestral taxa properly can mislead metrics intended to measure the quality of the fossil record from model phylogenies. Lane et al. (2005) have likewise demonstrated how the misidentification of ancestral taxa as sister taxa, as certainly happens with an unknown frequency in cladistic analyses, can lead to overestimates of past taxonomic richness. Stratocladistics is currently one of the few phylogenetic methods that can operationally identify ancestral taxa and thus holds considerable promise for our understanding of the evolutionary history of fossil organisms.

Despite the potential promise of stratocladistics, some issues have been raised that are yet unresolved (e.g., Nelson 1978; Smith 2000; Sum-

rall and Brochu 2003). Criticisms of the use of stratigraphic data in phylogenetic analysis are typically philosophical or theoretical in nature. Examples include whether stratigraphic data constitute phylogenetic or nonphylogenetic data (Sumrall and Brochu 2003) or positive or negative evidence (Heyning and Thacker 1999). However, others argue that any data about which a hypothesis makes predictions are appropriate for testing that hypothesis, and that we cannot limit analysis to only data that support hypotheses, or only those that contradict them (Wagner 2000).

Despite this debate, the general use and evaluation of stratocladistic methods have been hampered by the lack of an efficient, automated means of performing analyses (Fisher 1992; Fox et al. 1999; Bodenbender and Fisher 2001; Fisher and Bodenbender 2003; Sumrall and Brochu 2003). PAUP* (Swofford 2002) is the most widely used computer program for cladistic analysis, but it does not support stratigraphic data, nor does it search for optimal assignments of ancestors. MacClade (Maddison and Maddison 2005) supports stratigraphic data and can be used to perform searches for optimal assignments of ancestors on single trees, but its branch swapping capabilities are limited and do not include simultaneous searches for optimal assignments of ancestors. To date, published stratocladistic searches (Fox et al. 1999; Bloch et al. 2001; Bodenbender and Fisher 2001; Geisler and Uhen 2005) have been restricted to piecemeal analyses that involve iterations of traditional cladistic analysis using PAUP*, followed by additional manual branch swapping and ancestor assignment on the resulting trees (with the stratigraphic character included) for more optimal ones using MacClade (Fisher 1992). In doing this approach**,** the critical step of searching for optimal ancestor assignments is separated from that of branch swapping.

Although heuristic search strategies are never guaranteed to consider all possible solutions exhaustively, this manual search strategy is considerably limited because these components of the search are decoupled operationally, and it is quite likely that the optimal topology and assignment of ancestors might not be visited at all. Furthermore, the order in which taxa are assigned as ancestors can affect the debt incurred or saved by assignments of subsequent taxa as ancestors. Therefore, to find the optimal set of taxa assigned as ancestors, more than one sequence of taxon assignments should be performed. The number of possible sequences can be quite large when the ingroup includes enough taxa to be a computationally intensive phylogenetic problem, which is increasingly typical for analyses including fossil taxa. For such problems, if trees are evaluated one ancestral assignment at a time, the numbers of sequences that can reasonably be searched by hand is necessarily small. The vast number of possible combinations of trees and ancestral assignments render the manual search inexact and impractical for even small datasets, and an automated search is necessary.

Here, we present a new computer program entitled StrataPhy to perform full stratocladistic searches. StrataPhy is available from Appendix I or at the author's site www.life.uiuc.edu/marcot/strataphy/. We begin by describing the algorithm used to perform these searches, then describe how StrataPhy can be used in conjunction with other phylogenetic software to produce files for analysis, and finally, discuss analysis parameters in StrataPhy that are modifiable by the user in the current version. Future releases of StrataPhy will include additional features, which will be discussed briefly in the conclusions of this paper.

TREE SEARCH ALGORITHM

A more detailed description follows, but, in general, the tree search algorithm in StrataPhy consists of first constructing a starting tree, then rearranging it and simultaneously searching for optimal assignments of taxa as ancestors to find more optimal trees (Figure 1). After each rearrangement and assignment of ancestors, StrataPhy compares the debt of the new tree to the total debt of the optimal tree for the analysis so far. If the new tree has the same (optimal) debt it is saved for future swapping; if it has a lower debt, it is saved, and the old trees are deleted. StrataPhy then begins rearranging this new optimal tree. The search concludes when all rearrangements of all optimal trees have been made, without finding a more optimal one.

The tree search algorithm of StrataPhy is essentially a tree bisection and reconnection (TBR) branch swapping routine (see Swofford et al. 1996; Felsenstein 2004 for a more detailed description), with modifications to include searches for the optimal assignment of taxa as ancestors. StrataPhy constructs the initial starting tree using a random stepwise taxon addition sequence (Swofford et al. 1996; Felsenstein 2004), sequentially adding randomly selected taxa in the optimal (most parsimonious, considering morphologic and stratigraphic debt) location on the growing tree, until all taxa are

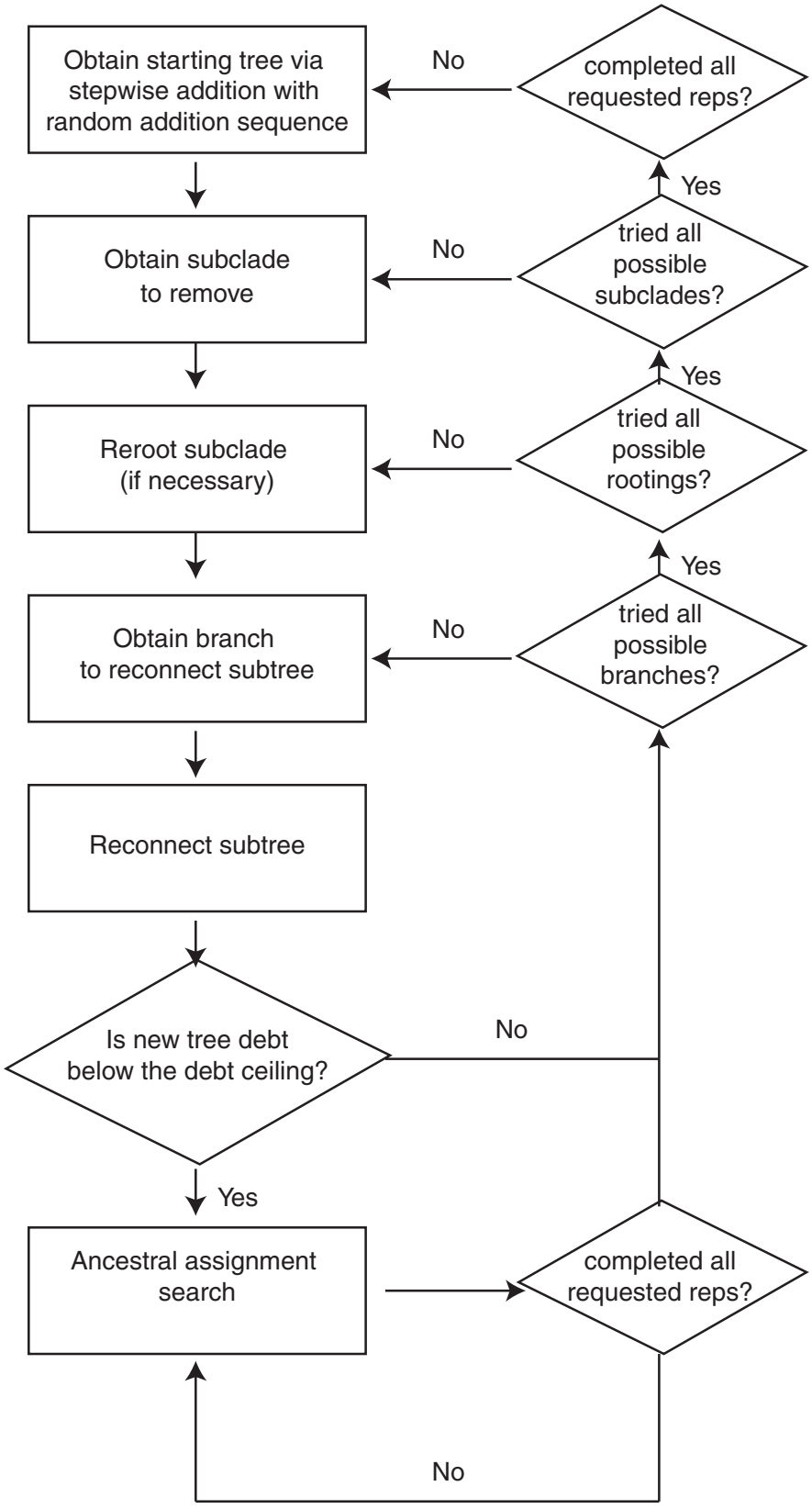

Figure 1. Flowchart demonstrating the stratocladistic tree search algorithm employed in StrataPhy. The box in the upper left represents the beginning of the algorithm.

added (Figure 2.1). The starting tree is rearranged using TBR branch swapping, which begins by breaking the initial tree into two subtrees (Figure 2.2 and 2.3). One of the subtrees is then attached systematically to each possible branch of the other subtree (Figure 2.4). Each time the subtree is reattached, a new topology is created, and the search for the optimal assignment of ancestors begins (described below; Figure 2.5), and the debt of the new tree is compared to the debt of the current optimal tree. When one subtree has been reattached to all possible branches of the other, the former is rerooted iteratively (Figure 2.6), and then tried on all branches on the latter subtree again.

As mentioned above, once the starting tree has been rearranged via TBR, StrataPhy then searches for optimal assignments of taxa as ancestral to others. StrataPhy offers two types of ancestral assignment searches: exhaustive and heuristic. The exhaustive search tries every possible combination of taxa fixed as ancestors or as terminal taxa. This requires 2^n combinations for each branch swap, where *n* is the number of taxa in the tree. It is therefore computationally intensive and very time consuming, but guaranteed to yield all optimal assignments of ancestral taxa. The heuristic ancestral assignment search is computationally faster because it excludes many possible combinations that are necessarily less optimal (i.e., increase debt). In this heuristic ancestor search algorithm, the lengths of the tree with a given taxon fixed as an ancestor and as a terminal taxon are compared. If the assignment as an ancestor increases the total debt, then no further combinations are attempted with that taxon fixed as an ancestor on that topology; the taxon is left as a terminal taxon, and the next taxon is tested. If it decreases total debt, the taxon is fixed as an ancestor, and the next taxon is evaluated; the combinations in which such taxa are not designated as ancestors are disregarded. It is possible that the assignment of a taxon as an ancestor incurs no additional total debt, either because the stratigraphic debt savings and the morphologic debt incurred are equal, or because they are both zero. In either event, the search continues on two separate trees with and without the taxon assigned as an ancestor.

In practice, the difference in debt when a taxon is fixed as an ancestor or as a terminal taxon is dependent on the ancestral status of taxa at surrounding nodes. The order in which taxa are tested as ancestors can change the effect the assignment has on total debt. To account for this, StrataPhy includes a user option that allows multiple random taxon addition sequences for testing taxa in ancestral positions, in a similar manner as when constructing the initial stepwise addition tree. In other words, for a single tree, several separate replicate attempts at fixing all taxa as ancestors can be performed, each with a different and random order in which taxa are evaluated. Multiple replicates increase the chances of the optimal assignment of ancestors being encountered, although, as in any heuristic algorithm, do not guarantee the optimal solution will be found.

Although faster than the exhaustive ancestral assignment search, the heuristic search is still computationally intensive. StrataPhy therefore employs a debt ceiling (Fisher 1992) to minimize the number of cladograms that are searched for optimal ancestors in this manner. When a single TBR search produces a new candidate topology that could be searched for optimal assignment of ancestors, StrataPhy determines the maximum stratigraphic debt savings possible for that candidate topology. If the difference between the total debt of the candidate topology (without fixed ancestors) and that of the globally optimal tree for the current search replicate is greater than the maximum possible stratigraphic savings, then the candidate topology cannot possibly be shorter than the optimal tree. In other words, the maximum debt savings possible by assigning taxa as ancestors could not possibly reduce the debt lower than the current optimal debt. At this point, StrataPhy abandons the ancestral assignment routine, and this topology is discarded. StrataPhy empirically determines the maximum possible stratigraphic savings with an algorithm that sequentially fixes as ancestors all taxa that save stratigraphic debt, without regard to morphologic debt incurred.

At the conclusion of each ancestral assignment search, the length of the current tree is compared to that of the best found over the entire search. As in other phylogenetic inference programs, if it is less than or equal to the best, it is saved, and a new round of TBR branch swapping begins on this new optimal tree. This amounts to a “hill-climbing” routine that swaps optimal trees until no better arrangements are found, discarding suboptimal ones along the way. The routine ends when the optimal trees cannot be rearranged to make more optimal ones. It is possible that single searches can become trapped on suboptimal “islands” (Maddison 1991), so StrataPhy includes the option for multiple replicate searches to better sample the tree space. As with heuristic cladistic

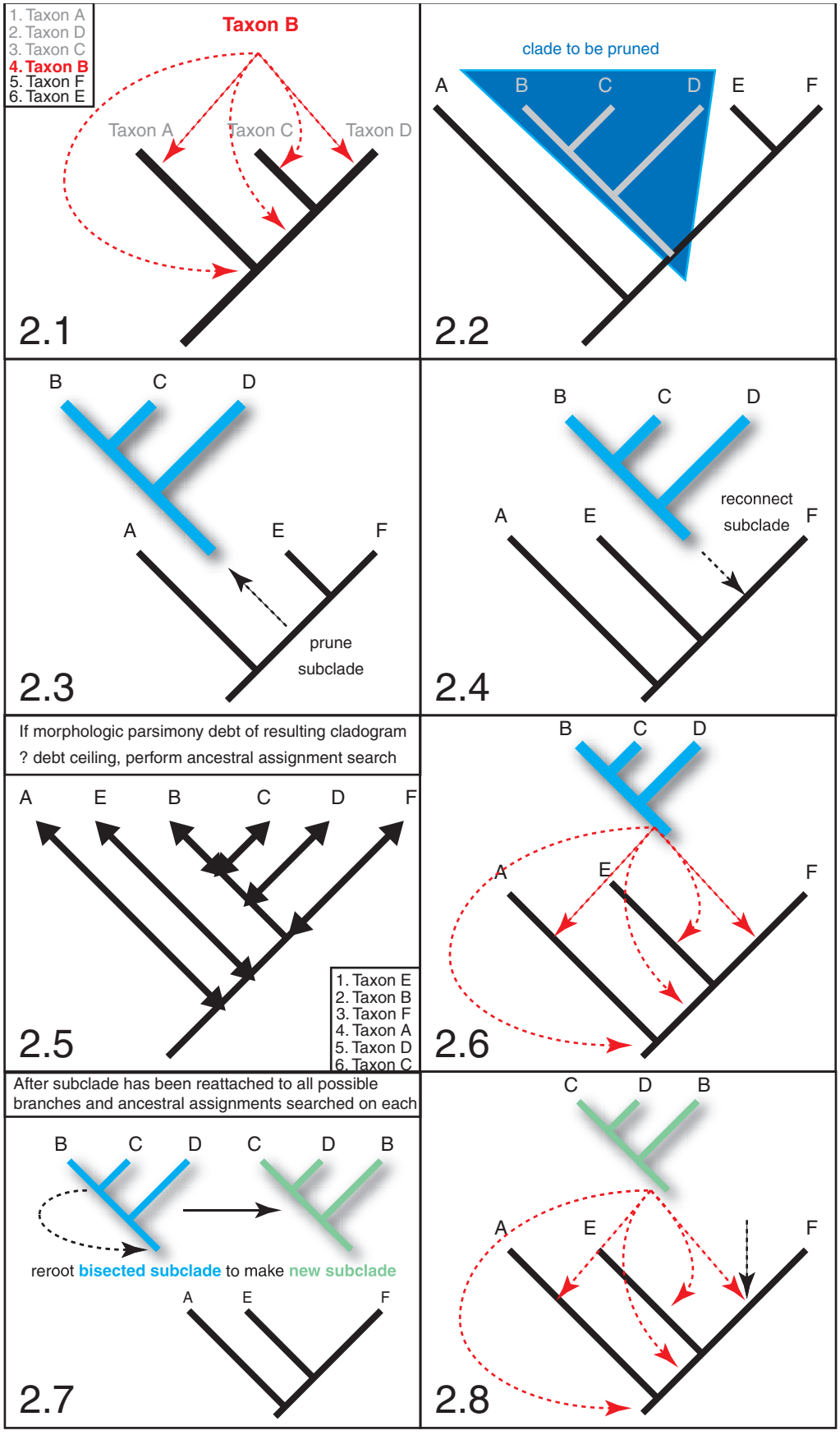
1. Taxon A
2. Taxon D
3. Taxon C
4. Taxon B
5. Taxon F
6. Taxon E
Taxon B
Taxon A
Taxon C
Taxon D
2.1
clade to be pruned
A
B
C
D
E
F
2.2
B
C
D
A
E
F
prune
subclade
2.3
B
C
D
A
E
reconnect
subclade
F
2.4
If morphologic parsimony debt of resulting cladogram
? debt ceiling, perform ancestral assignment search
A
E
B
C
D
F
1. Taxon E
2. Taxon B
3. Taxon F
4. Taxon A
5. Taxon D
6. Taxon C
2.5
B
C
D
A
E
F
2.6
After subclade has been reattached to all possible
branches and ancestral assignments searched on each
B
C
D
C
D
B
reroot bisected subclade to make new subclade
A
E
F
2.7
C
D
B
A
E
F
2.8

Figure 2 (on previous page). Diagram of the branch swapping and ancestor searching algorithms employed by StrataPhy. 2.1) StrataPhy begins by stepwise addition of taxa to build a starting cladogram. The list in the upper left portion of Figure 2.1 represents the random order in which taxa are added. Gray text represents taxa to be added, red text indicates the taxon currently being added, and black text represents taxa remaining to be added by the algorithm. Arrows indicate all the potential locations to which Taxon B could be added. Note that because stratocladistic searches necessarily search rooted trees, Taxon B can be added to the root of the cladogram below all other taxa. 2.2) A subclade is selected to be pruned from the cladogram (highlighted in blue). 2.3) The selected clade (now in light blue) is pruned from the rest of the cladogram. 2.4) The pruned clade is reattached to the original cladogram on a different branch. 2.5) Once attached, a new topology has been created, and the search for the optimal assignment of ancestors begins. The search will occur only if the new cladogram has a lower total debt than the debt ceiling (see text for full description). 2.6) The pruned subclade from step 2.3 is then reattached to all other possible branches, and the ancestor assignment search from step 2.5 is repeated. 2.7) After the subclade has been reattached to all possible branches, it is rerooted to produce a different set of cladistic relationships in the subclade (shown in green). 2.8) This new subclade is then reattached to every possible branch on the original cladogram (as in step 2.6), and each time the ancestor assignment routine from step 2.5 is repeated.

searches, more replicates increase the chance that optimal trees will be found.

The stratigraphic data are inherently linearly ordered and irreversible, which is to say young ancestors cannot give rise to older descendants during the evolution of actual clades, although a given empirical data set with an incomplete fossil record might yield such hypotheses as most parsimonious. The ordered and irreversible nature of the stratigraphic character means that when a stratigraphic character is included, StrataPhy must search rooted rather than unrooted networks. Searching rooted trees results in many more trees that must be considered, and therefore, even without searching for optimal ancestors, stratocladistic tree searching is considerably slower in StrataPhy than the cladistic searches in other available parsimony-based programs. When no stratigraphic character is included, trees are arbitrarily rooted on the first taxon in the matrix. If a stratigraphic character is included, the designation of an outgroup can significantly reduce the number of trees to be searched and correspondingly the analysis time. When outgroup taxa are specified, StrataPhy assumes that the optimal stratocladistic trees will be rooted on at least one of those taxa; StrataPhy will not consider trees in which ingroup taxa root the tree, no matter what their total debt. When no outgroup is specified, which is permissible within the logical framework of the stratocladistic optimality criteria, StrataPhy is forced to search all possible rootings of any candidate tree. Given that there are n-1 possible rootings of an unrooted tree with n taxa, each of which must be searched for optimal ancestor assignment in the absence of an outgroup, specifying an outgroup considerably reduces the possible analysis time. Specifying a single OTU as an outgroup saves time by allowing StrataPhy to skip the step of trying multiple rerootings of candidate trees produced during branch-swapping and therefore represents the fastest possible stratocladistic tree search.

The result of a phylogenetic analysis in StrataPhy is a NEXUS formatted text file containing the optimal trees. This tree file is most easily viewed using MacClade, because MacClade is currently the only program that graphically represents ancestral taxa. This tree file can therefore be used as the template for further studies of character evolution for which MacClade is intended (Maddison and Maddison 2005).

CALCULATION OF MORPHOLOGIC DEBT WITH POLYMORPHIC TAXA

In most instances, StrataPhy uses the same algorithms for calculating morphologic debt as MacClade (see Maddison and Maddison 2000). However, StrataPhy and MacClade differ in the way each calculates morphologic debt when polymorphic taxa are present, and particularly when they are fixed as ancestors. This issue applies only to terminal taxa with more than one observed state (e.g., states 0 and 1), and not to taxa that possess a single uncertain state (e.g., state 0 or 1, but not state 2). Differences in calculated debt arise under two specific conditions: 1) two sister-taxa are both polymorphic, or 2) a polymorphic taxon is fixed as an ancestor. If there are no instances of polymorphic sister-taxa or polymorphic taxa assigned to be ancestors, the calculation of morphologic debt in StrataPhy will be identical to that in MacClade. Here, we describe, in general terms, the novel algorithms used in StrataPhy to calculate morphologic debt when polymorphic taxa are present.

One main difference between the algorithms in the two programs is that StrataPhy allows poly-

morphism to be heritable (i.e., from a common ancestor), but minimizes its occurrence otherwise. For example, if two sister-taxa – which may be either terminal taxa or internal nodes – are both polymorphic (e.g., possess both states 0 and 1), then their ancestor is reconstructed to be polymorphic for those states as well (Figure 3.1). MacClade assumes that only one of their states (e.g., 0 or 1) is present in the ancestor (Figure 3.2). This causes the other state to evolve once in each descendant, for a total of two units of morphologic debt (Figure 3.2). In StrataPhy, the ancestor is reconstructed to be polymorphic for both states, yielding only a single step – the evolution of state 1 in the lineage prior to the most recent common ancestor of the two polymorphic descendants, if the ancestor of that lineages was not polymorphic (Figure 3.1).

As previously mentioned, StrataPhy minimizes the occurrence of polymorphism – if one of the descendants is not polymorphic, then the reconstructed ancestor is assumed to possess only one state, as well (Figure 3.3). A single unit of debt is added for all of the polymorphic taxon's states that do not match its sister-taxon. This matches the behavior of MacClade (Figure 3.4). In the event that two polymorphic descendants do not completely overlap in their states (Figure 3.5), the ancestor is reconstructed to possess all states common to both descendants; the ancestor is therefore polymorphic. If only one state overlaps between the two polymorphic descendants' state sets (Figure 3.7), then the ancestor is reconstructed with that single state and is not polymorphic. In these two instances, the debt incurred is simply the number of unshared states from each taxon. If both descendants are polymorphic, but there are no shared states (Figure 3.8), StrataPhy assumes that at only one of the states observed in the descendants is present in the ancestor, which is therefore not polymorphic. Accordingly, the debt is the number of states observed in the two descendants, minus one for the undetermined ancestral state.

As in MacClade, when a polymorphic taxon is fixed as an ancestor, the states observed in that taxon are assigned to the corresponding node, unless the taxon's state is uncertain or unknown. However, MacClade underestimates the morphologic debt in these circumstances – a property that is well documented in the MacClade manual (Maddison and Maddison 2000, p. 332), and is accompanied by a warning in the MacClade program. When a polymorphic taxon is fixed and its direct descendant (formerly, its sister-taxon) is not polymorphic (Figure 4.1), or if a non-polymorphic taxon is fixed, but its direct descendant *is* polymorphic (Figure 4.2), a unit of debt is added for each state not shared between the two taxa.

DATA FOR A STRATAPHY STRATOCLADISTIC SEARCH

Stratocladistic analysis requires additional data to supplement what is traditionally collected for cladistic analysis. Specifically, stratigraphic data and autapomorphies are necessary components of a stratocladistic data set.

As previously mentioned, one of the chief differences between stratocladistics and cladistics is the use of stratigraphic data in the process of phylogenetic inference in stratocladistics. These data are in the form of the temporal or stratigraphic intervals of first and last appearance of taxa. They are coded as discrete states in the character matrix, rather than as continuous values (e.g., absolute dates or meters of section). Taxa that span multiple intervals are coded as being polymorphic and possess all states corresponding to intervals in which they are sampled. Exactly how stratigraphic intervals should be defined for stratocladistic analysis, and how finely they should be divided remains an unresolved matter, and one for further research (see Smith 2000; Fisher et al. 2002).

Autapomorphies are often explicitly excluded from cladistic analysis, as they offer no information regarding sister-clade relationships among taxa. However, they are required for stratocladistic analysis because they potentially provide information about the optimality of hypotheses of ancestor-descendant relationships among taxa. Operationally, they affect the optimality of assignments of taxa as ancestors, in which consideration of a single taxon as an ancestor weighs the possible stratigraphic debt savings against the possible morphologic debt incurred. This morphologic debt will equal the number of autapomorphies, if character changes have a weight of one, so the exclusion of autapomorphies will bias the results toward finding more optimally assigned ancestors than would be the case if autapomorphies were included. Aside from this obvious issue of erroneously assigning ancestors, failure to include autapomorphies can be a computational challenge, as it can result in a large number of optimal trees when multiple taxa can be either ancestors or terminal taxa on a tree without a change in total debt. In such a case, for a given topology, numerous permutations

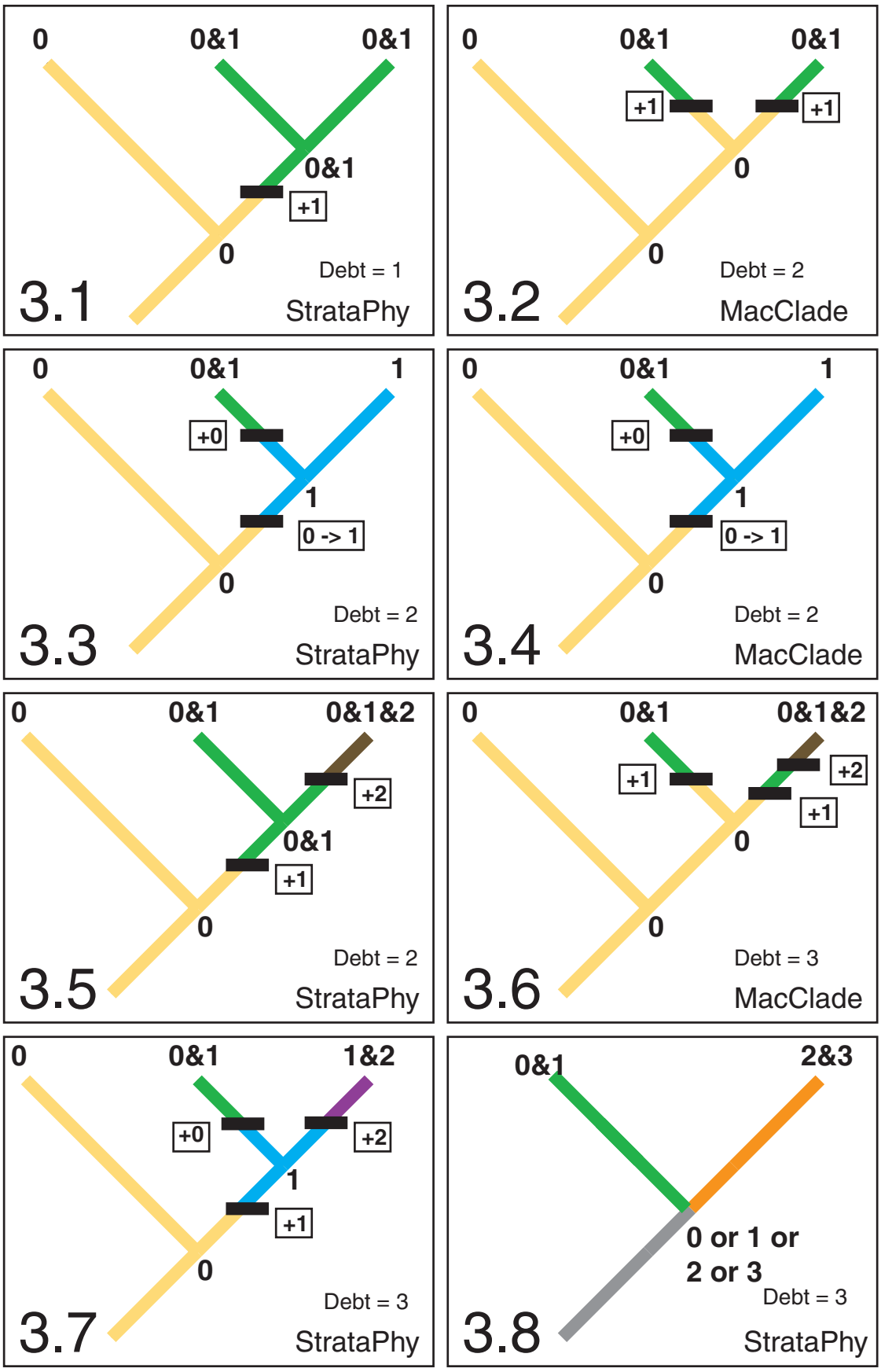
0
0&1
0&1
0&1
+1
0
3.1
Debt = 1
StrataPhy
0
0&1
0&1
+1
+1
0
0
3.2
Debt = 2
MacClade
0
0&1
1
+0
1
0 -> 1
0
3.3
Debt = 2
StrataPhy
0
0&1
1
+0
1
0 -> 1
0
3.4
Debt = 2
MacClade
0
0&1
0&1&2
+2
0&1
+1
0
3.5
Debt = 2
StrataPhy
0
0&1
0&1&2
+1
+2
+1
0
0
3.6
Debt = 3
MacClade
0
0&1
1&2
+0
+2
1
+1
0
3.7
Debt = 3
StrataPhy
0&1
2&3
0 or 1 or
2 or 3
3.8
Debt = 3
StrataPhy

Figure 3 (on previous page). Comparison of the calculation of morphologic debt for polymorphic taxa in StrataPhy and MacClade. Character states for a single reconstructed character are represented by colors: yellow (state 0), blue (state 1), green (states 0&1), red (state 2), purple (states 1&2), brown (states 0&1&2), orange (states 2&3). Tick marks represent points at which units of debt are incurred, and accompanying boxes represent state transitions. For example, "+1" (Figure 3.1) indicates the derivation of state 1 in a newly polymorphic lineage, and "0->1" (Figure 3.3) indicates a transition from state 0 to state 1. The debt is shown for each three-taxon scenario. 3.1) In StrataPhy the ancestor of two polymorphic taxa with identical states is reconstructed to be polymorphic for those states as well. The only debt incurred in this example is on the lineage leading to this polymorphic ancestor, because its ancestor was not polymorphic. 3.2) MacClade assumes that the ancestor of two polymorphic taxa was monomorphic, and therefore requires two changes – the derivation of state 1 in each descendant. 3.3 and 3.4) In both StrataPhy and MacClade, if only one sister-taxon is polymorphic, then the ancestor is assumed to have only one state. In this case calculation of debt in StrataPhy is the same as in MacClade. 3.5) In StrataPhy, if two sister-taxa are both polymorphic, and share some, but not all of their states, the ancestor is reconstructed as being polymorphic for only the shared states. 3.6) As above, MacClade assumes that the ancestor of two polymorphic taxa has only one state, and therefore adds units of debt for each polymorphic state in each descendant. 3.7) In StrataPhy, if two polymorphic sister-taxa share only one state, then the ancestor is reconstructed as having only that state, and is not polymorphic. Units of debt are added for each unshared polymorphic state. 3.8) In StrataPhy, if two polymorphic sister-taxa do not share any states, then the ancestor is assumed to have only one of the observed states, although exactly which cannot be determined from only these two taxa. Units of debt are added for each observed state, minus one, for the undetermined ancestral state.

of ancestral assignments of such taxa will be equally optimal. Searching vast numbers of equally optimal trees dramatically increases search times. Therefore, it is strongly suggested that users resist the temptation to append a single stratigraphic character to existing cladistic matrices for analysis in StrataPhy without first adding autapomorphies of included taxa.

FORMATTING NEXUS FILES FOR STRATAPHY ANALYSIS

StrataPhy reads files in the standard NEXUS format (Maddison et al. 1997) common to many modern computer programs for phylogenetic analysis. See Maddison et al. (1997) for a more detailed description of the NEXUS format and for the syntax of commands common to most NEXUS files. Here, we restrict our description of NEXUS syntax to only that necessary to perform stratocladistic analyses in StrataPhy. An example NEXUS file with a StrataPhy block is included in the Appendix 2. In general, NEXUS files are ASCII text files

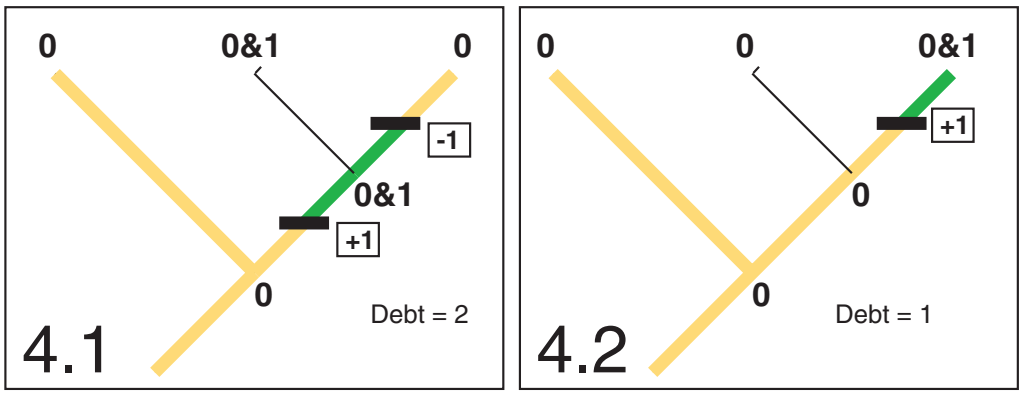

Figure 4. Morphologic debt for polymorphic taxa when some taxa are fixed as ancestors. Colors and symbols as in Figure 3. Fixed ancestors are represented by thin black lines. 4.1) When a polymorphic taxon is fixed as an ancestor, and its direct descendant is not, then units of debt must be added for each state not observed in the descendant. If the fixed taxon's ancestor is not polymorphic, then units of debt must also be added for each state reconstructed in the ancestor. 4.2) If a monomorphic taxon is fixed as an ancestor, and its direct descendant is polymorphic, units of debt must be added for each state not observed in the fixed ancestor.

and may be created in any text editor. Alternatively, they can be created and edited in programs explicitly designed to produce NEXUS files, such as MacClade or Mesquite (Maddison and Maddison 2004). NEXUS files for StrataPhy include labels of taxa, character data, as well as a stratigraphic character. Note that at the present time MacClade is the only program that fully supports stratigraphic data and the corresponding character type, as well as the assignment of taxa as ancestors. Ideally, StrataPhy and MacClade should be used in concert for the straightforward creation of data sets. In MacClade, stratigraphic characters can be specified, and will automatically be incorporated into the NEXUS file. Other programs such as Mesquite do not support stratigraphic data explicitly, and assumptions about stratigraphic characters must be added manually in an additional step.

StrataPhy currently supports both ordered and unordered categorical character types, as well as the stratigraphic character type in MacClade (Maddison and Maddison 2005). To make StrataPhy compatible with MacClade for NEXUS file creation and tree viewing and analysis, StrataPhy shares some critical restrictions on character states with MacClade. StrataPhy support a maximum of 34 character states: the numbers “0” through ”9” and letters “A” through “Z”, regardless of case, are all valid character states, except for “I” and “O”, which are invalid. Therefore due to restrictions imposed by MacClade, the stratigraphic character can include no more than 34 intervals.

StrataPhy allows differential user-defined weighting of characters, including the stratigraphic character. Character types and weights are specified using the TYPESET and WTSET options, respectively, in the ASSUMPTIONS block of the NEXUS data file, as is standard NEXUS format. Characters present in a data matrix contained in the NEXUS file can be excluded from analysis by creating an exclusion set (EXSET) in the ASSUMPTIONS block. Note that MacClade refers to exclusion sets as “Inclusion Sets.” This option records the character numbers of characters to be excluded prior to StrataPhy analysis. StrataPhy only performs stratocladistic analysis using the “active” TYPESET, WTSET, or EXSET set, which are each designated with an asterisk (*) in the NEXUS file (see Appendix 2). Typically, these sets are those active (i.e., selected) at the time the file was saved to disk. It is possible that a user might explicitly store a single TYPESET, for example, but if the user were to change the types of any character, then save the entire file to disk, the untitled TYPESET in effect at the time of the save would be that used by StrataPhy. Care must be taken to designate the TYPESET, WTSET, and EXSET immediately before saving NEXUS files in MacClade, and we encourage users to review their selections prior to analysis by opening NEXUS files in any standard text editor. Individual taxa in a matrix cannot be excluded from analysis within StrataPhy within the ASSUMPTIONS block, so taxa to be excluded must be deleted from the NEXUS file prior to analysis with StrataPhy.

The current version of StrataPhy utilizes a command line for inputting the file name of the NEXUS file to be analyzed (instead of a graphic user interface), so all parameter settings for the analysis must be specified within the NEXUS file prior to analysis. These parameters are set by placing commands in a new block labeled “STRATAPHY” in the NEXUS file. Syntax of the StrataPhy block is given below. Terms in italics would be replaced by user-defined values, and those in braces represent the range of possible parameters with the default underlined.

```
BEGIN STRATAPHY;
   MAXTREES = maximum-number-of-
      trees;
   SEARCHREPS = number-of-branch-
      swapping-replicates;
   ANCREPS = number-of-ancestral-
      search-replicates;
   ANCTYPE = { heuristic | exhaustive
      };
   OUTGROUP outgroup1 outgroup2
      outgroup3;
END;
```

MAXTREES – This command sets the maximum number of optimal trees to be saved in a heuristic search. It is also the maximum number of trees saved during single reps of a heuristic search, although if these are not globally optimal over all previous reps, they will not be saved. StrataPhy currently cannot adjust MAXTREES dynamically during a search, so this must be specified prior to the analysis. The default setting of MAXTREES is set to 1000.

SEARCHREPS – This command sets the number of heuristic search replicates to be performed. Each of these replicates begins with a tree built by random taxon addition. The default number of search replicates is 10.

ANCREPS – This command sets the number of replicate searches for optimal assignments of ancestors reps per tree produced by branch swapping. By default, StrataPhy performs only a single ancestral search replicate.

ANCTYPE – This command designates the type of search for optimal assignments of ancestors, either heuristic or exhaustive. Heuristic search (described above) is the default.

OUTGROUP – This command designates taxa that will be used as the outgroup. Taxon labels must exactly match the spelling of those in the translation table and matrix in the NEXUS file, and multiple outgroup taxa are separated by single spaces. If no outgroup is specified, StrataPhy assumes all taxa are in the ingroup. In this case, if a stratigraphic character is present, StrataPhy will search over all possible rooted trees, so it is recommended that at least one outgroup taxon is specified.

EMPIRICAL EXAMPLE

To demonstrate the potential utility of StrataPhy and to contrast its use with manual searches using PAUP and MacClade, we reanalyzed the Paleogene viverravid data set of Polly (1997) using StrataPhy. Our intention in reanalyzing Polly's data set is not comment on viverravid phylogeny, but to show how the automated procedure provided in a StrataPhy analysis can provide a more extensive search of the potential phylogenetic trees, in a reasonable amount of time.

The data matrix consists of 23 taxa (after redundant taxa were removed from the original matrix with 32 taxa, as in the original study), with 39 morphologic characters and a single stratigraphic character with 14 stratigraphic intervals. In the original analysis, Polly generated a set of most parsimonious cladograms using PAUP, and then manually manipulated them and designated ancestors using MacClade. Using this procedure, he identified two equally parsimonious phylogenetic trees with total length of 91 steps.

We performed a stratocladistic analysis with StrataPhy using 10 random taxon addition sequence replicates and a single heuristic ancestral search per replicate. The hypothetical outgroup included in the matrix was designated as the sole outgroup for the analysis. All morphologic characters were treated as unordered and given equal weight. All other settings were StrataPhy defaults. Whereas analysis time of different data sets will vary tremendously depending on various properties, this analysis was completed in 2 minutes and 22 seconds on a 1.67GHz Apple PowerBook G4 computer. StrataPhy discovered 24 most parsimonious phylogenies from a single "island" (*sensu* Maddison 1991) of length 89 – two steps shorter than those reported by Polly.

It should be noted that most of these 24 phylogenies differ in only the assignment of taxa as ancestors, as they correspond to only two cladistic topologies. The topologies of the most parsimonious trees from the StrataPhy analysis were not identical to that from Polly's analysis. In fact, the cladistic topologies of the most parsimonious phylogenies from the StrataPhy stratocladistic analysis are congruent with none of the most parsimonious cladograms produced in the PAUP cladistic search (Polly 1997). If the two most parsimonious trees are converted to their corresponding cladograms where observed taxa are not fixed as ancestors, the total morphological steps for the two trees are one and two steps longer, respectively, than the most parsimonious cladograms from the PAUP search. This result highlights the potential risk of missing the most parsimonious phylogeny inherent in the commonly employed technique of using only the most parsimonious cladograms from a PAUP search as a starting point for manual stratocladistic analysis in MacClade. The simultaneous branch-swapping and ancestral assignment search in StrataPhy allows users to rapidly survey a much broader set of possible optimal trees.

CONCLUSIONS

StrataPhy is still in development, and we are still adding functionality. Improvements for future releases will focus on three major themes. First, performance increases will come in the form of published "shortcuts" (e.g., Goloboff 1996, 1998, 1999; Ronquist 1998) that are not currently implemented in StrataPhy. This will also include the exploration of alternative tree search strategies besides TBR. The second main focus of future development will be the incorporation of probabilistic optimality criteria (Bayesian and likelihood; Huelsenbeck and Crandall 1997; Wagner 1998; Huelsenbeck and Rannala 2000; Wagner 2000). Finally, StrataPhy, which is currently written in C++, is being converted to a Java-based program to make it truly cross-platform and allow the development of a simple graphical user interface (GUI) with tree visualization that will be independent of MacClade.

Stratocladistics has not been explored as fully as warranted due to want of an automated applica-

tion. Thus some of the issues raised by critics that are not purely philosophical objections (e.g., Smith 2000; Sumrall and Brochu 2003; but see replies by Fisher et al. 2002; Fisher and Bodenbender 2003), and some aspects of the behavior of the method have not yet been explored in sufficient detail. We hope that StrataPhy will allow these issues to be addressed more fully and foster continued exploration of the use of stratigraphic data in phylogenetic inference.

Because not all clades have fossil records that are well suited to stratocladistic analysis, StrataPhy should be viewed as an additional tool for phylogenetic analysis to complement currently available methods, and not a replacement. Taken together, these methods will continue to help us understand the evolutionary histories of lineages and to determine the patterns that result from the diverse processes that have controlled the evolution of life as recorded in the fossil record and the extant biota.

ACKNOWLEDGEMENTS

We thank D. Fisher, A. Huttenlocker, J. Pardo, and P. Wagner for constructive discussion on the development of StrataPhy. J. Finarelli and an anonymous reviewer provided valuable comments that greatly improved the quality of this manuscript and discovered bugs in early versions of StrataPhy. The majority of this work was produced during a postdoctoral position to J.D.M. funded by the University of Minnesota Department of Geology and Geophysics.

REFERENCES

Bloch, J.I., Fisher, D.C., Rose, K.D., and Gingerich, P.D. 2001. Stratocladistic analysis of Paleocene Carpolestidae (Mammalia, Plesiadapiformes) with description of a new late Tiffanian genus. *Journal of Vertebrate Paleontology,* 21(1):119-131.

Bodenbender, B.E., and Fisher, D.C. 2001. Stratocladistic analysis of blastoid phylogeny. *Journal of Paleontology,* 75(2):351-369.

Brochu, C.A., Bryant, H.N., Theodor, J.M., O'Leary, M.A., Adrain, J.M., and Sumrall, C.D. 2001. Modern phylogenetics in paleontology: comments on Vermeij 1999. *Paleobiology,* 27(1):174-176.

Clyde, W.C., and Fisher, D.C. 1997. Comparing the fit of stratigraphic and morphologic data in phylogenetic analysis. *Paleobiology,* 23(1):1-19.

Farris, J.S. 1983. The logical basis of phylogenetic analysis. In Platnick, N.I., and Funk, V.A. (eds.), *Advances in Cladistics, Vol. 2. Proceedings of the Second Meeting of the Willi Hennig Society*. Columbia University Press, New York.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts.

Fisher, D.C. 1982. Phylogenetic and macroevolutionary patterns within the Xiphosurida. *Third North American Paleontological Convention Proceedings,* 1:175-180.

Fisher, D.C. 1992. Stratigraphic parsimony, p. 123-128. In Maddison, W.P., and Maddison, D.R. (eds.), *MacClade: Analysis of Phylogeny and Character Evolution*. Sinauer Associates, Inc., Sunderland, Massachusetts.

Fisher, D.C. 1994. Stratocladistics: morphological and temporal patterns and their relation to phylogenetic process, p. 133-172. In Grande, L., and Rieppel, O. (eds.), *Interpreting the hierarchy of nature: from systematic patterns to evolutionary process theories*. Academic Press, London.

Fisher, D.C., and Bodenbender, B.E. 2003. Blastoid stratocladistics - reply to Sumrall and Brochu. *Journal of Paleontology,* 77(1):195-198.

Fisher, D.C., Foote, M., Fox, D.L., and Leighton, L.R. 2002. Stratigraphy in phylogeny reconstruction - Comment on Smith (2000). *Journal of Paleontology,* 76(4):585-586.

Foote, M. 1996. On the probability of ancestors in the fossil record. *Paleobiology,* 22(2):141-151.

Fox, D.L., Fisher, D.C., and Leighton, L.R. 1999. Reconstructing phylogeny with and without temporal data. *Science,* 284:1816-1819.

Geisler, J.H., and Uhen, M.D. 2005. Phylogenetic relationships of extinct cetartiodactyls: results of simultaneous analyses of molecular, morphological, and stratigraphic data. *Journal of Mammalian Evolution,* 12(1/2):145-160.

Goloboff, P.A. 1996. Methods for faster parsimony analysis. *Cladistics,* 12:199-229.

Goloboff, P.A. 1998. Tree searches under Sankoff parsimony. *Cladistics,* 14:229-237.

Goloboff, P.A. 1999. Analyzing large data sets in reasonable times: solutions for composite optima. *Cladistics,* 15(415-428).

Harvey, P.H., and Pagel, M.D. 1991. *The Comparative Method in Evolutionary Biology. Oxford Series in Ecology and Evolution*. Oxford University Press, Oxford, UK.

Heyning, J.E., and Thacker, C.E. 1999. Phylogenies, temporal data, and negative evidence. *Science,* 285:1179a.

Huelsenbeck, J.P., and Crandall, K.A. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics,* 28:437-466.

Huelsenbeck, J.P., and Rannala, B. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science,* 276:227-232.

Huelsenbeck, J.P., and Rannala, B. 2000. Using stratigraphic information in phylogenetics. In Wiens, J.J. (ed.), *Phylogenetic Analysis of Morphological Data.* Smithsonian Series in Comparative Evolutionary Biology. Smithsonian Institution Press, Washington, D.C.

Hull, D.L. 1979. The limits of cladism. *Systematic Zoology,* 28(4):416-440.

Lane, A., Janis, C.M., and Sepkoski, J.J., Jr. 2005. Estimating paleodiversities: a test of taxic and phylogenetic methods. *Paleobiology,* 31(1):21-34.

Lieberman, B.S. 2005. Geobiology and paleobiogeography: tracking the coevolution of the Earth and its biota. *Palaeogeography, Palaeoclimatology, Palaeoecology,* 219:23-33.

Maddison, D.R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. *Systematic Zoology,* 40(3):315-328.

Maddison, D.R., and Maddison, W.P. 2000. *MacClade 4*. Sinauer Associates, Sunderland, Massachusetts.

Maddison, D.R., and Maddison, W.P. 2005. *MacClade*. Sinauer Associates, Sunderland, Massachusetts.

Maddison, D.R., Swofford, D.L., and Maddison, W.P. 1997. NEXUS: An extensible file format for systematic information. *Systematic Biology,* 46(4):590-621.

Maddison, W.P., and Maddison, D.R. 2004. Mesquite: a modular system for evolutionary analysis.

Nelson, G. 1978. Ontogeny, phylogeny, paleontology, and the biogenetic law. *Systematic Zoology,* 27(3):324-345.

Norell, M.A. 1992. Taxic origin and temporal diversity: the effect of phylogeny, p. 89-118. In Novacek, M.J., and Wheeler, Q.D. (eds.), *Extinction and Phylogeny*. Columbia University Press, New York.

Norell, M.A. 1993. Tree-based approaches to understanding history: comments on ranks, rules, and the quality of the fossil record. *American Journal of Science,* 293(A):407-417.

Novacek, M.J., and Norell, M.A. 1982. Fossils, phylogeny, and taxonomic rates of evolution. *Systematic Zoology,* 31(4):366-375.

Polly, P.D. 1997. Ancestry and species definition in paleontology: a stratocladistic analysis of Paleocene-Eocene Viverravidae (Mammalia, Carnivora) from Wyoming. *Contributions from the Museum of Paleontology, University of Michigan,* 30(1):1-53.

Ree, R.H., Moore, B.R., Webb, C.O., and Donoghue, M.J. 2005. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution,* 59(11):2299-2311.

Ronquist, F. 1998. Fast Fitch-parsimony algorithms for large data sets. *Cladistics* 14:387-400.

Sanderson, M.J., and Donoghue, M.J. 1996. Reconstructing shifts in diversification rates on phylogenetic trees. *Trends in Ecology and Evolution,* 11:15-20.

Slowinski, J.B., and Guyer, C. 1993. Testing whether certain traits have caused amplified diversification: an improved method based on a model of random speciation and extinction. *The American Naturalist,* 142(6):1019-1024.

Smith, A.B. 1988. Patterns of diversification and extinction in early Palaeozoic echinoderms. *Palaeontology,* 31:799-828.

Smith, A.B. 2000. Stratigraphy in phylogeny reconstruction. *Journal of Paleontology,* 74(5):763-766.

Stockmeyer Lofgren, A., Plotnick, R.E., and Wagner, P.J. 2003. Morphological diversity of Carboniferous arthropods and insights on disparity patterns through the Phanerozoic. *Paleobiology,* 29(3):349-368.

Sumrall, C.D., and Brochu, C.A. 2003. Resolution, sampling, higher taxa and assumptions in stratocladistic analysis. *Journal of Paleontology,* 77(1):189-194.

Swofford, D.L. 2002. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, Massachusetts.

Swofford, D.L., Olsen, G.J., Waddell, P.J., and Hillis, D.M. 1996. Phylogenetic inference, p. 407-514. In Hillis, D.M., Moritz, C., and Mable, B.K. (eds.), *Molecular Systematics*. Sinauer Associates, Sunderland, Massachusetts.

Wagner, P.J. 1996. Patterns of morphological diversification during the initial radiation of the "Archaeogastropoda," p. 161-169. In Taylor, J.D. (ed.), *Origin and Evolutionary Radiation of the Mollusca*. Oxford University Press, Oxford, England.

Wagner, P.J. 1998. A likelihood approach for evaluating estimates of phylogenetic relationships among fossil taxa. *Paleobiology,* 24(4):430-449.

Wagner, P.J. 2000. Phylogenetic analyses and the fossil record: tests and inferences, hypotheses and models, p. 341-371. In Erwin, D.H., and Wing, S.L. (eds.), *Deep Time: Paleobiology's Perspective*. The Paleontological Society, Lawrence, Kansas.

APPENDIX 1

StrataPhy is a computer program for stratocladistic analysis written by Jonathan Marcot and David Fox. This is an early release, and development is ongoing. Therefore, we request two things:

1. Any problems/bugs be reported to the developers.
2. Researchers seek permission to publish results from StrataPhy analyses (so that we may ensure that no problems with the software discovered after your download have caused erroneous results).

Documentation is currently rudimetary, but accompanies the downlaods below. A proper manual is currently in the works. StratPhy is currently available for Mac and Windows, and the source code will be made available upon final release. This release of StrataPhy is version 0.3.5 (posted 7 March 2008). Newer releases will be posted to the authors' website as they become available.

StrataPhy for Mac (PPC)
StrataPhy for Mac (Intel)
StrataPhy for Windows

APPENDIX 2

An example NEXUS file with StrataPhy block.

```
#NEXUS

BEGIN TAXA;
    DIMENSIONS  NTAX=6;
    TAXLABELS Taxon_1 Taxon_2 Taxon_3 Taxon_4 Taxon_5 Taxon_6;
END;

BEGIN CHARACTERS;
    DIMENSIONS  NCHAR=7;
    FORMAT SYMBOLS= " 0 1 2 3 4 5" MISSING=? GAP=- ;
MATRIX
Taxon_1   0000000
Taxon_2   1000001
Taxon_3   1100002
Taxon_4   1110003
Taxon_5   1112004
Taxon_6   1112105
;
END;
BEGIN ASSUMPTIONS;
    TYPESET * UNTITLED = ord: 1-6, strat: 7;
    TYPESET unord_only = unord: 1-6, strat: 7;
    WTSET * UNTITLED = 1.00: 1-7;
    WTSET variable = 2.00: 1-3, 1.00: 4-7;
    EXSET UNTITLED = 4-6;
    EXSET * stored_1 = 4;
END;
BEGIN TREES;
    TRANSLATE
        1    Taxon_1,
        2    Taxon_2,
        3    Taxon_3,
        4    Taxon_4,
        5    Taxon_5,
        6    Taxon_6
    ;
    TREE  * Tree1 =  [&R] (((((6)5)4)3)2)1;

END;

BEGIN STRATAPHY;
        MAXTREES = 1000;
        SEARCHREPS = 10;
        ANCREPS = 1;
        ANCTYPE = heuristic;
        outgroup Taxon_1 Taxon_2;
END;
```