

# VIEW FROM THE FIELD

*J. Paleont.*, 74(5), 2000, pp. 763–766  
Copyright © 2000, The Paleontological Society  
0022-3360/00/0074-763\$03.00

Beginning with this issue, the *Journal of Paleontology* will be publishing an occasional section entitled “View from the Field”. “View from the Field” will provide a forum for concise commentary on current issues in paleontology, including but not limited to discussion of methods, important new discoveries, and major conceptual advances. Contributions should be brief (no more than 12 double-spaced typewritten pages). Submissions will be vetted for appropriateness by the Editors and subject to peer review prior to acceptance. Enquiries and submissions should be directed to the Journal of Paleontology Special Projects Editor, Dr. Jonathan Adrain, Department of Geoscience, 121 Trowbridge Hall, University of Iowa, Iowa City, IA 52242. Phone (319) 335-1539; <jonathan-adrain@uiowa.edu>

## STRATIGRAPHY IN PHYLOGENY RECONSTRUCTION

ANDREW B. SMITH

Department of Palaeontology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK  
<a.smith@nhm.ac.uk>

**A**CCCESS to the dimension of time makes paleontology unique as a discipline, and it is stratigraphical data that allows paleontologists to tackle a wide range of evolutionary questions unanswerable by neontologists. Some of these need only a vague and imprecise hypothesis of evolutionary relationships. For example, considerable headway has been made in documenting the evolution of morphological disparity with only the crudest of phylogenetic information (Foote, 1997). However, it is undoubtedly true that more precise and probing questions can be formulated if accurate phylogenies were available. But how do we construct such phylogenies?

Trees indicating ancestor-descendent relationships used to be constructed in a vague and arbitrary way, using a combination of morphological similarity and stratigraphic proximity. As working hypotheses these served a purpose, but these days phylogenetic hypotheses need to be formulated with much more rigor in order to be taken seriously. First there must be an optimality criterion—a yardstick by which competing explanations of the data can be judged (e.g., minimal tree length or maximum likelihood under a given model). Then a wide range of possible solutions (trees) are objectively evaluated against this optimization criterion for a given set of observations. Thus phylogenetic hypotheses are judged by their fit against empirical data.

Paleontologists have access to two categories of data, the distribution of character states amongst fossil taxa and the temporal distribution of those taxa in the fossil record. Unfortunately, neither can be obtained without ambiguity and error. Character state definitions of complex morphological features are horrendously arbitrary in comparison to molecular sequence data, while sampling and preservational biases in the fossil record mean that observed ranges differ, sometimes considerably, from true ranges. Thus, although we can specify an optimality criterion with little problem, there remain significant obstacles in deriving accurate phylogenies from non-ideal data.

Stratigraphic data cannot generate phylogenetic hypotheses on their own—there is no information that helps us sort trilobites from graptolites or herring from mayflies. A paleontologist must use morphological data either alone, or in combination with stratigraphic data to reconstruct phylogeny. Current debate focuses on the relative merits of these two approaches.

### APPROACHES TO ESTIMATING PHYLOGENY

*Stratigraphy-free methods.*—Parsimony is the only widely employed stratigraphy-free approach currently used by paleontologists, although distance and maximum likelihood are alternative methods popular with molecular biologists. Parsimony methods have few assumptions—they set out to explain the observed distribution of characters amongst taxa in the most economical way, assuming evolutionary descent with modification. Its operational criterion is minimal tree-length (=minimal evolutionary steps, a step being a single transformation from one character state to another) under a specific set of assumptions about how character states change. In contrast to molecular data, no biologically-justified method has been proposed that can estimate or model, a priori, how frequently morphological character states transform over time relative to one another. Thus although parsimony allows each character state transition to be assigned a finite “cost”, for morphological data an equal probability model is almost universally applied at least initially.

Random errors are introduced from character definition and because trees are constructed using small numbers of characters, while systematic error arises from non-independence of characters and by reversal and parallel change that introduce homoplasy. Although parsimony performs well in recovering known viral phylogenies (Hillis et al., 1994), simulation experiments have shown that high levels of homoplasy can decrease the accuracy of phylogenetic inference under the parsimony criterion (Helsenbeck and Hillis, 1993; Wagner, 1999). Unfortunately there is no way of inferring levels of homoplasy independent of a phylogenetic hypothesis, so the accuracy of stratigraphic-free phylogenies of fossils is never known.

*Methods combining stratigraphic and morphologic data.*—Because there are unknown levels of systematic error associated with parsimony estimates, some workers have advocated using stratigraphic distribution to “improve” phylogenetic accuracy. The criterion for selecting a phylogenetic hypothesis is then the tree that best explains the observed stratigraphic distribution and character distribution overall. How “best” is defined depends upon the method of analysis adopted.

1) Stratocladistics and related approaches.—These methods begin with a parsimony analysis of morphological data to identify optimal and suboptimal cladograms. Each cladogram is then calibrated against stratigraphic occurrence data and the match of the observed record to the inferred record assessed. This can be used

to select one from amongst a number of equally most-parsimonious alternatives (Smith, 1994) or to argue for preferring a topology that would be rejected under the parsimony criterion alone.

Wagner (1995) used confidence intervals on observed fossil ranges to reject the maximum parsimony tree, favoring a suboptimal tree in which inferred sister-groups had ranges with overlapping 95 percent confidence intervals. However, the ad hoc nature of assessing the relative weight to give to stratigraphic fit as opposed to character congruence makes this technique very weak.

A more widely used procedure is to calculate the missing record inferred when optimal and suboptimal cladograms are mapped onto the biostratigraphical record, and to select the hypothesis that offers the best compromise between minimizing morphological tree length and minimizing gaps in the fossil record. Stratocladistics, as advocated by Fisher (1992, 1994) codes stratigraphic occurrence as an ordered character and adds this to the parsimony tree-length score. The "best" solution is then the one that explains the observed distribution of taxa and characters in the most economical way, i.e., with fewest assumptions of homoplasy and missing segments of fossil record (stratigraphic debt).

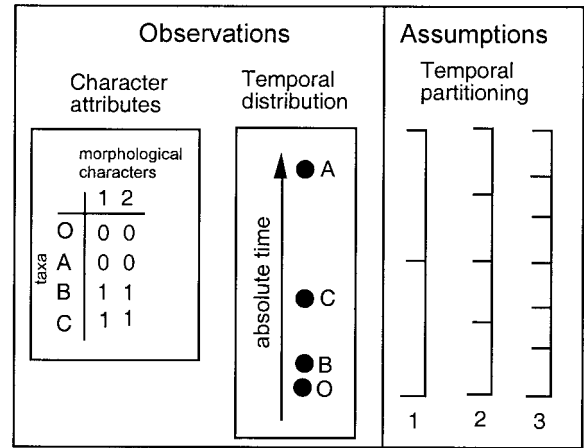
2) Stratolikelihood.—With the same goal in mind Wagner (1998) has developed a technique of assessing "best" in terms of Maximum Likelihood theory. Maximum Likelihood has the useful property that probabilities are additive and so hypotheses can be tested for their fit against multiple optimality criteria.

Once again parsimony analysis of morphological data provides the starting point by identifying the set of optimal and suboptimal trees. There are two strands to Wagner's approach. First he estimates the quality of the fossil record from the observed distribution of taxonomic ranges (cf. Foote, 1998) and from this derives the maximum likelihood scores for extinction rate and sampling intensity for the data. Next he performs a series of simulations constructing phylogenetic trees with a data set of the same dimensions and character types as the original. Extinction and origination rates are specified and trees are computer-generated across a broad spectrum of character change probabilities. This provides a likelihood estimate of observing trees of the maximum parsimony length, given a data matrix of the original size and composition under the range of evolutionary models used. The criterion for "best" is then the solution with maximum likelihood of generating a tree of the observed length and stratigraphic debt under the defined model.

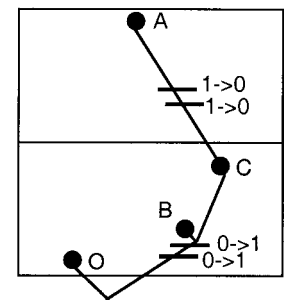
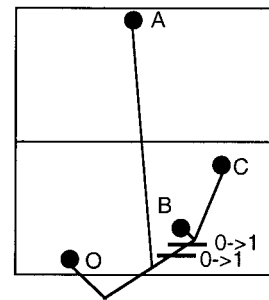
MODELING THE REAL WORLD

Stratigraphic data are included as part of the optimization criterion on the assumption that fossils are, in general, preserved in the right temporal order. With fine enough subdivision of time, however, stratigraphic debt can be made sufficiently large to overturn any phylogeny based on morphology in favor of a minimal gap tree (Fig. 1). Conversely, if longer time intervals are used many fewer cladograms need be rejected. Adding stratigraphical match to the optimization criterion thus increases the uncertainty and potential for error when reconstructing phylogeny because no objective threshold for this metric exists.

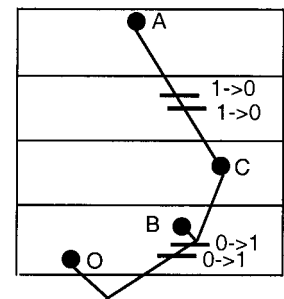
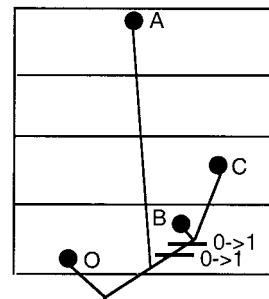
So do the advantages of stratigraphic data outweigh the disadvantages? Support for using a combination of stratigraphic and



Assumption 1



Assumption 2



Assumption 3

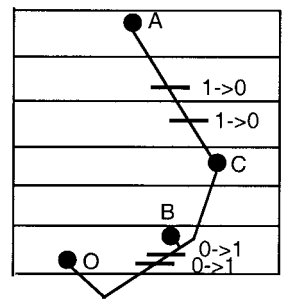
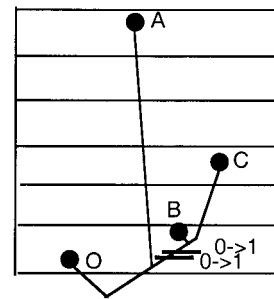


FIGURE 1—The effect of time partitioning on the optimality criterion (total score = tree length + stratigraphic debt) in stratocladistics. Lower scores are better. Whether the morphology-based tree (left-hand) or minimal-gap tree (right-hand) is favored depends on the arbitrary time-scale chosen.

morphological data in phylogeny reconstruction has come from two recent observations: simulation experiments comparing algorithm performance under known conditions, and estimates of the completeness of the fossil record.

*Simulation experiments.*—Simulation experiments, in which computer-generated trees are used to test the accuracy of different methods of analysis at recovering the correct tree, have proved extremely useful for assessing conditions that cause methods of phylogenetic reconstruction to fail (e.g., Huelsenbeck and Hillis, 1993). Simulation experiments form one of the two strands of the stratolikelikelihood technique (Wagner, 1998) and the basis for a recent claim of the superiority of stratocladistic techniques (Fox et al., 1999). However, simulation studies test the accuracy of methods only with respect to the specific set of assumptions on which the model was constructed. It is very easy to unwittingly bias the results in favor of one or other method unless the complete range of parameters is explored (Hillis et al., 1994). The simulation experiments of both Fox et al. and Wagner were of limited scope, so how close do the assumptions in their models match reality?

There are several worrisome features of the Fox et al. (1999) simulations. For example, parameters ensure that every time segment of the tree is morphologically discrete (0.1 or greater character transformation rate per step over 50 characters). As these segments are also the terminal taxa of the analysis it means that no taxon overlaps with any other, thereby prohibiting any sampling error that might reverse the order of occurrence of two taxa. Furthermore stratigraphic debt intervals then exactly match taxon durations so that when a taxon is sampled its complete range is sampled. So when Fox et al. (1999, p. 1819) claim that their “results lay to rest the notion that an incomplete fossil record yields no clues for inferring phylogeny” it is only within the parameters of their very specific and unrealistic model. The fossil record I am familiar with is much more complex than this.

There are also major problems with Wagner’s (1998) simulation experiment used to calculate the likelihood mismatch between true tree length and tree length as estimated from parsimony. Wagner used a model with the probability of character transformation set between 0.05 and  $>0.25$  per step (mean 0.15) for 23 characters evolved over 12 steps. Under all but the lowest estimates most characters can then be expected to have undergone reversion, and the mean change per character per branch in the simulation is higher than any found in real data sets of comparable size (cf. Wagner, 1999, fig. 3). The tree-length distributions that result are only weakly skewed suggesting that very little hierarchical signal resides in these data (cf. Hillis, 1991). What Wagner has demonstrated is that under high levels of homoplasy parsimony usually fails to find the correct tree. But the original data matrix had a highly skewed tree-length distribution; thus his simulation experiment does not provide a realistic optimality criterion against which to test the original phylogeny.

Clearly these specific models are unsatisfactory and misleading, but the point I wish to stress is that all such models are simplistic in comparison to real data and require estimates of unknown parameters. Computer simulation as a means of exploring how specific parameters effect the performance of an algorithm is entirely justified (e.g., Wagner, 1999), but as a means of establishing what those parameters might be in the real world is foolhardy. If simulation experiments form a keystone of the optimization criterion, as in stratolikelikelihood, then they represent an additional major potential source for error.

*Completeness of the fossil record.*—If there is a relatively good fossil record then methods that incorporate stratigraphic data might be expected to recover the correct tree more often than a stratigraphy-free analysis. Although early estimates suggested that less than 10 percent of species were preserved in the fossil record (Valentine, 1970), more recent findings suggest that we have

achieved surprisingly high levels of completeness. Support for this view come from gap analysis (Paul and Donovan, 1998), historical sampling patterns (Paul and Donovan, 1998; Maxwell and Benton, 1990; Benton, 1998) and frequency distribution analysis (Foote and Raup, 1996; Foote, 1997; Solow and Smith 1997; Foote and Sepkoski, 1999).

Clearly our knowledge of the fossil record is now remarkably good. However, all of these techniques simply ask to what extent we have sampled the available rock record. They pointedly fail to address how representative that fossil record is of what once existed. Even when applied to the global record of higher taxa (e.g., Foote and Sepkoski, 1999) the method is summarizing data from a patchwork of small windows provided by the preserved and accessible portions of sedimentary basins—a minute portion of what once existed. By compiling range data into a single cohort, Foote’s (1997) technique asks the question “What is the extinction rate and preservation potential averaged over all our windows of fossiliferous strata at generic or family level?” It does nothing to answer the question what proportion of the fossil record is revealed by those windows. Thus estimates for completeness of the fossil record are local estimates and thus absolute maxima, as Foote (1996) was careful to emphasise. Estimation of the global completeness of the fossil record remains poorly constrained and error-prone despite high local completeness (see Foote, 1996).

#### CONCLUSIONS

The past 150 years has witnessed an exponential growth in the intensity of paleontological collecting, and has brought us to a point where our sampling of the fossils from the rock record on the whole is really very good. Yet that sampled record is only a portion of what once existed, and varies in quality and representation through time.

All methods of phylogenetic tree reconstruction estimate branching order and sister-group relationships of fossils from small amounts of imperfect morphological data and are thus prone to error. Unfortunately, the inclusion of stratigraphic data can only add to the uncertainties involved in this estimation process. The fossil record does not provide error-free stratigraphical ranges, even locally, and systematic errors can be introduced according to how finely time is divided when deriving a measure of stratigraphic debt. There have to be very compelling reasons for favoring phylogeny estimates that use morphological and stratigraphical data, each contributing an unknown amount of error, in favor of those based on morphology alone. To date none have been provided.

Finally, if temporal data are involved in defining the optimization criterion by which we select our phylogenetic hypothesis, then it becomes circular to use phylogenetic trees to investigate questions concerning rates of character evolution and other time-related aspects. Adding stratigraphic data needlessly inflates the potential sources of error associated with phylogenetic reconstruction and lessens the value of the resultant trees.

#### REFERENCES

- BENTON, M. J. 1998. The quality of the fossil record of the vertebrates, p. 269–303. *In* S. K. Donovan and C. R. C. Paul (eds.), *The Adequacy of the Fossil Record*. J. Wiley & Sons, Chichester.
- FISHER, D. C. 1992. Stratigraphic parsimony, p. 124–129. *In* W. P. Maddison and D. R. Maddison, *MacClade: Analysis of phylogeny and character evolution*, Version 3. Sinauer, Sunderland, MA.
- FISHER, D. C. 1994. Stratocladistics: Morphological and temporal patterns and their relation to phylogenetic process, p. 133–171. *In* L. Grande and O. Rieppel (eds.), *Interpreting the Hierarchy of Nature: From Systematic Patterns to Evolutionary Process Theories*. Academic Press, San Diego.

- FOOTE, M. 1996. On the probability of ancestors in the fossil record. *Paleobiology*, 22:141–151.
- FOOTE, M. 1997. The evolution of morphological diversity. *Annual Review of Ecology and Systematics*, 28:129–152.
- FOOTE, M. 1998. Estimating taxonomic durations and preservation probability. *Paleobiology*, 23:278–300.
- FOOTE, M., AND D. M. RAUP. 1996. Fossil preservation and the stratigraphic ranges of taxa. *Paleobiology*, 22:121–140.
- FOOTE, M., AND J. J. SEPKOSKI. 1999. Absolute measures of the completeness of the fossil record. *Nature*, 398:415–417.
- FOX, D. L., D. C. FISHER, AND L. R. LEIGHTON. 1999. Reconstructing phylogeny with and without temporal data. *Science* 284:1816–1819.
- HILLIS, D. M. 1991. Discriminating between phylogenetic signal and random noise in DNA sequences, p. 278–294. *In* M. M. Miyamoto and J. Cracraft (eds.), *Phylogenetic Analysis of DNA Sequences*. Oxford University Press, New York.
- HILLIS, D. M., J. P. HUELSENBECK, AND C. W. CUNNINGHAM. 1994. Application and accuracy of molecular phylogenies. *Science*, 264:671–677.
- HUELSENBECK, J. P., AND D. M. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. *Systematic Biology*, 42:247–264.
- MAXWELL, W. D., AND M. J. BENTON. 1990. Historical tests of the absolute completeness of the fossil record of tetrapods. *Paleobiology*, 16:322–335.
- PAUL, C. R. C., AND S. K. DONOVAN. 1998. An overview of the completeness of the fossil record, p. 111–131. *In* S. K. Donovan and C. R. C. Paul (eds.), *The Adequacy of the Fossil Record*. J. Wiley & Sons, Chichester.
- SMITH, A. B. 1994. *Systematics and the Fossil Record: Documenting Patterns of Evolution*. Blackwell's Science, Oxford, 223 p.
- SOLOW, A. R., AND W. SMITH. 1997. On fossil preservation and the stratigraphic ranges of taxa. *Paleobiology*, 23:271–277.
- VALENTINE, J. W. 1970. How many marine invertebrate fossil species? A new approximation. *Journal of Paleontology*, 44:410–415.
- WAGNER, P. 1995. Stratigraphic tests of cladistic hypotheses. *Paleobiology*, 21:153–178.
- WAGNER, P. 1998. A likelihood approach for evaluating estimates of phylogenetic relationship among fossil taxa. *Paleobiology*, 24:430–449.
- WAGNER, P. 1999. The quality of the fossil record and the accuracy of phylogenetic inferences about sampling and diversity. *Systematic Biology*, 49:65–86.

ACCEPTED 7 JUNE 2000