Pacific Symposium on Biocomputing 6:547-558 (2001)

SCALING OF ACCURACY IN EXTREMELY LARGE PHYLOGENETIC TREES

O. R. P. BININDA-EMONDS[a,b]
Section of Evolution and Ecology, University of California, Davis, CA 95616, USA

S. G. BRADY[a]
Center for Population Biology and Department of Entomology, University of California, Davis, CA 95616, USA

J. KIM
Department of Ecology and Evolutionary Biology and Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06511, USA

M. J. SANDERSON
Section of Evolution and Ecology, University of California, Davis, CA 95616, USA

The accuracy of phylogenetic inference was examined in simulated data sets up to nearly 10,000 taxa, the size of the largest set of homologous genes in existing molecular sequence databases. Even with a simple search algorithm (maximum parsimony without branch swapping), the number of characters needed to estimate 80% of a tree correctly can scale remarkably well at optimal substitution rates (on the order of log *N*, where *N* is the number of taxa). In other words, the number of taxa in an analysis can be doubled and only an arithmetic increase in the number of characters is required to maintain the same level of accuracy. Even substitution rates that are much higher than normally used in phylogenetic studies did not affect the scaling too adversely. However, scaling is usually worse than log *N* for more stringent levels of accuracy. Moreover, errors are not distributed randomly throughout the tree. Shallow nodes are remarkably easy to reconstruct and display favourable log-linear scaling. The deepest nodes are extremely difficult to reconstruct accurately, even with branch swapping, and the scaling is poor. Therefore, the strategy of sequencing large numbers of homologous genes may not always provide global solutions to extreme phylogenetic problems and alternative strategies may be required.

1. Introduction

The size and scope of phylogenetic analysis has changed dramatically in recent years. Advances in DNA sequencing technology now permit the assembly of very large sets of homologous character data, such as the Ribosomal Database Project (RDP II), which contains over 10,000 sequences of aligned small subunit rDNAs [1]

[a] Both authors contributed equally to this work.
[b] Current address: Institute of Evolutionary and Ecological Sciences, Leiden University, Kaiserstraat 63, Postbox 9516, 2300 RA Leiden, The Netherlands

sampled across all of life. Phylogenetic studies of hundreds [2-6] to upwards of several thousand organisms [7] have been undertaken, made feasible by new algorithms such as parsimony jackknifing,[8] different search strategies such as compartmentalization,[9] and faster computers. Typically, however, the number of characters used in these studies (i.e., the length of the sequences) has been of the same order as the number of sequences, N. Let C_X be the number of characters needed to reconstruct a percentage, X, of the bipartitions of a tree correctly. If C_X scales linearly with N, but the ultimate goal of systematics is to reconstruct the tree of life with its 10,000,000 extant species or more, then whole-genome quantities of sequence data would be needed for *each* of these species. Understanding of whether C_X scales "well", meaning *better* than linearly, will therefore impinge on sequencing strategies in the very large phylogenetic studies that are likely in the near future.

Linear scaling is intuitively reasonable based on the argument that, even in the absence of homoplasy (seemingly the ideal condition for phylogenetic inference), at least one character is needed to identify one clade in a tree using maximum parsimony.[10, 11: 351] Linear scaling is also implied by the observation that, again in the absence of homoplasy, a constant number of characters per clade is needed for every clade to be supported at some fixed level in bootstrap tests of tree reliability (e.g., three characters for 95% support [12]). Surprisingly, however, homoplasy can actually add information,[13, 14] and information from theoretical arguments have suggested that a theoretical lower bound on "complete accuracy", C_{100}, scales as log N, which is better than linear, for a broad class of tree generation models and a Markov model of substitution.[15, 16] However, no tree-building algorithm has been shown to achieve this lower bound, and scaling of algorithms currently in wide use with real data, such as maximum parsimony, has not been investigated. Nonetheless, recent simulation studies have shown that some *specific* large trees are unexpectedly easy to infer,[14] and that some taxon sampling schemes can improve accuracy even with a fixed amount of sequence data by breaking up long, misleading branches.[17-19] However, these results may be contingent on properties of the data or the specific parameter values defining a class of simulated trees.[20] Theoretical arguments on different classes of trees suggest that reconstructing phylogenies should get increasingly difficult as the number of taxa is increased.[15, 16, 21-23]

Relatively little is known empirically about how the accuracy of tree-building algorithms scales as the number of taxa is increased. Yet, the scaling properties of accuracy are important because the solution to a large phylogenetic problem may require the deployment of radically different strategies for sequencing, mapping, and genomic analyses, depending on whether it is more desirable to add new taxa or new sequences to a given data set.[18, 24] Here we use simulation to investigate how phylogenetic accuracy in the broad sense scales over an extremely large range of taxa up to about the size of the Ribosomal Database Project at ~10,000 taxa. We also examine how accuracy and the scaling thereof varies according to the region of the tree being examined.

2. Methods

To study the scaling properties of accuracy in phylogenetic inference, we generated a class of model trees according to a stochastic Yule birth process, conditioned on a fixed number of terminal taxa and a fixed time between the root of the tree and the present.[25] This model guarantees that the age distribution of nodes (and hence the fundamental "shape") of trees was invariant to number of taxa, isolating possible confounding factors that might influence scaling properties. Model trees were constructed using the default parameters of the YULE_C procedure in the computer program *r8s* (available from http://loco.ucdavis.edu/r8s/r8s.html). This model is different from some others used to study accuracy in that it implements complete sampling of a clade of designated size rather than random subsampling from a much larger clade.[e.g., 19]

We investigated how accuracy scales over an extremely large range of taxa, N, that varied on a $\log_2$ scale from 4 to 8,192. Nucleotide sequences were evolved down the model trees according to a standard Markov process model, including site-to-site variation and different transition-transversion rates using the computer program Seq-Gen 1.1.[26] We modified Seq-Gen so that it could simulate more than 1,000 sequences (the default upper limit) and would accept a user-input random number seed to allow us to exactly replicate any runs. Sequences were generated under a Kimura 2-parameter model [27] with several different values for transition/transversion ratio (ti:tv), site-to-site rate heterogeneity, and rates of evolution. We initially used ti:tv ratios of 2.0 or 8.0, with rate heterogeneity being either present (with shape parameter of 0.5) or not. However, the majority of our results were obtained using ti:tv of 2.0 with rate heterogeneity.

Branch lengths were determined assuming a model of substitution that departs from a molecular clock. Branch-specific rates of evolution were determined by drawing random normal variates (mean of 1.0 and standard deviation of 0.5, truncated outside of [0.1, 2.0]) and multiplying by an overall tree-wide rate of substitution. Branch lengths were determined by multiplying branch-specific rates by branch durations generated by the Yule process model (see above).

The number of characters was normally varied over a range from 200 to 6,000. Overall average rates of evolution were varied from 0.01 to 0.50 substitutions/site, measured along a path from the root to a tip of the tree. These parameters span the range of values generally considered useful in phylogenetic work.[28, 29] Phylogenetic trees were inferred with maximum parsimony, a method in widespread use for real data,[27] using a fast heuristic algorithm (random addition sequence and no branch swapping) in PAUP* v4.0b2.[30]

We obtained values of C_x from the simulations by fitting a quadratic regression to the data on accuracy versus number of characters and then determining the value of C needed to achieve a pre-specified accuracy from the regression coefficients (see

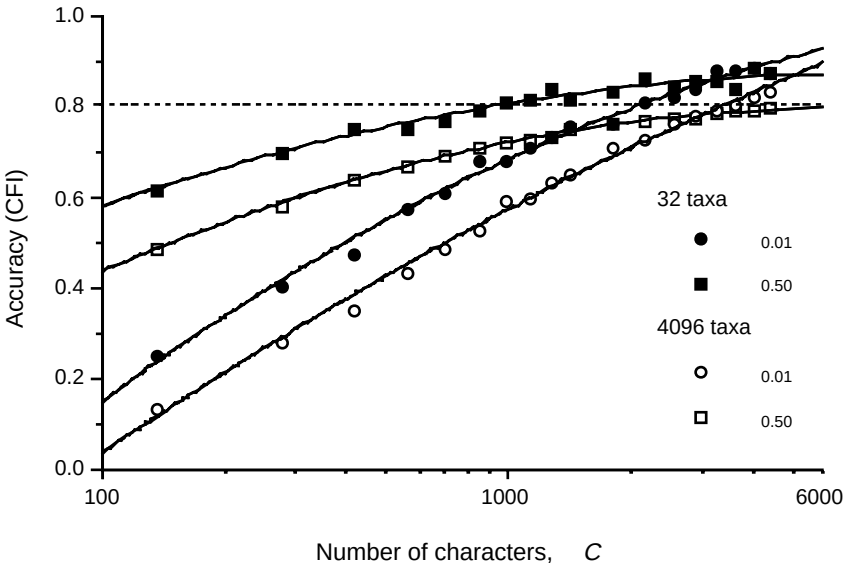

Fig. 1. Quadratic curves used to interpolate the number of characters required to achieve 80% accuracy (C_{80}) for two different rates of evolution for trees of 32 and 4,096 taxa.

Fig. 1). Each value is the mean over a number of replicate simulations (25 replicates for $N \bullet 32$; 10 for $N = 64$ or 128; 3 for $N \bullet 256$; the lower number of replicates for higher values of N did not affect confidence levels adversely). Accuracy was examined at different depths in a tree by classifying nodes (bipartitions) into "levels". Classification was based on the size of the bipartitions (where size is min(j, N–j), and j is the number of taxa in one of the two partitions). The kth level includes all clusters of size 2^k to 2^{k+1}–1 (with the last level defined such that $2^{k+1} = N/2$; those rare bipartitions that divide the tree into two groups of exactly $N/2$ taxa are added to the last level). We define "shallow" bipartitions as those from the first level (always three or fewer taxa), "deep" as those from the second level and upwards (always four or more taxa), and "deepest" as those from the last level, which varies in size depending on the size of the tree. On a rooted version of these trees, the size of groups is a reasonable surrogate for depth in the tree, because the expected clade size is a monotonic function of the rate of diversification, λ, (assumed constant) and the age of the clade, T, so that $E(N) = e^{\lambda T}$.

Most recent studies of accuracy have used measures based on the number of common bipartitions (= proportion of correctly inferred clades), such as the partition metric, d_s,[31] or consensus-fork index, CFI.[32, 33] We used the CFI instead of d_s because it yielded greater discrimination given that our model trees are always

strictly bifurcating. Low values for d_S can only occur if the estimated trees are largely resolved but disagree strongly with the model tree. For poorly resolved estimated trees, d_S tends towards 50% (i.e., recognizes the clades from the model tree). In contrast, both polytomies on the estimated trees and clades whose membership differs between the model and estimated trees are counted as incorrect using the CFI. More discussion regarding the importance of the metric used to measure accuracy can be found elsewhere.[19]

3. Results and Discussion

The scaling of accuracy with taxa depends on the specified accuracy level and the rate of evolution (Figs. 2 and 3). Scaling is quite favorable at an accuracy level of 70% (Fig. 2). Regardless of the rate of evolution, C_{70} scales log-linearly with N, although the slope for the faster rate is slightly greater. Even for 8,192 taxa, no more than 1,000 characters were needed to achieve 70% accuracy at either rate. Thus a doubling in the number of taxa requires only an arithmetic increase in the number of characters. This result directly contradicts the intuition about linear scaling of

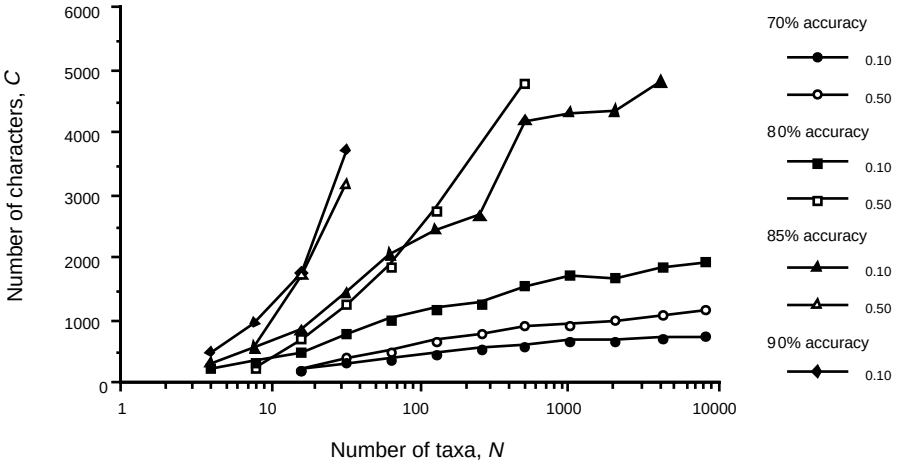

Fig. 2. Number of characters needed to achieve different levels of accuracy, as a function of numbers of taxa for two different rates of substitution. Accuracy is measured by the consensus fork index (CFI), which is the proportion of bipartitions in common between the estimated and true (model) trees. Rates of substitution are per site along the path from root to tip of the model trees.

accuracy, and shows that at least over the range examined in these simulations, it is possible for a simple heuristic maximum parsimony algorithm to achieve the theoretical lower bound on accuracy as long as 100% accuracy is not demanded. However, as the desired accuracy level becomes more stringent, the scaling becomes less favorable, particularly at the faster rate (Fig. 2). At higher specified accuracies, the scaling is worse than log-linear (at C_{80} for the faster rate and C_{90} for the slower rate), and it was often not possible to reach the desired accuracy level even with 6,000 characters. Better accuracy given the same number of characters was generally obtained with models that incorporated a reduced transition-transversion ratio, no rate variation across sites, or an assumption of a molecular clock, especially at higher rates. However, the fundamental scaling trends were robust to all these modifications.

The form of the scaling function depends on the rate of substitution (Fig. 3). An "optimal rate" of ~0.05 – 0.10 substitutions per site produces log-linear scaling of C_{80} with a shallow slope. These rates are comparable to rates observed, for example, for nonsynonymous substitutions in many chloroplast genes sampled at the level of seed plants. For an 8,192 taxon tree, C_{80} was fewer than 2,000 characters. Rates that were either slower or faster than optimal displayed poorer performance. This is particularly true of the fastest rate of 0.50, which scaled worse than log-linearly, and the slowest rate of 0.01, which, although scaling "better" than log-linearly, generally required more than three times as many characters to achieve C_{80} as did the optimal rates. Although the optimal rate will depart somewhat from theoretical values in real data, a good strategy for selecting potential genes for phylogenetic analysis may be to err in the direction of genes that evolve slightly faster than optimal (except in the

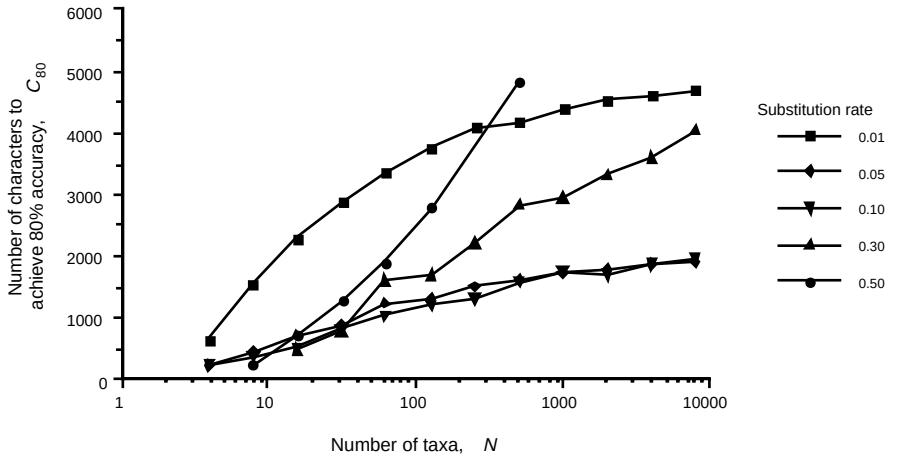

Fig. 3. Scaling of C_{80}, the number of characters needed to achieve 80% accuracy, as a function of numbers of taxa for five different rates of substitution.

very largest trees). The decrease in performance using faster genes is much more gradual than that for slower genes.[34; also Fig. 5 below]

It may seem counterintuitive that rates that are slightly faster than optimal demonstrate better scaling than those that are slightly slower. However, limited amounts of additional homoplasy may sometimes be beneficial because homoplasies (i.e., convergences or reversals) can act as homologies at less inclusive levels. Consider the case presented in Fig. 4. Without any homoplasy (*A*), it requires a minimum of 14 binary characters, one for each informative node on the tree, to

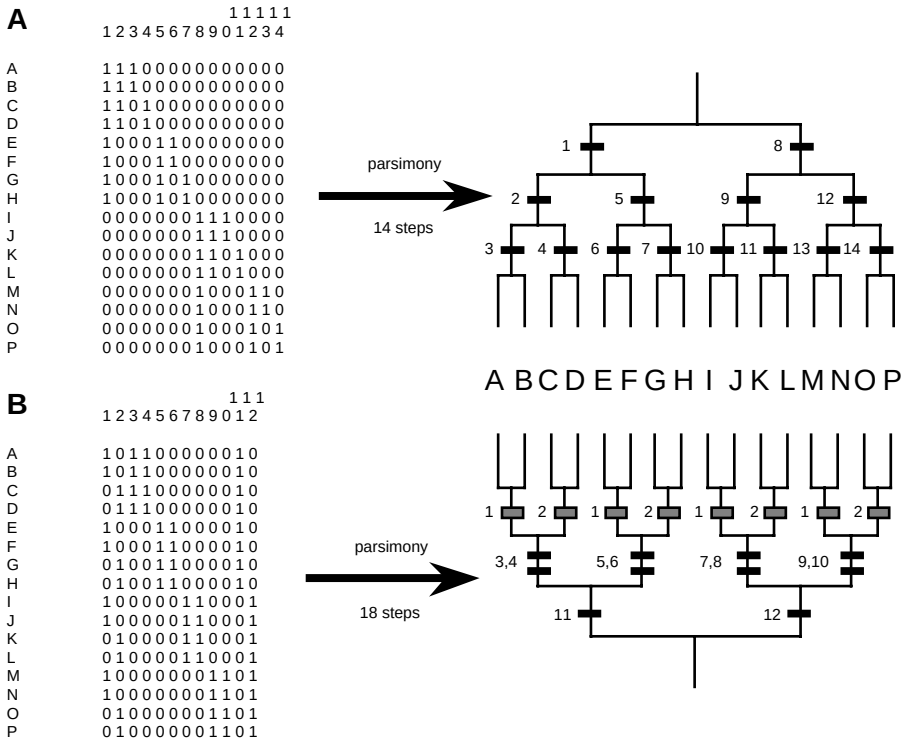

Fig. 4. How homoplasy can improve character scaling. In matrix *A*, which contains no homoplasy, 14 characters are required (one for each informative node). Matrix *B* does contain homoplasy in the form of convergence (characters 1 and 2; hatched bars), but it is outweighed by characters 3–12 (solid bars) and requires two fewer characters to specify the same most parsimonious tree as *A*.

reconstruct this tree accurately. Now consider two homoplastic characters marked

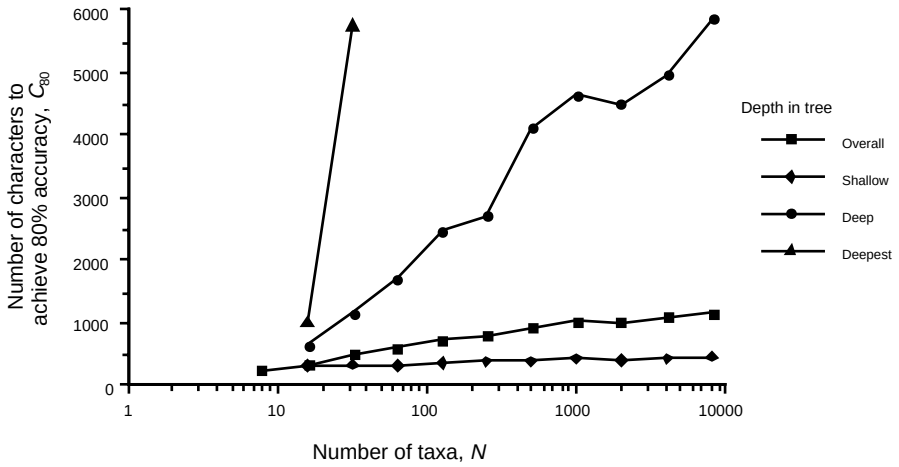

Fig. 5. Scaling of C_{80} in different regions of the tree (rate of substitution = 0.10). See text for definitions of depth categories.

with hatched bars with a distribution as given in *B*. Although they are homoplastic with respect to the entire tree, each character supports four clades if there is sufficient evidence to outweigh them (here, the remaining characters, marked with solid bars). Thus, even with the two homoplastic characters, only 12 binary characters are required altogether, a savings of two characters. The reduction in the number of characters required could be even greater for DNA sequence data given that it is a four-state character and can thus accommodate more homoplasy per character than the binary character used in our example. The effect should also be more pronounced in larger trees because of the potential for a single homoplastic character to support that many more clades.

Scaling of accuracy varies dramatically across the tree (Fig. 5). Shallow nodes were by far the easiest to reconstruct, requiring fewer than 500 characters regardless of the size of the tree. Deeper nodes were more difficult. Although the scaling was still log-linear, the slope for “deep” nodes was fairly steep—nearly 6,000 characters were required for trees with 8,192 taxa. The “deepest” nodes on a tree were even more difficult to reconstruct. Even with 30,000 characters, 80% accuracy could not be obtained in the deepest nodes of a 64 taxon tree. These nodes were difficult regardless of the rate of evolution (Fig. 6). For all parts of the tree, the optimum rate was about 0.05 – 0.10, with accuracy being substantially lower at slower rates. In most cases, accuracy decreased only marginally as the rate was increased from the optimum (cf. preceding paragraph and Fig. 3); only the deepest nodes displayed a large decrease. There was virtually no effect of tree size on the accuracy of shallow

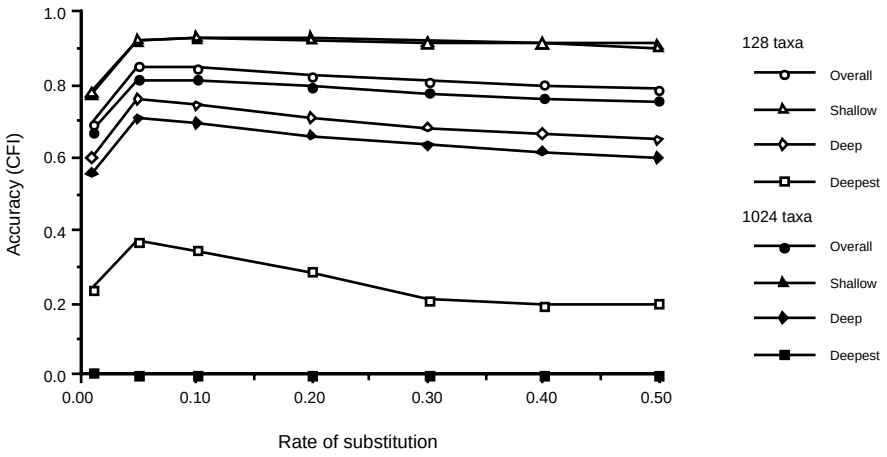

Fig. 6. Accuracy in different regions of trees of 128 or 1,024 taxa as a function of the rate of substitution.

nodes and only marginal decreases in accuracy for the larger tree for overall and deep nodes. However, tree size greatly affected the accuracy of reconstructing the deepest nodes. For 1,024 taxon trees virtually none of the deepest nodes were inferred correctly, regardless of the rate of evolution. Note that shallow nodes comprise the largest proportion of a tree's nodes (e.g., in a perfectly symmetrical tree, 75% of all nodes will possess four or fewer terminal taxa). Thus, earlier findings that measure accuracy based on overall fraction of bipartitions correct [14, 19] should be interpreted with a degree of caution: the high overall levels of accuracy may be accounted for mainly by shallow groups that are relatively easy to reconstruct.

One factor that contributes to the extreme difficulty of inferring the deepest nodes in a tree is the "greediness" of the simple search algorithm used, which is fast but crude. Once a cluster is constructed in the sequential addition process, taxa are not removed. Any errors made during earlier addition events will accumulate progressively and be reflected in high probabilities of errors at the deepest nodes. At substantial cost in running times, rearrangements in topologies ("branch swapping") can be made, which often improves accuracy. Because of the computational burden imposed by branch swapping, this was examined only in trees of 128 taxa. The improvement is slight for shallow nodes but significant for the deepest nodes (Fig. 7).

Unfortunately, search strategies involving branch swapping quickly become impractical in large data sets. Our basic search strategy, which involves the

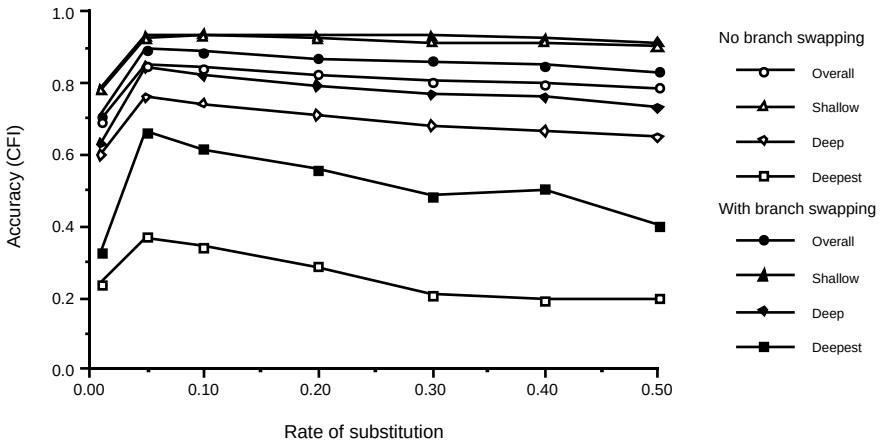

Fig. 7. The effect of branch swapping on accuracy of reconstructing different regions of a 128 taxon tree (with 2,000 characters). Branch swapping used the tree-bisection-reconnection algorithm in PAUP* 4.0 b2,[29] with the maximum number of trees saved set to 100.

sequential addition of randomly chosen taxa (optimized at each step) has a running time that scales as $O(N^2)$, which is probably about as good as any nontrivial tree-building algorithm can be. Heuristic search strategies involving branch swapping often scale as a higher order polynomial, or, in the worst case, require a search (in exponential time) of the entire space of trees. Reported experiments with large data sets confirm this behavior. Branch swapping in a parsimony search using 500 nucleotide sequences of the chloroplast gene *rbc*L lasted about four weeks [2] before being terminated prior to completion. When Rice *et al*. [35] re-analyzed the data, it required 11.6 months of CPU time, mainly continued branch swapping (again, not to completion), to derive a solution that was slightly shorter (five steps or 0.03%). In extremely large data sets, the necessity to branch swap on the one hand, coupled with its computational burden on the other hand, argues against a strategy of obtaining large numbers of sequences from many taxa and combining them in a single phylogenetic analysis. Instead, to achieve accurate inferences across the entire tree, a compartmentalization strategy [9] may be necessary in which well supported subtrees are identified by quick heuristic searches (e.g., parsimony jackknifing [8]) and their topologies determined through detailed analysis to form synthetic representative taxa which are then used as higher taxa in a final, more exhaustive, phylogenetic analysis.

These results were derived using a model in which trees scale in size in a way that preserves the age distribution of node times and relative branch lengths. Other models can introduce scaling that is better or worse for accuracy. A (backward Yule

process) model in which node times are normalized to the age of the root generates trees that are progressively easier to estimate as they grow—leading to better scaling properties—because they tend to be progressively more top-heavy with a preponderance of recent divergence events. Alternatively, a model in which a very large clade is sampled randomly can produce topologies that are unusually difficult when the number of taxa in the final sample is small, because internodes are very short deep in the tree leading to many long branches.[e.g., 19] Patterns of mass extinction or differential background extinction are likely to affect tree topologies in interesting ways that have not yet been examined. For example, non-homogeneous extinction could render some early nodes easier to estimate than expected if extinction were constant because it may leave some branches with enough information to signal the monophyly of those groups. It is unlikely that one pattern of scaling of accuracy will hold true across all models of scaling of tree size. Overall, however, shallow nodes will be surprisingly easy to infer, whereas deep ones will be exceedingly hard, which will pose important methodological challenges to the use of burgeoning comparative sequence databases for large-scale phylogeny reconstruction.

Acknowledgements

We thank Byron Adams, Bob Kuzoff, Steve Nadler, Ashleigh Smythe, Phil Ward, and the UCD Phylogenetics Discussion group for comments. OBE was supported by an NSERC Postdoctoral Fellowship.

References

1. B.L. Maidak et al., *Nucleic Acids Res.* **27,** 171 (1999).
2. M.W. Chase et al., *Ann. Mo. Bot. Gard.* **80,** 528 (1993).
3. Y. Van de Peer and R. de Wachter, *J. Mol. Evol.* **45,** 619 (1997).
4. R.M. Bush et al., *Mol. Biol. Evol.* **16,** 1457 (1999).
5. P.S. Soltis et al., *Nature* **402,** 402 (1999).
6. V. Savolainen et al., *Syst. Biol.* **49,** 306 (2000).
7. M. Källersjö et al., *Pl. Syst. Evol.* **213,** 259 (1998).
8. J.S. Farris et al., *Cladistics* **12,** 99 (1996).
9. B.D. Mishler, *Am. J. Phys. Anthropol.* **94,** 143 (1994).
10. J.A. Hendrickson, Jr., "*A methodological analysis of numerical cladistics"* (Ph.D. dissertation, University of Kansas, 1967).
11. P.H.A. Sneath and R.R. Sokal, *Numerical taxonomy: the principles and practice of numerical classification* (W.H. Freeman and Company, San Francisco, 1973).

12. J. Felsenstein, *Evolution* **39,** 783 (1985).
13. M. Källersjö et al., *Cladistics* **15,** 91 (1999).
14. D.M. Hillis, *Nature* **383,** 130 (1996).
15. P.L. Erdös et al., *Random Struc. Alg.* **14,** 153 (1999).
16. P.L. Erdös et al., *Theoret. Comput. Sci.* **221,** 77 (1999).
17. M.D. Hendy and D. Penny, *Syst. Zool.* **38,** 297 (1989).
18. A. Graybeal, *Syst. Biol.* **47,** 9 (1998).
19. B. Rannala et al., *Syst. Biol.* **47,** 702 (1998).
20. Z. Yang and N. Goldman, *Trends Ecol. Evol.* **12,** 357 (1997).
21. J. Kim, *Syst. Biol.* **45,** 363 (1996).
22. J. Kim, *Syst. Biol.* **47,** 43 (1998).
23. K. Strimmer and A. von Haesler, *Syst. Biol.* **45,** 516 (1996).
24. S. Poe and D.L. Swofford, *Nature* **398,** 299 (1999).
25. S.M. Ross, *Stochastic processes* (Wiley, New York, 1996).
26. A. Rambaut and N.C. Grassly, *Comput. Appl. Biosci.* **13,** 235 (1997).
27. D.L. Swofford et al. in *Molecular systematics*, Eds. D.M. Hillis et al. (Sinauer Associates, Inc., Sunderland, Massachusetts, 1996).
28. W.-H. Li, *Molecular evolution* (Sinauer Associates, Inc., Sunderland, Massachusetts, 1997).
29. R.G. Olmstead and J.D. Palmer, *Am. J. Bot.* **81,** 1205 (1994).
30. D.L. Swofford, *PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4* (Sinauer Associates, Sunderland, Massachusetts, 1999).
31. D.F. Robinson and L.R. Foulds, *Math. Biosci.* **53,** 131 (1981).
32. D.H. Colless, *Syst. Zool.* **29,** 288 (1980).
33. D.H. Colless, *Syst. Zool.* **30,** 325 (1981).
34. Z. Yang, *Syst. Biol.* **47,** 125 (1998).
35. K.A. Rice et al., *Syst. Biol.* **46,** 554 (1997).