Points of View

Syst. Biol. 46(3):554–563, 1997

Analyzing Large Data Sets: *rbc*L 500 Revisited

KENNETH A. RICE,[1,3] MICHAEL J. DONOGHUE,[1] AND RICHARD G. OLMSTEAD[2]

[1]*Harvard University Herbaria, 22 Divinity Avenue, Cambridge, Massachusetts 02138, USA*
[2]*Department of Botany, University of Washington, Seattle, Washington 98195, USA*

In 1993, Mark Chase and 41 coauthors published phylogenetic analyses of two very large data sets of nucleotide sequences of the chloroplast gene *rbc*L, which encodes the large subunit of ribulose 1,5-bisphosphate carboxylase. Their paper was important for several reasons. These analyses were (and still are) among the largest ever attempted using parsimony. The assembly of such a large number of sequences clearly demonstrated a high level of cooperation on the part of the botanical systematics community. Furthermore, a number of important new hypotheses regarding seed plant phylogeny emerged from this study, and it has helped to orient many subsequent phylogenetic analyses. Increasingly, the Chase et al. trees are being used in quantitative comparative analyses (e.g., Barraclough et al., 1996; also see Donoghue and Ackerly, 1996, and associated papers).

We reanalyzed one of the Chase et al. data sets for two reasons. First, we wanted to explore the general methodological and theoretical issues raised by very large data sets. It is critical that these issues be addressed now because the number of large data sets is increasing rapidly. Second, in view of its importance, we wanted to discover the effects of long search times and alternative search strategies on this data set in particular. We have no desire to quibble over the details of these analyses or their implications for angiosperm phylogeny; instead, we want to focus attention on the special challenges posed by large data sets.

THE CHASE ET AL. ANALYSES

Chase et al. performed two analyses. Search I included 476 sequences, with 1,398 nucleotide sites/characters. A UNIX version of PAUP 3.0r (Swofford, 1990) was used in this analysis, and a transition–transversion step matrix was employed (Albert et al., 1993). Search I ran for approximately 200 hr, yielding 500 shortest trees that were saved and reduced to a consensus tree. Search II included 500 sequences, also of length 1,398. This search was performed with PAUP 3.0s on a Macintosh Quadra computer, and all characters and codon positions were treated equally. A starting tree was obtained by the CLOSEST addition procedure, followed by several runs using NNI and TBR branch swapping (Swofford et al., 1996). This analysis was carried out for approximately 4 weeks, at which time 3,900 trees were saved. These trees were of length 16,305 with uninformative characters excluded using PAUP. One of these trees was shown as the B series trees by Chase et al., with an indication of nodes that were not present in the strict consensus tree.

[3] Present address: Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; E-mail: krice@saul.cis.upenn.edu.

Our Analyses

We obtained the Chase et al. search II data matrix from the authors in machine-readable form, along with the search II consensus tree and the single minimum-length tree presented in the original paper. The data set, as received, included 501 taxa. We made several changes to reproduce the search II data set exactly. First, we removed the *Fuchsia* sequence and switched the labels for *Myrica* and *Celtis*, as did Chase et al. Second, the first 30 bases (the annealing site of the forward amplification primer) were excluded using a PAUP exclusion set, yielding a final length of 1,398 bases. Alignment of *rbc*L sequences is unambiguous because there are no insertions or deletions in this gene. There are missing data for some taxa, especially near the beginning and end of the sequence, owing mainly to different methods of obtaining the sequences (e.g., cloning versus direct sequencing of PCR products).

Using PAUP 3.1.1 (Swofford, 1993) we were able to confirm the length of 16,305 steps reported by Chase et al. when the foregoing modifications were made and uninformative characters were excluded. When uninformative characters are included, tree length is 16,538 steps. In the following discussion, we use only lengths that include all characters because there is no universally accepted method of partitioning characters into informative and uninformative subsets. MacClade (3.04; Maddison and Maddison, 1992), for example, gives a length of 16,225 for the search II data set when "uninformative" characters are excluded.

The issue of informative versus uninformative characters arises again in connection with the branch lengths reported by Chase et al. for the B series trees. Chase et al. calculated branch lengths using PAUP's ACCTRAN optimization. Uninformative characters were included in the analysis, and the branch lengths shown in the figures should therefore sum to 16,538. However, when we added these branch lengths we obtained a length of 16,429, 109 steps fewer than expected. We then repeated the ACCTRAN optimization using PAUP and found a number of small differences that account for the missing 109 steps: the *Symphoricarpos* branch length should read 19 instead of 14 (+5), the *Cornus walteri* branch should be 13 instead of 11 (+2), the *Humulus* branch should be 32 instead of 3 (+29), and the *Cabomba* branch should be 16 instead of 1 (+15; in this case, a 6 appears below the *Cabomba* branch). Removal by Chase et al. of the two rather different *Canella* sequences (labeled A and B), leaving only the branch leading to their common ancestor, accounts for the remaining 58 missing steps (*Canella* A, 14 steps; *Canella* B, 44 steps).

All of our searches were done with PAUP 3.0s running in parallel at minimum priority under L. Oliver and J. Weigert's freeware "Screen" background session manager on three Sun Workstations (one 60-MIPS and two 100-MIPS machines). The searches consumed approximately 90% of the CPU cycles, even though the machines were in daily use by interactive users who did not experience noticeable performance degradation. Running PAUP under a background session manager allowed searches to be brought to the foreground and detached from any remote terminal. This arrangement allowed us to perform longer and more thorough searches than those used in previous efforts.

Chase et al. conducted a single PAUP heuristic search using CLOSEST addition order, STEEPEST DESCENT, a combination of NNI and TBR branch swapping, and a combination of MULPARS and NOMULPARS tree retention. As noted by Baum (1994), the search II strategy guaranteed that only a single island of trees could be found (Maddison, 1991; Page, 1993). In our analyses we did not explore the efficacy of the several search strategies that have previously been proposed for large data sets (e.g., Maddison et al., 1992; Olmstead and Palmer, 1994). We did, however, try eight independently seeded heuristic searches with random taxon addition order, TBR branch swapping, and MULPARS. We also used consensus trees as negative constraints (Swofford, 1993) in

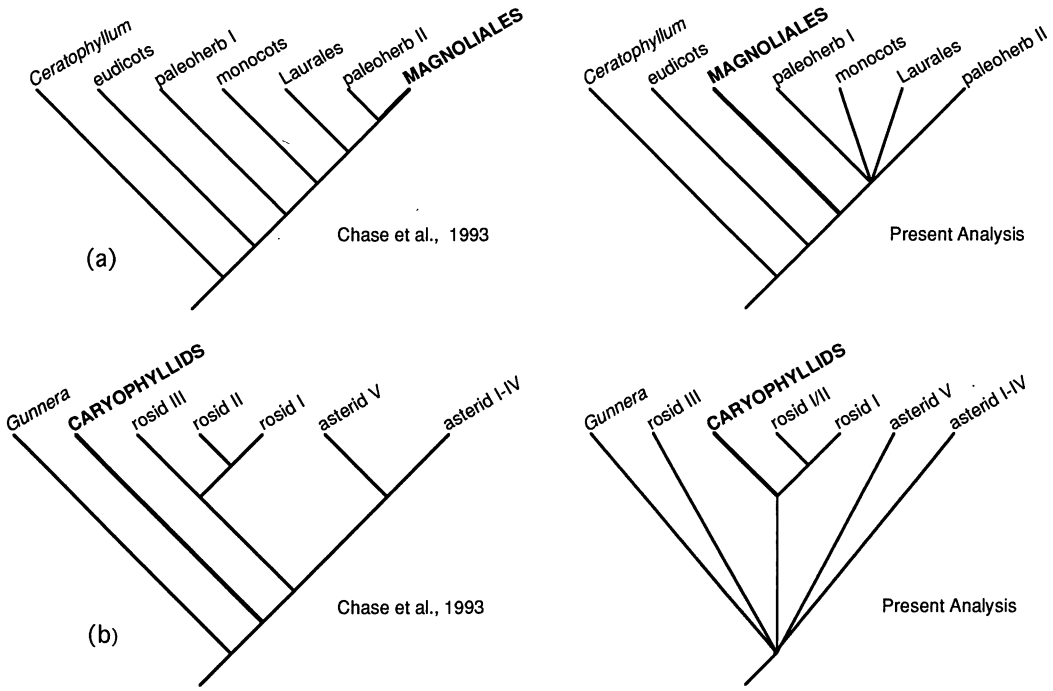

FIGURE 1. Examples of differences between the Chase et al. (1993) B series trees of length 16,538 versus the 16,533-step trees obtained in the present analysis. (a) Shift in the position of Magnoliales. (b) Shift in the position of caryophyllids and the rosid groups (neither rosid I nor rosid II is monophyletic in the present analysis).

two searches that were started with trees obtained from prior analyses (tree B of Chase et al. and a 16,533-step tree obtained from a random addition sequence). This approach was intended to break out of a large neighborhood of equal-length trees without having to search through all trees. As an example of the effectiveness of this approach, when the Chase et al. B tree was used as a starting tree and the strict consensus of the 3,900 trees they obtained was used as the constraint, shorter trees were obtained in <48 hr on a Macintosh Quadra similar to the computer on which the original analysis was run.

RESULTS

Our searches used approximately 11.6 months of CPU time and examined approximately 27.9 billion trees, swapping 2,497 trees to completion, including 283 minimum-length trees (as compared with 3 in the Chase et al. analysis). Searches starting from three of the eight independently seeded replicates identified 21,774 trees shorter than the published tree. We found 8,975 trees of length 16,533, 5 steps shorter than the published trees. These trees, the data matrix, a single tree of 16,533 steps, and a consensus of most-parsimonious trees can be obtained on the Worldwide Web at http://herbaria.harvard.edu/~rice/treezilla/ or by anonymous ftp from ftp://herbaria.harvard.edu/pub/rice/treezilla/ or by sending a diskette to one of us (K.A.R.).

Comparison of a strict consensus of our most-parsimonious trees to the Chase et al. search II consensus reveals many areas of agreement. For example, Gnetales are monophyletic in both consensus trees, and they are still the sister group of angiosperms (in agreement with analyses summarized by Doyle and Donoghue [1993] and Crane et al. [1995]; but see Chaw et al., 1997). As shown in Figure 1, the aquat-

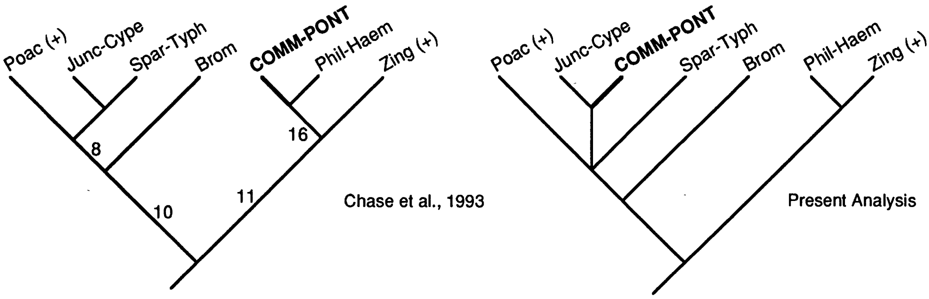

FIGURE 2. Different placement of the Commelinaceae–Pontederiaceae (COMM-PONT) clade in the Chase et al. (1993) B series trees of length 16,538 versus the 16,533-step trees obtained in the present analysis. Numbers associated with the branches indicate the number of character state changes reconstructed in an ACCTRAN optimization. Poac (+) = Poaceae, Eriocaulaceae, Restionaceae, and Flagellariaceae; Junc-Cype = Juncaceae and Cyperaceae; Spar-Typh = Sparganiaceae and Typhaceae; Brom = Bromeliaceae; Phil-Haem = Philydraceae and Haemodoraceae; Zing (+) = Zingiberales, i.e., Zingiberaceae, Strelitziaceae, Marantaceae, Costaceae, Musaceae, Lowiaceae, and Heliconiaceae.

ic plant *Ceratophyllum* remains the sister group of the rest of the angiosperms (see Les et al., 1991), which are then split into a tricolpate (or eudicot) clade (sensu Donoghue and Doyle, 1989; Doyle and Hotton, 1991) and a clade consisting of paleoherbs (including monocots; i.e., sensu Donoghue and Doyle, 1989) and magnoliids. Monocots form a clade, within which *Acorus* is the sister group of the rest (see Duvall et al., 1993). Asterid groups I–IV still form a clade and maintain the same relationships to one another (see Olmstead et al., 1993).

There are 65 topological differences between our best trees and those presented by Chase et al. These differences are widely distributed throughout the tree (i.e., toward the tips, closer to the base, etc.), although there are several sections of the tree in which many changes are concentrated (e.g., within asterid III and rosid I). Some of the changes entail small, relatively local rearrangements, whereas in other cases large clades shift position by a number of nodes. The more major rearrangements include (1) movement of the Magnoliales from within the monosulcate clade (with water lilies and associated taxa) to a basal position within monosulcates; (2) movement of the caryophyllids (the sister group of the rosids and asterids of Chase et al.) to a position nested within rosids, which renders the rosid group (sensu Chase et al.) paraphyletic; and (3) connection of several rosid I clades (e.g., Myrtales) to rosid II groups, such that neither rosid I nor II is monophyletic (Fig. 1). In each of these cases, our results are more similar to those found in the search I analysis of Chase et al.

Many of the differences between our trees and the search II trees involve branches that are supported by only one or two character state changes in the Chase et al. trees. (Branch lengths are the only readily available measure of support; bootstrap and decay analyses were not performed because these are prohibitively time consuming with data sets of this size. Heuristics such as parsimony jackknifing and fast bootstrapping [i.e., without branch swapping] could be used but were not reported by Chase et al.) However, a number of clades that do not appear in the strict consensus of our most-parsimonious trees are marked in the Chase et al. ACCTRAN optimization by greater than the average number of changes on an internal branch (11.7 steps). The movement of the Commelinaceae–Pontederiaceae clade to a position near sedges, cattails, and grasses provides an example of a rearrangement involving branches that seemed to be well supported judging by branch lengths (Fig. 2). The connection of this clade to *Phylidrum* (Phylidraceae) plus *Anigozanthos* (Haemodoraceae) is supported in the Chase et al. ACCTRAN optimization by 16 steps, and this branch is united in turn with the Zingiberales by 11 character changes. A more extreme case concerns the branch uniting one clade of conifers with the anthophytes (Gnetales + angiosperms), which is marked by 22 steps in the Chase et al. tree but does not appear in our more-parsimonious trees. These results support the recommendation of Chase et al. that branch lengths should not be used as a measure of confidence or robustness, despite the evident temptation to do so (e.g., Barraclough et al., 1996).

IMPLICATIONS

It is reasonable to ask whether our trees are really any better than the Chase et al. trees. The trees in our analysis are five steps shorter and are therefore certainly preferable to the degree that investigators believe maximum parsimony to be a reasonable criterion for judging the quality of competing hypotheses (Farris, 1983; Sober, 1988). Those who accept parsimony as an optimality criterion should (until shorter trees are discovered) use our trees instead of those of Chase et al. in designing further phylogenetic analyses, in studies of character evolution, etc.

However, both analyses may be woefully inadequate. For example, two of eight independently seeded random addition searches found 2,265 trees of length 16,536, and 7,429 trees of length 16,535 (respectively 2 and 3 steps shorter than the published tree). These sets of trees differ in significant ways from both the published tree and from the 16,533-step trees, indicating that maximum parsimony (at run times within reach of today's hardware) has poor asymptotic performance when applied to the present data set. That is, additional independently seeded runs are predicted to find trees shorter than our trees of length 16,533, and these trees may differ from ours in significant ways. Under these circumstances, one may wish to refrain from using any of these results and should be suspicious of future parsimony analyses of large *rbc*L data sets.

The heuristic searches in the present analysis examined more than 27.9 billion trees. This sample might appear at first glance to be adequate, but of course it is not. There are about 1.01×10^{1280} distinct trees for 500 taxa (see Felsenstein, 1978a), and 27.9 billion trees constitute a vanishingly small fraction of the entire set. Of course, the size of the search space need not in itself cause a problem because there are local similarity algorithms, such as neighbor joining (Saitou and Nei, 1987), that yield approximations of minimum-distance trees very quickly. However, a factorially large search space poses genuine difficulties for parsimony and maximum likelihood because the globally best solution cannot be obtained by a systematic composition of best solutions to local subproblems. Tree reconstruction algorithms such as maximum parsimony and maximum likelihood are probably members of the set of combinatorially difficult tasks called NP-complete problems (Day, 1983). Problems in this set are widely believed to have no solution for which the expected running time is less than an exponential function of the "size" of the problem (in this case, the number of sequences). There is no mathematical proof of the nonexistence of faster-than-exponential solutions, but an overwhelming majority of computability theorists believe that such solutions do not exist (Garey and Johnson, 1979).

What should investigators do when there are genuinely interesting problems involving hundreds or thousands of taxa? One approach, which we emphatically do not recommend, is to simply continue to add taxa to data sets that are already too large for stable analysis under parsimony or maximum likelihood. This is the main way in which our recommendations diverge from those of Chase et al. They noted that adding more taxa has the potential to improve inferences of relationships, and they envisioned studies that include up to four times as many *rbc*L sequences as the present data set (Chase et al., 1993:543). Adding more taxa in an attempt to break up "long branches" (see Felsenstein, 1978b; Zharkikh and Li, 1993; Hillis et al., 1994) or to increase the sampling density in poorly represented clades is likely to compound the intractable problems encountered with the present data set. Moreover, the probability of inconsistent estimation due to long branches may increase with the addition of taxa (Kim, 1996).

There is, however, a reasonably good chance that the reconstruction problem would become easier with the addition of more characters (Hillis, 1996). Adding more characters should, other things being equal, yield more synapomorphies per clade under the assumption that synapomorphy will, on average, covary more than homoplasy across lineages. Under these circumstances the tree obtained by the initial addition sequence may be closer to maximally parsimonious, and branch swapping may be somewhat faster because there will be (on average) fewer equally parsimonious trees of a given length. Because the number of characters needed to achieve reliable results will often be greater than the number obtained from individual genes (Hillis, 1996), this line of reasoning leads inevitably to combining data from different sources (where combination is deemed reasonable on other grounds; see de Queiroz et al., 1995).

Adding more characters may ameliorate, but in no way solves, the central problem: the optimization of an NP-complete objective function on a data set of large size. When an investigator is confronted with such a problem, there are at present three strategies: (1) reduce the fraction of the search space that must be traversed, (2) reduce the computational complexity of the problem to be solved, or (3) reduce the size of the problem.

In the first case, one can use heuristic search methods to reduce the fraction of the search space sampled, with the hope of finding a globally best solution in a reasonable amount of time. However, studies such as the present one and those of Maddison et al. (1992) and Templeton (1992) have shown that longer running times often find shorter trees. That is, present heuristics may not converge on the globally shortest tree in a reasonable amount of time, which suggests that today's shortest tree will be replaced by tomorrow's even shorter trees if faster hardware and longer run times are employed. A partial solution may be found in the search for better heuristic search algorithms. Some possible avenues of exploration include the judicious use of negative constraint trees and stochastic relaxation algorithms, such as simulated annealing and threshold accepting (Aarts and Korst, 1989; Dueck and Scheuer, 1990).

A second possibility is to reduce the computational complexity of the task by solving a less combinatorially difficult problem such as neighbor joining or parsimony jackknifing (Farris et al., in press). The prospect of hour-long instead of month-long run times is enticing, but using such methods requires abandoning the notion that maximum parsimony or maximum likelihood is the criterion that we are optimizing. Abandoning parsimony and maximum likelihood requires that we rethink the hypotheticodeductive or probabilistic underpinnings of the discipline (e.g., Farris, 1983). There are other reasons for preferring parsimony and maximum likelihood: sets of optimal and nearly optimal trees can be used to explore the neighborhood of a local optimum (Penny et al., 1995), and sets of optimal trees can be reduced to a consensus tree to give an indication of the uncertainty of an analysis. Parsimony jackknifing yields a single, often unresolved tree containing clades that are expected to appear in a strict consensus of all most-parsimonious trees (Farris et al., 1996). It is said to identify clades that are expected to be supported by at least one uncontradicted synapomorphy (Albert et al., unpubl. data; leaflet distributed at the 1995 AIBS/ASPT meeting, San Diego, CA), but this expectation does not always appear to be met in the *rbc*L data set. For example, eudicots appear as a clade in the *rbc*L parsimony jackknife tree (Albert et al., unpubl. data), but none of the 16 characters that change unambiguously on the eudicot branch are uncontradicted (e.g., reversals occur within eudicots). In any case, it is clear that parsimony jackknifing is simply not attempting to solve the problem of finding all globally optimal trees.

A third possibility is to find ways of reducing the problem size such that global maximum-parsimony or maximum-likelihood solutions can be found or such that heuristic parsimony samples a sizable fraction of the search space. Exemplar taxa can be used, i.e., one or a few terminal taxa are chosen to take the place of a major group. More attention needs to be paid to suitable sampling strategies because some are almost certainly better than others (e.g., Yeates, 1995). For example, one could choose as exemplars taxa with the fewest derived states in a group in an effort to minimize branch lengths. Still, if we confine ourselves to choosing among terminal taxa, we will inevitably lengthen branches, with all the attendant problems. Moreover, the use of exemplars discards potentially useful information about inferred ancestral states within a clade. For these reasons, the exemplar strategy probably will be unsatisfactory in the long term.

A more promising avenue is one that might best be called the inferred ancestral states (IAS) approach (cf. the placeholder approach of Donoghue, 1994; compartmentalization of Mishler, 1994; groundplan of Yeates, 1995). Here, large presumptive clades are replaced with a hypothetical ancestor inferred by optimizing characters to the base of a tree for the clade of interest (Fig. 3). The inferred ancestor is then allowed to stand in for the presumptive clade in a larger analysis, which gives computational performance equivalent to replacing a clade by a single exemplar. Inferred ancestors should incorporate polymorphisms where there are equivocal optimizations at the node in question, or the effects of equally parsimonious alternatives might be examined. Doyle et al. (1994) provided a concrete example of this approach in angiosperms (also see Kellogg and Campbell, 1987; Donoghue and Doyle, 1989).

To experiment with the IAS method, we automated it for the Chase et al. data. A chain of programs takes as its input the 500-taxon by 1,398-character matrix and a set of files each containing a preselected NEXUS-format subtree from our first 16,533-step tree. The program then generates a NEXUS file for each subtree and passes the set of these to PAUP. PAUP then optimizes characters on each of the subtrees, generating an inferred ancestral *rbc*L sequence for each, using standard IUPAC ambiguity codes to represent equivocal optimizations. Finally, the program removes the appropriate sequences from the matrix and replaces them with an inferred ancestral sequence. A suite of UNIX programs to accomplish the IAS procedure in connection with PAUP is available as un-

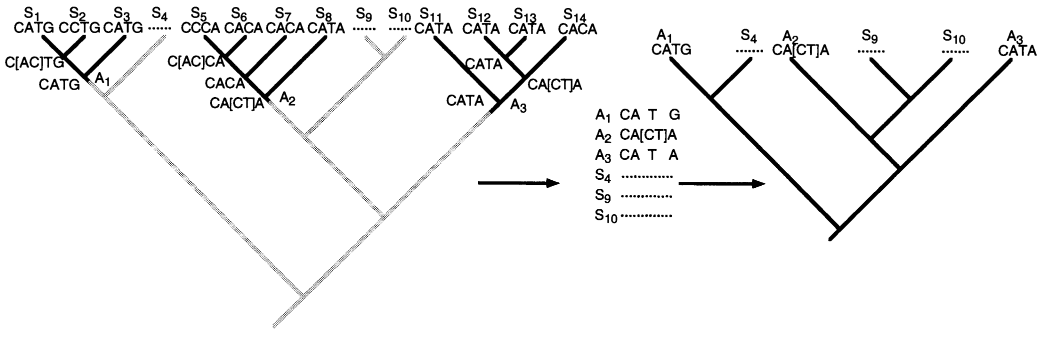

FIGURE 3. An outline of the inferred ancestral states (IAS) procedure. Three presumptive clades of sequences (S_{1-3}; S_{5-8}; S_{11-13}) are reduced to the inferred ancestral sequences labeled A_1, A_2, and A_3 through an optimization procedure. These ancestral sequences are used along with the remaining sequences (S_4, $S_{9,}$ S_{10}) in a new (reduced) analysis. Relationships shaded gray in the larger tree need not be "known" when the ancestral sequences are calculated.

supported source code from one of us (K.A.R.).

Using these programs, we extracted subtrees for 13 large and presumably monophyletic groups found in all of the *rbc*L trees: Asterales, Capparales, Crassulaceae, Cycadales, Lamiidae, Malvales, monocots, Myrtales, Pinaceae, Piperales, Ranunculales, Sapindales, and the Solanaceae–Boraginaceae clade. Altogether, these subtrees contain 229 taxa, and the reduction procedure took about 2 min. The program's output is a new data matrix containing the 271 taxa that remained after the 13 putative clades were extracted, with 13 new sequences representing the inferred ancestors for each of the clades, for a total of 284 sequences. The new data matrix was used for a series of heuristic searches with the same parameters as the 500-taxon searches, and the trees obtained were fully concordant with those obtained from our 500-taxon searches. Reducing the number of taxa from 500 to 284 reduces the number of trees in the problem space from approximately 1×10^{1280} to 2×10^{656}; a reduction of 624 orders of magnitude, implying a proportional increase in the fraction of the problem space searched per unit time. The IAS procedure also speeds branch swapping substantially. For example, whereas a 500-taxon tree was swapped to completion in 52.73 hr on a 100-MIPS workstation (ca. 442.2 million trees evaluated), our 284-taxon tree required only 2.6 hr (ca. 2.3 million trees evaluated). Similar results might also be obtained by simply constraining particular subtrees to be monophyletic, a procedure which would be possible when the phylogeny of the subtree is to be based only on the data under consideration (as opposed to being based on other data).

An inferred ancestor produced by IAS fairly represents both our knowledge and our uncertainty about ancestral character states in particular subtrees while insuring that branches are not lengthened artifactually. If the group being reduced is indeed monophyletic, and if relationships in the subtree are in fact globally parsimonious, then it is theoretically possible (although perhaps not possible in practice) to find globally parsimonious trees using just the inferred ancestral states, and the overall position of the subtree in the more inclusive analysis should be the same as if all of the terminal taxa were included from the outset (Maddison et al., 1984). IAS could allow additional taxa to be added into the calculations within particular subtrees to try to obtain the best possible inferences of ancestral states. Likewise, as stressed by Mishler (1994), more characters could be used for better resolution of relationships within particular subtrees, including characters that are not available or cannot be homologized confidently over the entire data set. At the limit, the phylogenetic relationships assumed to hold within a particular subtree might be based entirely on data other than those being analyzed.

The key to successful application of the IAS approach lies in conducting independent analyses to test hypotheses of monophyly and assumptions about relationships within subtrees (Donoghue, 1994). Confirmation of monophyly might be obtained from other sources, such as morphology, or from other computational methods, such as bootstrapping or parsimony jackknifing. Sensitivity analyses designed to test the robustness of the conclusions obtained through this procedure should also be conducted (e.g., Kellogg and Campbell, 1987; also see Donoghue and Ackerly, 1996). For example, one could test whether other plausible relationships within particular subtrees make a difference in ancestral state assignments and whether any such changes in ancestral states have a substantial impact on the more inclusive analysis.

CONCLUSIONS

The large *rbc*L analyses of Chase et al. provide a point of departure for studies of the theoretical and methodological issues surrounding the analysis of large data sets. The computational difficulties are immense, and adding more taxa to the *rbc*L analysis will simply compound the problem. Instead, for those of us who wish to maintain parsimony and maximum likelihood as criteria for choosing among alternative phylogenetic hypotheses, it will be necessary both to add more characters and to develop better, more explicit methods for reducing the size of analyses. Although the use of exemplars is appealing in its simplicity, it suffers from the same problems as do other methods that knowingly ignore relevant data. More promising are methods such as IAS, which at least indirectly make use of all the available data and take into account whatever is "known" about phylogenetic relationships within well-supported subtrees. Such approaches are in their infancy, and much additional work is needed to ascertain and formalize the best methods. In the meantime, we urge caution in using the results of very large phylogenetic analyses, including our own.

Acknowledgments

We thank Mark Chase and Brent Mishler for sending us the original *rbc*L data set, Victor Albert for his leaflet on the parsimony jackknife analysis, and David Baum, David Hibbett, Toby Kellogg, Michael Sanderson, and Lisa Vawter for helpful discussion. We are also grateful to David Swofford for providing the UNIX version of PAUP and for answering numerous questions.

References

Aarts, E., and J. Korst. 1989. Simulated annealing and Boltzmann machines: A stochastic approach to combinatorial optimization and neural computing. Wiley, New York.

Albert, V. A., M. W. Chase, and B. D. Mishler. 1993. Character-state weighting for cladistic analysis of protein coding DNA sequences. Ann. Mo. Bot. Gard. 80:752–766.

Barraclough, T. G., P. H. Harvey, and S. Nee. 1996. Rate of *rbc*L gene sequence evolution and species diversification in flowering plants (angiosperms). Proc. R. Soc. Lond. B 263:589–591.

Baum, D. 1994. *rbc*L and seed-plant phylogeny. Trends Ecol. Evol. 9:39–41.

Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall, R. A. Price, H. G. Hills, Y.-L. Qiu, K. A. Kron, J. H. Rettig, E. Conti, J. D. Palmer, J. R. Manhart, K. J. Sytsma, H. J. Michaels, W. J. Kress, K. G. Karol, W. D. Clark, M. Hedrén, B. S. Gaut, R. K. Jansen, K.-J. Kim, C. F. Wimpee, J. F. Smith, G. R. Furnier, S. H. Strauss, Q.-Y. Xiang, G. M. Plunkett, P. S. Soltis, S. M. Swensen, S. E. Williams, P. A. Gadek, C. J. Quinn, L. E. Eguiarte, E. Golenberg, G. H. Learn, Jr., S. W. Graham, S. C. H. Barrett, S. Dayanandan, and V. A. Albert. 1993. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbc*L. Ann. Mo. Bot. Gard. 80: 528–580.

Chaw, S.-M., A. Zharkikh, H.-M. Sung, T.-C. Lau, and W.-H. Li. 1997. Molecular phylogeny of extant gymnosperms and seed plant evolution: Analysis of nuclear 18S rRNA sequences. Mol. Biol. Evol. 14:56–68.

Crane, P. R., E. M. Friis, and K. R. Pederson. 1995. The origin and early diversification of angiosperms. Nature 374:27–33.

Day, W. H. E. 1983. Computationally difficult parsimony problems in phylogenetic systematics. J. Theor. Biol. 103:429–438.

de Queiroz, A., M. J. Donoghue, and J. Kim. 1995. Separate versus combined analysis of phylogenetic evidence. Annu. Rev. Ecol. Syst. 26:657–681.

Donoghue, M. J. 1994. Progress and prospects in reconstructing plant phylogeny. Ann. Mo. Bot. Gard. 81:405–418.

Donoghue, M. J., and D. D. Ackerly. 1996. Phylogenetic uncertainties and sensitivity analyses in comparative biology. Philos. Trans. R. Soc. Lond. B 351:1241–1249.

Donoghue, M. J., and J. A. Doyle. 1989. Phylogenetic analysis of angiosperms and the relationships of Hamamelidae. Pages 17–45 *in* Evolution, systematics and fossil history of the Hamamelidae (P. Crane and S. Blackmore, eds.). Clarendon Press, Oxford, England.

Doyle, J. A., and M. J. Donoghue. 1993. Phylogenies and angiosperm diversification. Paleobiology 19: 141–167.

Doyle, J. A., M. J. Donoghue, and E. A. Zimmer. 1994. Integration of morphological and ribosomal RNA data on the origin of angiosperms. Ann. Mo. Bot. Gard. 81:419–450.

Doyle, J. A., and C. L. Hotton. 1991. Diversification of early angiosperm pollen in a cladistic context. Pages 169–195 *in* Pollen and spores: Patterns of diversification (S. Blackmore and S. Barnes, eds.). Clarendon Press, Oxford, England.

Dueck, G., and T. Scheuer. 1990. Threshold accepting: A general purpose optimization algorithm appearing superior to simulated annealing. J. Comp. Physics 90:161–175.

Duvall, M. R., M. T. Clegg, M. W. Chase, W. D. Clark, W. J. Kress, H. G. Hills, L. E. Eguiarte, J. F. Smith, B. S. Gaut, E. A. Zimmer, and G. H. Learn, Jr. 1993. Phylogenetic hypotheses for the monocotyledons constructed from *rbc*L sequence data. Ann. Mo. Bot. Gard. 80:607–619.

Farris, J. S. 1983. The logical basis of phylogenetic analysis. Pages 7–36 *in* Advances in cladistics, Volume 2 (N. Platnick and V. Funk, eds.). Columbia Univ. Press, New York.

Farris, J. S., V. A. Albert, M. Källersjö, D. Lipscomb, and A. G. Kluge. 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12:99–124.

Felsenstein, J. 1978a. The number of evolutionary trees. Syst. Zool. 27:27–33.

Felsenstein, J. 1978b. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

GAREY, M. R., AND D. S. JOHNSON. 1979. Computers and intractability: A guide to the theory of NP-completeness. W. H. Freeman, San Francisco.

HILLIS, D. M. 1996. Inferring complex phylogenies. Nature 383:130–131.

HILLIS, D. M., J. P. HUELSENBECK, AND C. W. CUNNINGHAM. 1994. Application and accuracy of molecular phylogenies. Science 264:671–677.

KELLOGG, E. A., AND C. S. CAMPBELL. 1987. Phylogenetic analyses of the Gramineae. Pages 310–322 *in* Grass systematics and evolution (T. Soderstrom, K. Hilu, C. Campbell, and M. Barkworth, eds.). Smithsonian Institution Press, Washington, D.C.

KIM, J. 1996. General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa. Syst. Biol. 45:363–374.

LES, D., D. K. GARVIN, AND C. F. WIMPEE. 1991. Molecular evolutionary history of ancient aquatic angiosperms. Proc. Natl. Acad. Sci. USA 88:10119–10123.

MADDISON, D. R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. Syst. Zool. 40:315–328.

MADDISON, D. R., M. RUVOLO, AND D. L. SWOFFORD. 1992. Geographic origins of human mitochondrial DNA: Phylogenetic evidence from control region sequences. Syst. Biol. 41:111–124.

MADDISON, W. P., M. J. DONOGHUE, AND D. R. MADDISON. 1984. Outgroup analysis and parsimony. Syst. Zool. 33:83–103.

MADDISON, W. P., AND D. R. MADDISON. 1992. MacClade: Interactive analysis of phylogeny and character evolution, version 3.04. Sinauer, Sunderland, Massachusetts.

MISHLER, B. D. 1994. Cladistic analysis of molecular and morphological data. Am. J. Phys. Anthropol. 94: 143–156.

OLMSTEAD, R. G., B. BREMER, K. M. SCOTT, AND J. D. PALMER. 1993. A parsimony analysis of the Asteridae sensu lato based on *rbc*L sequences. Ann. Mo. Bot. Gard. 80:700–722.

OLMSTEAD, R. G., AND J. D. PALMER. 1994. Chloroplast DNA systematics: A review of methods and data analysis. Am. J. Bot. 8:1205–1224.

PAGE, R. D. M. 1993. On islands of trees and the efficiency of different methods of branch swapping in finding most-parsimonious trees. Syst. Biol. 42:200–210.

PENNY, D., M. A. STEEL, P. J. WADDELL, AND M. D. HENDY. 1995. Improved analyses of human mtDNA sequences support a recent African origin for *Homo sapiens*. Mol. Biol. Evol. 12:863–882.

SAITOU, N., AND M. NEI. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

SOBER, E. 1988. Reconstructing the past: Parsimony, evolution, and inference. MIT Press, Cambridge, Massachusetts.

SWOFFORD, D. L. 1990. PAUP: Phylogenetic analysis using parsimony, version 3.0. Illinois Natural History Survey, Champaign.

SWOFFORD, D. L. 1993. PAUP: Phylogenetic analysis using parsimony, version 3.1.1. Illinois Natural History Survey, Champaign.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

TEMPLETON, A. 1992. Human origins and analysis of mitochondrial DNA sequences. Science 255:737–739.

YEATES, D. K. 1995. Groundplans and exemplars: Paths to the tree of life. Cladistics 11:343–357.

ZHARKIKH, A., AND W.-H. LI. 1993. Inconsistency of the maximum-parsimony method: The case of five taxa with a molecular clock. Syst. Biol. 42:113–125.

Received 5 August 1996; accepted 18 October 1996
Associate Editor: David Cannatella

Syst. Biol. 46(3):563–569, 1997

Stratigraphic Indices and Tree Balance

REBECCA HITCHIN AND MICHAEL J. BENTON

Department of Geology, University of Bristol, Wills Memorial Building, Queens Road, Bristol BS8 1RJ, England;
E-mail: r.hitchin@bristol.ac.uk (R.H.), mike.benton@bristol.ac.uk (M.J.B.)

Siddall (1996) claimed that (1) the number of internal nodes in a cladogram (n) is correlated with tree balance (Heard's [1992] index of imbalance, Im), (2) Huelsenbeck's (1994) stratigraphic consistency index (SCI) is correlated with the reciprocal of n (Siddall, 1996: fig. 1), and (3) SCI is therefore correlated with Im (Siddall, 1996: fig. 6).

Siddall tested these assertions with a database of 14 cladograms taken from Huelsenbeck's (1994) original SCI analysis and