6 Taxon Sampling versus Computational Complexity and Their Impact on Obtaining the Tree of Life

O. R. P. Bininda-Emonds
Institut für Spezielle Zoologie und Evolutionsbiologie mit Phyletischem Museum, Friedrich-Schiller-Universität Jena, Germany

A. Stamatakis
Swiss Federal Institute of Technology, School of Computer and Communication Sciences, Lausanne, Switzerland

CONTENTS

ABSTRACT

The scope of phylogenetic analysis has increased greatly in the last decade, with analyses of hundreds, if not thousands, of taxa becoming increasingly common in our efforts to reconstruct the tree of life and study large and species rich taxa. Through simulation, we investigated the potential to reconstruct ever larger portions of the tree of life using a variety of different methods

(maximum parsimony, neighbour joining, maximum likelihood and maximum likelihood with a divide-and-conquer search algorithm). For problem sizes of 4, 8, 16 … 1,024, 2,048 and 4,096 taxa sampled from a model tree of 4,096 taxa, we examined the ability of the different methods to reconstruct the model tree and the running times of the different analyses. Accuracy was generally good, with all methods returning a tree sharing more than 85% of its clades with the model tree on average, regardless of the size of the problem. Unsurprisingly, analysis times increased greatly with tree size. Only neighbour joining, by far the fastest of the methods examined, was able to solve the largest problems in under 12 hours. However, the trees produced by this method were the least accurate of all methods (at all tree sizes). Instead, the strategy used to sample the taxa had a larger impact on both accuracy and, somewhat unexpectedly, analysis times. Except for the largest problem sizes, analyses using taxa that formed a clade generally both were more accurate and took less time than those using taxa selected at random. As such, these results support recent suggestions that taxon number in and of itself might not be the primary factor constraining phylogenetic accuracy and also provide important clues for the further development of divide-and-conquer strategies for solving very large phylogenetic problems.

6.1 INTRODUCTION

Reconstructing the tree of life accurately and precisely represents the holy grail of phylogenetics and systematics. However, the impact of obtaining the tree goes well beyond these research fields to include all of the life sciences because, as it was nicely put recently by Rokas and Carroll[1], the conclusions we make as phylogeneticists form part of the assumptions underlying the analyses of the other biologists. Evolutionary information is now becoming increasingly included in fields as diverse as comparative biology, genomics and pharmaceutics. In the past decade, the increasing accumulation of phylogenetic data, made possible by the molecular revolution, has brought the dream of realising a highly comprehensive tree of life tantalisingly close.

Currently, however, the continued lack of suitable phylogenetic data represents a proximate hindrance in our efforts to reconstruct the tree of life. Although whole genomic data is becoming available at an increasing rate (but more so for prokaryotic organisms with their smaller genomes), molecular sampling has generally been sparse and restricted largely to model organisms and model genes[2,3]. However, even with the prospect of abundant whole-genome sequence data, the ultimate hindrance is the sheer size of the tree of life itself, which has been estimated to comprise anywhere from 3.6 million to 100+ million species (but most commonly 10–15 million)[4].

It has long been appreciated that the number of possible phylogenetic trees increases superexponentially with the number of taxa[5]. For example, there are three distinct rooted phylogenetic trees for three species, 15 for four species, 105 for five species, and so on. For only 67 species, the number of possible trees is on the order of 10 to the power of 111 trees, a number that just exceeds the volume of the universe in cubic Ångstroms (a comparison first heard by the first author from David Hillis). Phylogenetic analyses are now routinely conducted on data sets of this size and larger (up to hundreds of taxa). Albeit comparatively rare, analyses of thousands of taxa have also been performed, mostly as proof of concepts for new algorithmic implementations. These include a neighbour joining (NJ) analysis of nearly 8,000 sequences[6], a maximum likelihood (ML) analysis of 10,000 taxa[7], and a maximum parsimony (MP) analysis of 13,921 taxa[8,9]. However, we are unsure of the prospects of achieving a correct or nearly correct answer for studies of these size, given the literally astronomical size of ‘tree space’.

Compounding this limitation is the fact that the general problem of reconstructing a tree (or a network, given that the tree of life is not always tree-like) from a given data set is one of a set of non-deterministic polynomial time (NP) problems for which no efficient solution is known or, more pessimistically, one for which no such solution potentially exists (NP-complete)[10]. Thus, the analysis of larger data sets requires a disproportionately longer time (or disproportionately more computer resources) and/or the use of increasingly less efficient heuristic search strategies, with both factors

impacting negatively on our ability to recover the best solution for that given data set. Fortunately, several studies using empirical and/or simulated data have shown that even phylogenetic analyses at the high end of the scale currently examined are both tractable and show acceptable, if not surprising, accuracy with shorter sequence lengths than might be expected[11–13], thereby reinforcing some theoretical work in the latter area[14,15]. Additionally, advances in computer technology and architecture such as parallel and distributed computing and programs that exploit them efficiently in combination with the continual development of faster search strategies promise to make even larger phylogenetic problems increasingly tractable. However, the NP-completeness of the phylogeny problem represents a fundamental limitation in our efforts to unearth the tree of life.

As such, we face a dilemma in attempting to reconstruct the tree of life (or even major portions thereof). Smaller problems are computationally easier to solve, but at the extreme, have been demonstrated to be susceptible to the adverse effects of taxon sampling and, for parsimony in particular, long branch attraction[16] (for a review of the latter, see Bergsten[17]). In these cases, the fact that DNA has only four character states can lead to a high number of convergent changes (noise) along two long branches leading to unrelated taxa. These convergent changes can pull the two branches together, thereby leading the phylogenetic analysis astray. Thus, the general consensus is that, given a suitable sampling strategy[18], the addition of species to a phylogenetic analysis is usually beneficial in terms of accuracy because it ameliorates the effects of these two problems[19,20] (see Rosenberg and Kumar[21] for a contrary view). At some point, however, the computational complexity of the phylogeny problem must begin to outweigh the benefits of adding taxa. Although it is not stated explicitly in the literature, it seems that the general expectation is that phylogenetic accuracy shows a convex distribution with respect to the number of taxa in the analysis, with taxon sampling and computational complexity limiting accuracy when species numbers are low and high, respectively.

It remains to be demonstrated whether or not this expectation is true and, if so, at what point accuracy is maximised, while simultaneously considering the running time of the analysis. Establishing the latter could be especially important to the further development of the so-called 'divide-and-conquer' search strategies such as quartet puzzling[22] and disk-covering[9,23,24]. These strategies generally seek to solve large phylogenetic problems by breaking them down into numerous smaller subproblems that are computationally easier to solve precisely because they are smaller with respect to both the number of taxa and the evolutionary distance between those taxa. The results from the subproblems are then combined to provide an answer for the initial, global problem. As such, divide-and-conquer strategies essentially attempt to bridge the gap between the problems of taxon sampling and computational complexity. However, it is unknown what the optimal sizes of the subproblems should be in order to achieve the greatest accuracy in the shortest time possible. To date, subproblem sizes have usually been determined empirically on a case-by-case basis.

Thus, the goal of this chapter is to extend on previous analyses examining the scalability of phylogenetic accuracy with respect to the number of species in the analysis (the 'size' of the analysis). Specifically, we use simulation to investigate the changes in various parameters (accuracy, resolution and running time) related to the analysis of increasingly larger phylogenetic problems under different optimisation criteria (NJ, MP and ML) and methods of data set selection (random or clade sampling). Our results elucidate the prospects for phylogenetic analyses of very large phylogenetic problems, as might be needed to infer the tree of life or study large and species rich taxa, and provide additional insights into the potential of divide-and-conquer search strategies within this context.

6.2 MATERIALS AND METHODS

6.2.1 Simulation Protocol

The simulation protocol used was modelled on that followed by Bininda-Emonds et al.[13] to examine the scaling of accuracy in very large phylogenetic trees. For each run, a model tree of 4,096 taxa was generated according to a stochastic Yule birth process using the default parameters of the

YULE_C procedure in the program r8s v1.60[25]. Branch lengths on the tree were modelled assuming a model of substitution that departs from a molecular clock. Specifically, 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 with branch durations obtained from the Yule process model.

A model data set was then created by evolving a nucleotide sequence down the model tree using a standard Markov process model as implemented in Seq-Gen v.1.2.7[26]. The sequence length was 2,000 bp, which is of sufficient length for simulated data with its stronger signal to achieve good accuracy for even the largest tree examined herein[13], but is also short enough to keep running times within acceptable limits. Sequences were generated under a Kimura 2-parameter model[27] with a transition/transversion ratio (ti:tv) of 2.0, site-to-site rate heterogeneity (that is, Gamma model) with shape parameter of 0.5, and an overall average rate of evolution of 0.1 substitutions/site, measured along a path from the root to a tip of the tree. No invariant sites were explicitly modelled.

The model data set was then sampled to create test data sets where the number of taxa varied on a $\log_2$ scale from 4 to 2,048. No sampling of characters was performed so that the sequence length was always 2,000 bp. Taxon sampling was accomplished by either selecting taxa at random (random sampling) or by selecting a single clade from the model tree of the same size as the number of taxa to be retained (clade sampling); all other taxa were pruned from the test data sets. The expectation is that clade sampling should result in improved accuracy, given that it minimises the evolutionary diameter of the problem; this is the logic underlying the disk-covering family of divide-and-conquer methods[28]. By contrast, random sampling will tend to result in an increased number of long branches and/or extend the diameter of the problem, especially when the proportion of taxa sampled is very low. Both factors have been demonstrated to reduce the accuracy of phylogenetic inference.

Clade sampling requires the model tree to possess at least one clade for all the test sizes. Because this situation was difficult to achieve, clades that were within ±2.5% of the desired size were used when there was no clade of exactly the size desired. When multiple clades for a given size existed, one was chosen at random. If the model tree did not contain clades of all the desired sizes, it was discarded, and a new model tree was generated.

Each subsampled data set (for both random and clade sampling) as well as the full data set were analysed using three optimisation criteria, each of which accounted for the model of evolution Kimura 2-parameter + Gamma (K2P + G) as far as possible: MP, NJ, and ML. For the four largest matrices (512; 1,024; 2,048; and 4,096 taxa), a ML analysis in conjunction with a disk-covering divide-and-conquer framework (ML-DCM3) was also used. Bayesian analysis was not examined due to time and memory constraints[29]. Because Bayesian analysis samples from the posterior distribution of trees, it is necessarily significantly slower than the other methods examined here, especially if a high number of generations is employed to ensure reliable results. Even without Bayesian analysis, each replicate required just over five days to complete.

Thus, the results for each individual run were based on data matrices all derived from the same model set of molecular data evolved along the same model tree. This procedure differs substantially from that used by Bininda-Emonds et al.[13], in which model trees of the desired problem size were generated (that is, there was no sampling performed). Additionally, for each subproblem size and sampling strategy, the same alignment was analysed by each of MP, NJ, ML and where appropriate ML-DCM3. In total, 50 runs were conducted, comprising nearly eight CPU months of analysis time.

6.2.2 Phylogenetic Analysis

MP analyses used PAUP* v4.0b10[30] with transversions weighted twice as much as transitions. Different search strategies were employed depending on the size of the alignment. Below 16 taxa, a branch-and-bound search was used, thereby guaranteeing that all optimal trees were found. For matrices with ≥16 taxa, various heuristic searches were used depending on the size of the problem:

a thorough heuristic (<256 taxa), the parsimony ratchet (<1,024 taxa[31]), and finally a greedy heuristic (≥4,096 taxa). The thorough heuristic consisted of 100 random addition sequences with TBR branch-swapping, with a maximum of 10,000 trees being retained at any time during the analysis. The parsimony ratchet consisted of 10 batches of 100 iterative weighting steps, with 25% of the characters receiving a weight of two at each step. Thereafter, all equally most parsimonious trees were used as starting trees for a heuristic search using TBR branch swapping and limited to one hour of CPU time. Each replicate used the same command file for the ratchet, which was created using the Perl script PerlRat v1.0.9a. However, for the largest matrices, even the parsimony ratchet proved to be too slow during the test phase, especially because of the use of a step matrix to account for the ti:tv ratio. Therefore, a greedy heuristic was used, consisting of a single simple stepwise addition sequence followed by TBR branch-swapping with a maximum of 10 trees being retained at any time.

NJ analyses used QuickTree[32] using a Kimura translation to determine the pairwise distances.

ML analyses used RAxML-V (Randomized Axelerated Maximum Likelihood)[33], which is one of the fastest and most accurate programs for ML-based phylogenetic inference. A key feature of RAxML is its comparatively low memory consumption[29], which in combination with its advanced search algorithms and accelerated likelihood function[33,34] makes it uniquely suitable for ML analyses of large numbers of taxa. All RAxML analyses used the default hill-climbing search option (–f c) using an HKY85 substitution model with an estimate of 50 distinct per site evolutionary rate categories (CAT). This HKY + CAT model is essentially empirically equivalent to the better known HKY + I + G model, but requires fewer floating point operations and memory.

Finally, we also performed ML analyses using a divide-and-conquer search algorithm at the largest problem sizes (512 or more taxa) using RAxML in concert with the Recursive Iterative Disk Covering Method (Rec-I-DCM3)[9]. This combination of methods has been more formally referred to as Rec-I-DCM3(RAxML); however, we use the simpler ML-DCM3 throughout this chapter. Based on an initial 'guide tree' containing all taxa (here, the starting tree for the ML analyses as computed by RAxML), Rec-I-DCM3 intelligently decomposes the data set into smaller subproblems that overlap in their taxon sets. These subproblems are then solved using RAxML (using the same parameters as above), with the respective subtrees merged into a comprehensive tree with the Strict Consensus Merger[23]. This global tree was then further improved using RAxML (using the fast hill climbing heuristic; option –f f) to construct the new guide tree. The processes of decomposition, subproblem inference, subtree merging and global refinement were repeated for three iterations. The maximum size of the subproblems was 25% of the size of the full data set, as suggested in the user notes to Rec-I-DCM3.

A time limit of 12 hours was imposed on each individual analysis. This limit was never invoked for the NJ analyses and only for the largest matrices for MP (4,096 only), ML (2,048 and 4,096, but not always for both sizes) and ML-DCM3 (2,048 and 4,096). The use of a time limit will obviously impact accuracy negatively and potentially penalise the more computationally intensive ML analyses to a greater extent. However, the reality is that shortcuts of various types (for example, time limits or less thorough search strategies) must be employed when analysing very large matrices, so this constraint might represent a reasonable one. To judge the effects of imposing a time limit, one additional run was performed for the full data set of 4,096 taxa with all methods being allowed to run to completion.

In all cases, the inferred tree was held to be the strict consensus of all equally optimal solutions. All analyses were conducted on a cluster of unloaded 2.4-GHz Opteron 850 processors, each with 8 GB of RAM, located at the Department of Informatics at the Technical University of Munich. All programs used (including those used to simulate the data) were compiled as needed for this platform.

6.2.3 Variables Examined

Results were analysed with respect to three variables that are particularly relevant to the phylogenetic analysis of very large data sets: resolution, accuracy and running time. Resolution is the

number of clades on the inferred tree relative to the total number of clades on a fully bifurcating tree of the same size ($n - 2$ for an unrooted tree, where n = number of taxa). Resolution varies between 0 and 1, with the former value indicating a completely unresolved bush and the latter indicating a fully resolved tree. This parameter reflects the decisiveness of the analysis and is most relevant for the MP analyses. NJ always returns a single, fully resolved tree, and ML analyses invariably do so as well.

Accuracy was measured as the ability to reconstruct the model tree. In computer science, the optimality score of an analysis (either in isolation or in relation to that of the model tree) is also often used as a proxy for accuracy. However, the use of three different optimality criteria in this study prevents such an approach, and the comparison to a known 'true' tree is perhaps more intuitive to biologists. Accuracy was quantified using both the consensus fork index (CFI[35,36]) of the strict consensus of the inferred and model trees and the symmetric difference (or partition metric) between these two trees (d_S[37]). The CFI indicates the proportion of clades shared between the two trees, whereas d_S indicates the number of clades found on one tree or the other, but not both. To make these values comparable, d_S was normalised according to the number of taxa on the trees (by dividing by $2n - 6$, where n = number of taxa[38]) and subtracted from one to derive a similarity measure equivalent to CFI. Although it is not strictly accurate, we continue to refer to this metric as d_S for convenience.

CFI and d_S differ most importantly in how they treat polytomies in the inferred tree (the model tree is always fully bifurcating). CFI treats all polytomies as errors, whereas d_S essentially ignores them because they do not specify any unique clades. Thus, in comparing a fully resolved tree with a fully unresolved one, CFI = 0 and $d_S = 0.5$. As such, the difference between CFI and d_S is again most relevant for the MP analyses, which are the only ones expected to produce trees that are not fully resolved. For the comparison of two fully resolved trees, CFI = d_S.

Finally, the running time for each analysis was recorded in seconds. Again, an upper limit of 12 hours (43,200 seconds) was imposed on all analyses. However, analysis times could still substantially exceed this limit in some cases due to the discrete nature of the stopping mechanisms. For instance, a search can be terminated only after the completion of an iteration or calculation of an optimality score, both of which can represent long-running operations at the largest tree sizes.

For each variable, results were compared using a multivariate analysis of variance (ANOVA), with the method of analysis and sampling strategy as factors, and the size of the data set as a covariate. The level of significance was $\alpha = 0.05$. Fisher's protected least significant difference (PLSD) test was used to determine significant differences between categories within a factor.

6.2.4 Software Availability

The following software and/or source code used in this study are freely available at the following URLs:

- PerlRat.pl: www.uni-jena.de/~b6biol2/ProgramsMain.html
- RAxML: diwww.epfl.ch/~stamatak (under 'software')
- Rec-I-DCM3: www.cs.njit.edu/usman/RecIDCM3.html

6.3 RESULTS

6.3.1 Resolution

Resolution was always one for each individual NJ, ML, and ML-DCM3 analysis. MP produced trees that were significantly less resolved ($P < 0.0001$ for all pairwise comparisons) and, except for a tree size of four with random sampling, were never fully resolved on average (Figure 6.1A). Nevertheless, the MP trees were generally well resolved at all tree sizes, with the average resolution being always greater than 0.90. Resolution for MP differs significantly with tree size ($P < 0.0001$),

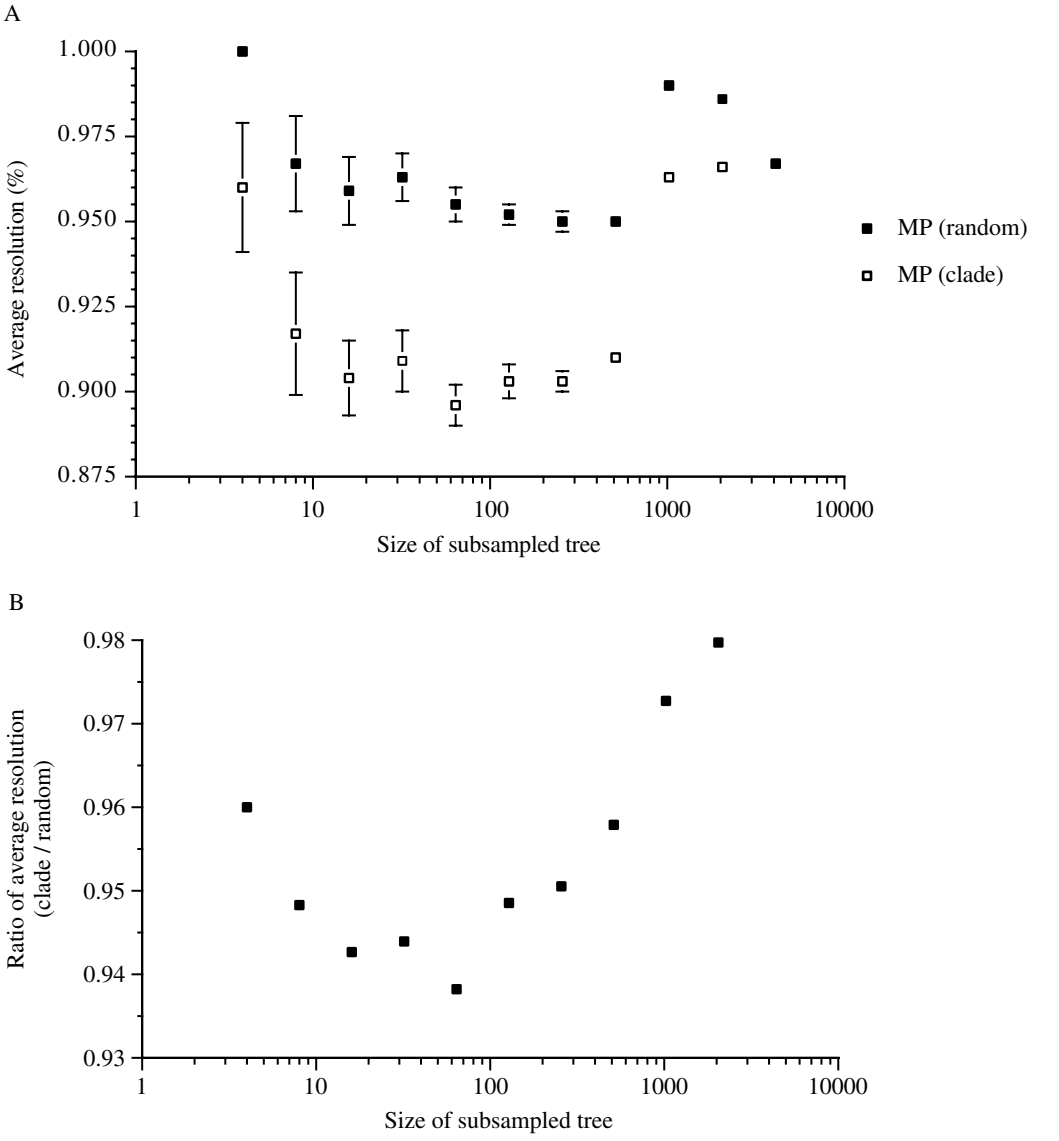

FIGURE 6.1 Resolution of trees inferred using MP from data sampled from a model matrix of 2,000 bp for 4,096 taxa. (A) Average resolution over 50 individual runs; error bars represent standard errors. (B) Ratio of average resolutions from clade sampling as compared to random sampling. Resolution for all other optimisation criteria was always 1.

showing a concave pattern that is noticeably higher at extremely small and extremely large tree sizes. In the latter case, however, this is an artefact of only 10 trees being retained in analyses of 1,024 or more taxa. Otherwise, it appears that resolution reaches a plateau of about 0.90 for clade sampling and 0.95 for random sampling. The average resolution for the MP analyses using clade sampling was always significantly less than that for random sampling ($P < 0.0001$); the ratio of the values for clade versus random sampling fell between 0.94 and 0.98 at all tree sizes (Figure 6.1B). All methods yielded fully resolved trees, or nearly so for MP, in the time-unlimited analyses (Table 6.1).

TABLE 6.1
Statistics Relating to a Time-Unlimited Analysis of the Full Dataset of 4,096 Taxa

Optimisation Criterion/Method	Resolution	Accuracy		Time (seconds)
		CFI	$(1 - d_S)$	
MP	1.000	0.903	0.917	69,392
NJ	1.000	0.857	0.857	193
ML (fast hill climbing)	1.000	0.912	0.912	38,737
ML (standard hill climbing)	1.000	0.923	0.923	303,450
ML-DCM3	1.000	0.921	0.921	195,371

6.3.2 Accuracy

Accuracy, whether measured by CFI or d_S was generally good at all tree sizes and for all methods (Figure 6.2A and Figure 6.3A). In all cases, accuracy was greater than 80% on average and often better than 90%. Tree size had a variable impact on accuracy. It did not influence accuracy for either ML-DCM3 ($P = 0.4812$; although only four sizes were tested for this method), MP as measured by d_S ($P = 0.4132$), or ML ($P = 0.1995$), but had a significant effect for both NJ ($P = 0.0244$) and MP as measured by CFI ($P = 0.0087$). However, the only clear trend is for NJ under clade sampling where accuracy decreases with the size of the problem. In all the remaining cases, the curves are reasonably flat and/or sigmoidal. Except for ML-DCM3, allowing all methods to run to completion in the time-unlimited analyses produced significantly more accurate results when compared to the 12-hour limited analyses ($P < 0.0001$ according to a one sample *t*-test).

The different optimisation criteria/methods used also had an impact on the accuracy of the solutions. When CFI was used to measure accuracy (Figure 6.2A), ML and ML-DCM3 were not significantly different ($P = 0.0763$), and neither were MP and NJ ($P = 0.6982$). However, the trees derived using the former methods were significantly more accurate than those from the latter ($P < 0.0001$). When d_S was used (Figure 6.3A), ML trees were statistically indistinguishable from those from either MP ($P = 0.7037$) or ML-DCM3 ($P = 0.0618$), although the latter two were significantly different from one another ($P = 0.0340$). NJ yielded significantly worse trees in all cases ($P < 0.0001$).

Only the MP analyses showed a difference in accuracy as measured by the two metrics (compare Figure 6.2A and Figure 6.3A), with the analogous values for d_S being either equal to, or more commonly, greater than those for CFI. The effect was the most pronounced for clade sampling, which also produced solutions that were less resolved than were those from random sampling (Figure 6.2B and Figure 6.3B).

For both NJ and MP (d_S only), the sampling strategy had a significant effect on accuracy ($P < 0.0001$), with clade sampling generally leading to increasingly accurate solutions as the size of the problem decreased. However, the two sampling strategies showed similar performance with respect to accuracy for trees of 512 or more taxa. No effect was present for MP when accuracy was measured using CFI ($P = 0.1248$). Likewise, there was no significant trend for ML with respect to the sampling strategy ($P = 0.0698$). Random sampling produced slightly, but significantly more accurate trees with ML-DCM3 at the three relevant problem sizes examined for it (512; 1,024; and 2,048 taxa; $P < 0.0001$).

6.3.3 Running Time

Except for NJ, no method obtained a solution for the full model data set (4,096 taxa) within the 12-hour time limit. The running times for the unlimited MP, ML and the three iteration ML-DCM3 analyses of 4,096 taxa (see Table 6.1) were 1.6, 7.0 and 4.5 times longer than the limit of 12 hours.

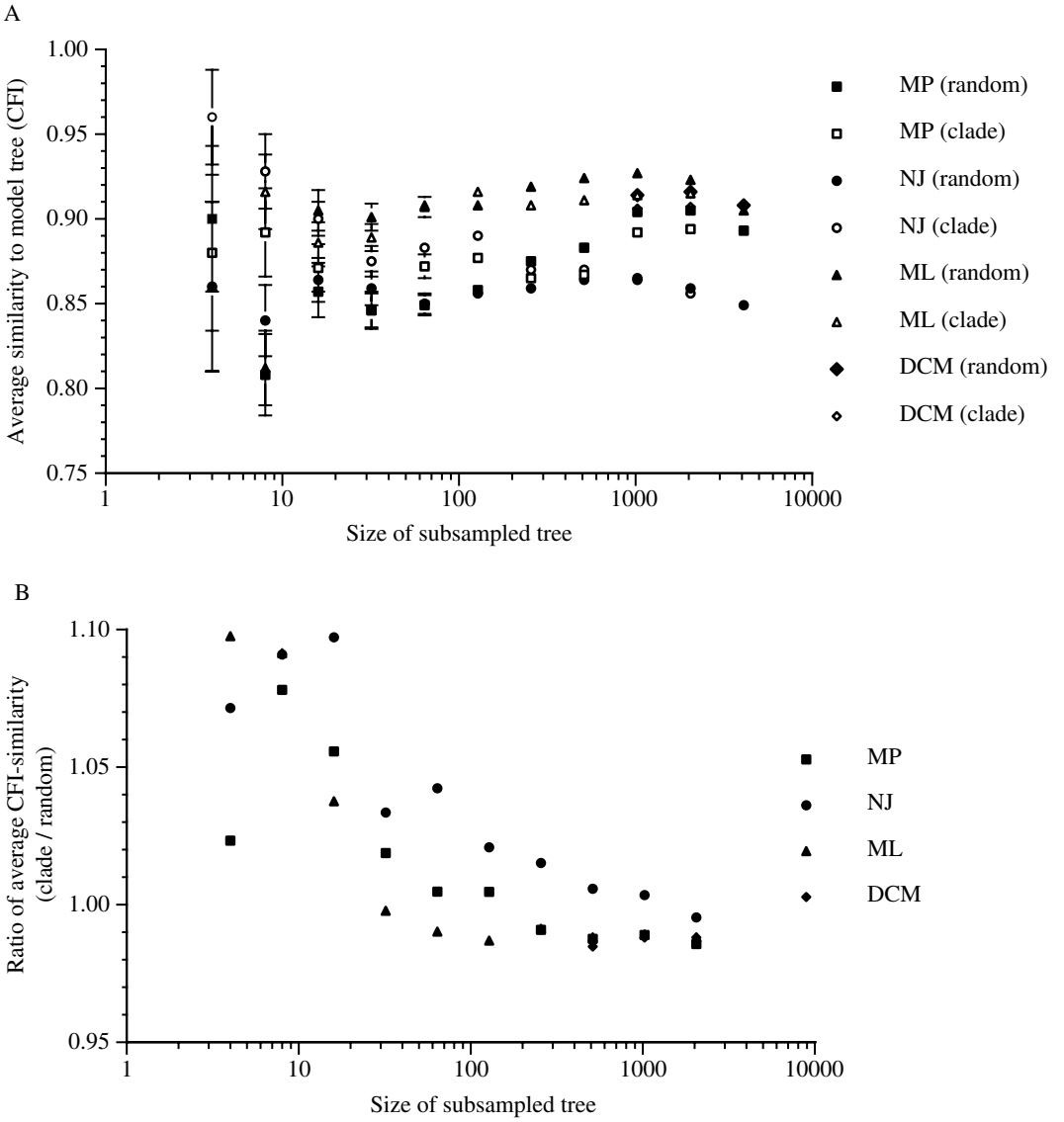

FIGURE 6.2 Phylogenetic accuracy of trees inferred using different methods from data sampled from a model matrix of 2,000 bp for 4,096 taxa. Accuracy was measured as the value of the CFI between the inferred tree and the model tree upon which the data were simulated; both trees were pruned so as to have identical taxon sets. (A) Average accuracy over 50 individual runs; error bars represent standard errors. (B) Ratio of average accuracy from clade sampling as compared to random sampling.

Running times were significantly influenced by all three factors and covariates examined, either in isolation or in combination (all $P < 0.0001$). Fisher's PLSD tests also revealed highly significant differences (all $P < 0.0001$) between all pairs of categories within the factors of sampling strategy and method of analysis.

For all optimisation criteria, running times increased approximately linearly with tree size on a log-log scale (Figure 6.4A and Figure 6.4C). For each doubling in tree size, the running time of NJ increased by a factor of about three on average (random sampling: 3.22 ± 0.60 (mean $\pm$ SE); clade sampling: 3.33 ± 0.61). MP showed both the largest and most variable increases in running

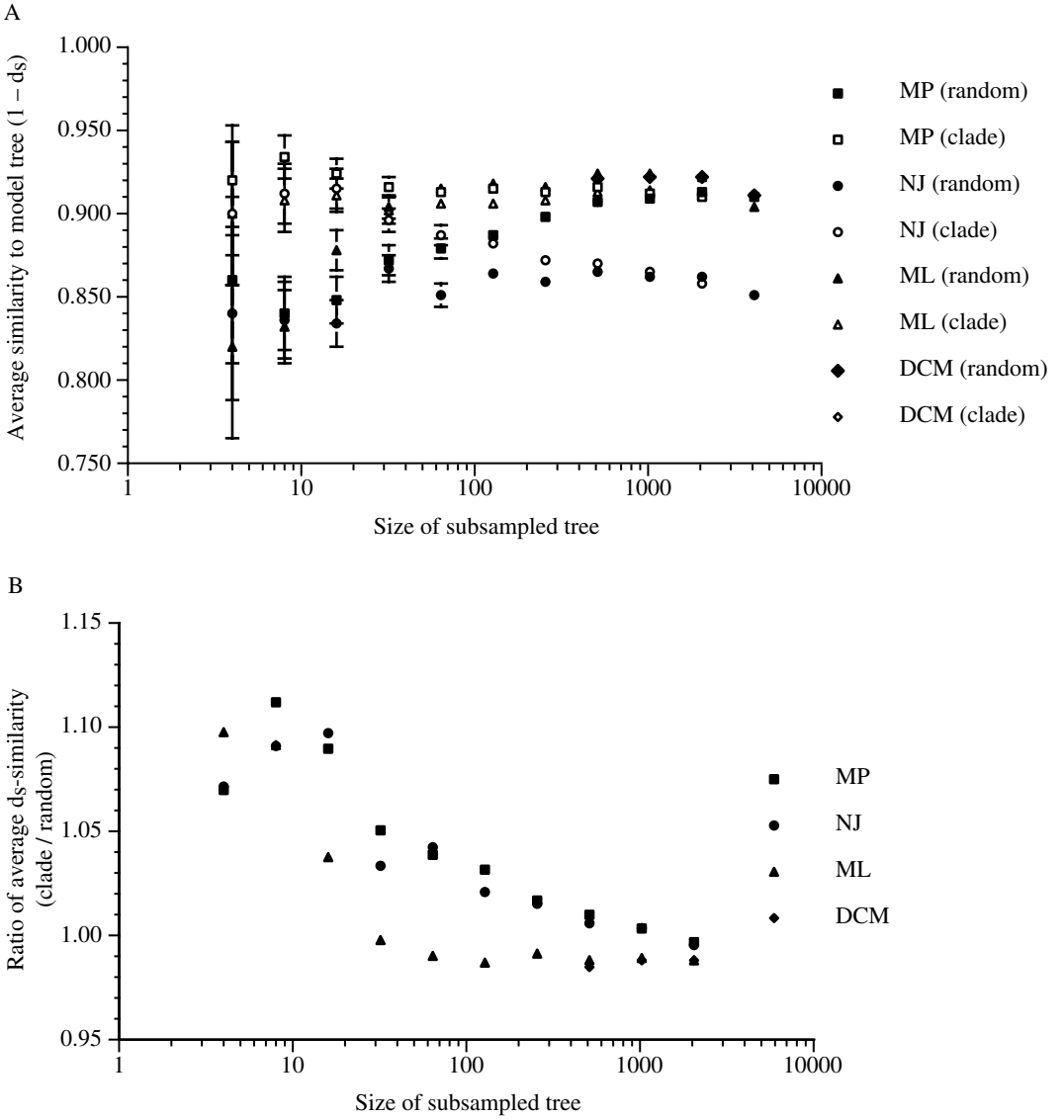

FIGURE 6.3 Phylogenetic accuracy of trees inferred using different methods from data sampled from a model matrix of 2,000 bp for 4,096 taxa. Accuracy was measured as one minus the normalised value of the partition metric between the inferred tree and the model tree upon which the data were simulated; both trees were pruned so as to have identical taxon sets. (A) Average accuracy over 50 individual runs; error bars represent standard errors. (B) Ratio of average accuracy from clade sampling as compared to random sampling.

time (random sampling: 6.54 ± 2.31; clade sampling: 12.26 ± 5.84). The largest increases for MP occurred for the comparisons 8–16 and 64–128 taxa (random sampling) and 16–32, 32–64 and 64–128 taxa (clade sampling). Many of these high rates corresponded with either the adoption of a new, less thorough search strategy or when a given search strategy was apparently becoming 'overloaded' for a given problem size. Finally, the rate increases for ML were intermediate between NJ and MP and with low variation (random sampling: 4.34 ± 0.69; clade sampling: 5.03 ± 0.87).

Compared to the other methods (Figure 6.4A), NJ always produced the shortest running times ($P < 0.0001$), with the differences becoming the most marked at tree sizes of 16 taxa or greater.

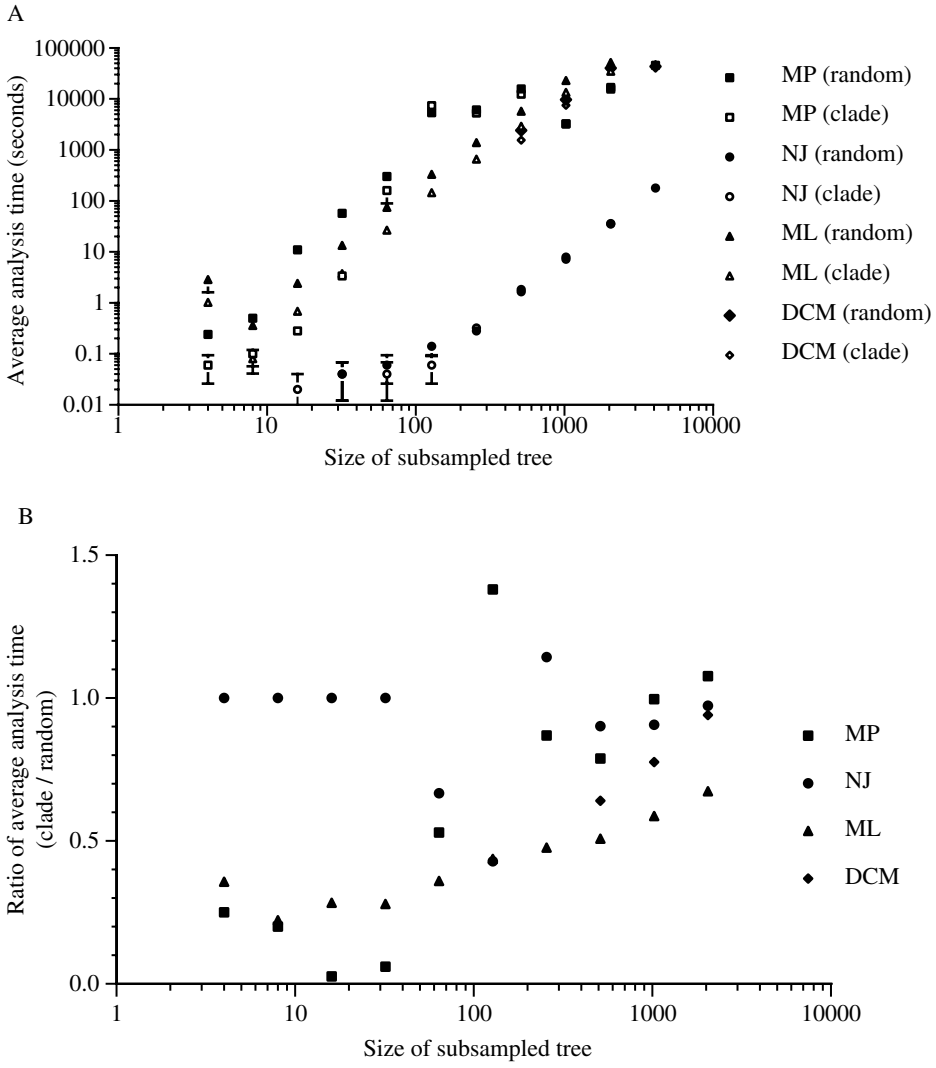

FIGURE 6.4 Analysis times for trees inferred using different methods from data sampled from a model matrix of 2,000 bp for 4,096 taxa. The time used for 4,096 taxa derives from the single time-unlimited analysis for MP, ML and ML-DCM3. (A) Average analysis time over 50 individual runs; error bars represent standard errors. (B) Ratio of average analysis time from clade sampling as compared to random sampling. (C) Ratio of average analysis time for a given sample size as compared to previous sample size.

The running times of MP and ML were roughly comparable, although MP was significantly faster ($P < 0.0001$) on the whole. MP tended to run faster for the smallest and largest tree sizes with clade sampling, whereas ML generally ran faster for all tree sizes under random sampling. In those cases where the time limit was not exceeded, ML-DCM3 was faster on average than ML, but slower than MP (in both cases by a factor of two to three).

Analysis times under clade sampling were almost always less than those for random sampling, with the differences becoming smaller with increasing tree size (Figure 6.4B). For ML, running times under clade sampling were faster by a factor of at least two for all tree sizes except four and 2,048. The most marked differences for MP were limited to tree sizes of 64 or fewer taxa, where the differences were the largest for all the methods examined (including a factor of nearly 40 with

C

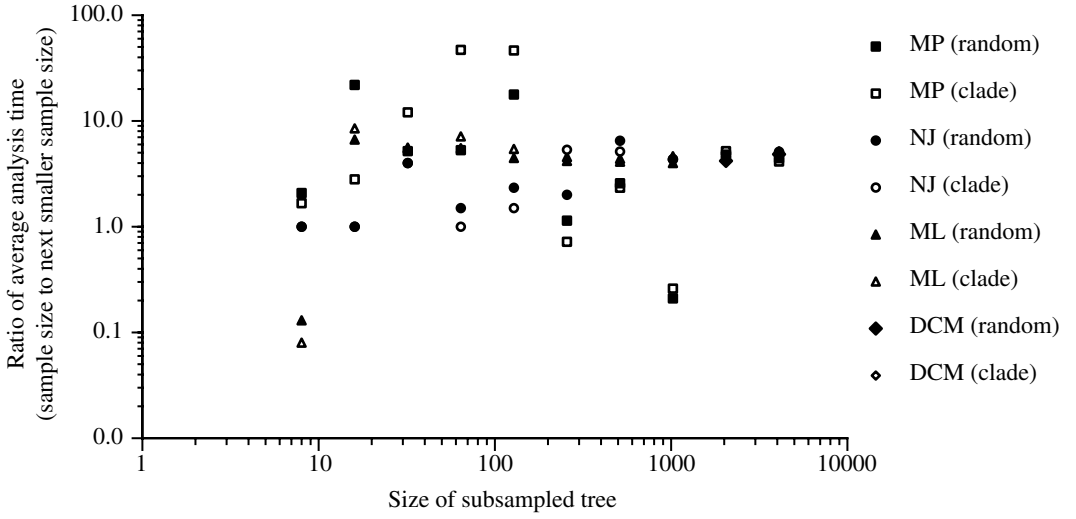

FIGURE 6.4 (Continued).

16 taxa). Beyond 128 taxa, the respective running times for the different sampling strategies were approximately equal for MP. For 256 and 512 taxa, at least, this result reflects the more deterministic nature of the parsimony ratchet, which must perform a set number of iterations.

6.4 DISCUSSION

Our simulation study produced three key findings, some unexpected:

- That accuracy is at best only weakly influenced by the size of the problem
- That the methods of inference examined produce solutions of comparable, and good, accuracy
- That the sampling strategy employed has a significant effect on both accuracy (to a point) and, more strongly, on the running time of the analysis

Naturally, caveats abound. Our study used simulated data, which tend to be 'cleaner' and contain more phylogenetic signal than real sequence data. Moreover, the model of evolution used (HKY + G) is less complicated, and therefore less computationally intensive, than those more commonly used. Finally, the methods of inference were refined to match the known model as precisely as possible. Thus, our results might, in isolation, represent a best case scenario. However, the comparative aspects of our study should be accurate.

6.4.1 Accuracy and Speed

Although it is widely accepted that larger phylogenetic problems are more difficult to solve and should therefore show decreased accuracy, there is now accumulating evidence to suggest that the performance dropoff is not as severe as many would believe, at least up to problem sizes of about 10,000 taxa[11,13]. Our finding that accuracy is essentially flat with respect to tree size (Figure 6.2A and Figure 6.2B) lends further support to these latter findings, indicating that good accuracy is achievable even for very large phylogenetic problems and within a reasonable timeframe. Moreover, our work indirectly supports recent work by Rokas, Carroll and colleagues[1,39] that strongly argues

that the amount of sequence data (number of genes), and not the number of taxa, is the more critical factor influencing phylogenetic accuracy. In our case, the alignment length of 2,000 bp was specifically chosen because it apparently contained sufficient phylogenetic signal for all problem sizes examined here[13], thereby minimising the effects of sequence length. In real terms, 2,000 bp would represent perhaps one or two genes of lengths that are typically used in phylogenetic analysis. Although this might be an insufficient number of genes to achieve good accuracy[1,39] it must again be remembered that the simulated sequence data often contain much more signal than real data due to the absence of gaps and a lack of noise that would arise because of alignment errors.

Similarly, good evidence exists that the performance of NJ with respect to accuracy, although acceptable, is generally inferior to that produced by ML and weighted MP[40]. This difference in performance, as well as the vastly shorter running times of NJ, derives from the absence of branch-swapping in NJ to correct for suboptimal topologies created during the tree construction process. As such, NJ by itself seems ideally suited as a method to very quickly generate a relatively accurate starting tree for subsequent and more computationally intensive branch-swapping. Exactly such an approach is implemented in PHYML[41] and could equally well be applied to MP or within a distance framework (minimum evolution, ME).

Our results also attest to the recent advances in heuristic search strategies, particularly in a ML framework. Despite the increased complexity of ML as compared to MP (which has long been viewed as an obstacle to ML analyses), both accuracy and running times were comparable between the two optimisation criteria. Moreover, the more complex nature of the likelihood surface means that, at most, only a few equally optimal solutions are usually found (and typically only a single solution)[42], thus providing a more resolved solution than is usually the case for the analogous MP analyses (Figure 6.1A). Many would see this as being a desirable feature in that the ML estimate of the phylogeny of a large, species-rich group could be argued to be more decisive and definitive than the MP estimate.

In addition, whereas MP running times were kept in check by applying increasingly faster heuristic search algorithms, all ML searches used the same standard hill climbing searches in RAxML. A less thorough, but faster hill climbing strategy also exists, which was used to optimise the global tree in the ML-DCM3 analyses based on previous empirical work showing it to work the best of all methods in this context. As revealed by the single analysis of the full data set, the use of this option makes the ML analyses faster (by a factor of 7.8) with virtually no loss in accuracy (see Table 6.1). In fact, the 'fast' ML analyses were both faster and more accurate than the MP analyses performed. However, it should also be realised that faster implementations of MP searches than those used here, such as those implemented in TNT (available from http://www.zmuc.dk/public/phylogeny/TNT)[43] also exist. Rec-I-DCM3 has also been used in conjunction with TNT[9], boosting performance even further. A faster implementation of NJ than that used here was also recently published[6]. At the same time, it should be pointed out that the computational 'arms race' is still ongoing, with the latest version of RAxML, RAxML-VI-HPC (v2.1), showing significant speed improvements over the version used here, particularly for very large data sets.

Altogether, these findings bode well, not only for reconstructing very large phylogenies, but also for estimating support for the groups present in those phylogenies[44]. Analogous to our finding that accuracy is relatively flat with respect to the size of the problem and, in the case of MP, to the use of greedier heuristic search strategies, Salamin et al.[44] found that estimated bootstrap frequencies are apparently robust to the use of less effective branch-swapping methods during the tree searching operations (for example, in decreasing order of searching thoroughness, TBR versus SPR versus NNI; see Swofford et al.[27] for descriptions). Again, the use of NJ offers a means to quickly generate a reasonably accurate starting tree for further branch-swapping operations. Finally, an additional option for quickly determining support values in a ML framework is the use of the resampling of estimated log-likelihood (RELL) approximation[45,46], which apparently can estimate bootstrap proportions for a given tree more accurately than a true bootstrap analysis that uses fast heuristics to search through tree space[47].

6.4.2 The Importance of Sampling Strategy

Instead of sample size, it was the form of the sampling strategy that had the greatest impact on both phylogenetic accuracy (Figure 6.2A and Figure 6.2B) and, perhaps more unexpectedly, the running time of the analyses (Figure 6.4A and Figure 6.4B). Given that we will always be working with samples from the entire tree of life, the sampling strategy used, therefore, becomes a crucial consideration in phylogenetic systematics.

The influence of sampling strategy on accuracy has long been discussed, but more in the context of sampling density and the diameter of the problem. The former variable relates to how many taxa for a given group are included in the analysis, whereas the latter roughly corresponds to how much evolutionary history the sampled group contains. An especially stimulating paper was that of Kim[48], which showed that adding taxa to an analysis usually decreased phylogenetic accuracy. However, the protocol used by Kim added taxa outside the reference group, thereby expanding the phylogenetic diameter of the problem and decreasing the overall sampling density. Subsequent studies designed to address Kim's findings instead added the taxa within the reference group. As such, the phylogenetic diameter of the problem was unchanged, and the sampling density was increased. These studies instead demonstrated the general benefit of adding taxa to the analysis (see [18] and references therein). Moreover, taxa that were added specifically to break up any long branches (and therefore making the greatest increase in local density) were shown to improve accuracy to the greatest extent. Thus, a general, long-standing recommendation for phylogenetic analyses is to add taxa in such a way as to best represent the overall diversity of the group and/or to potentially break up any long branches[18].

The two sampling strategies used herein likewise can be viewed with respect to sampling density and evolutionary diameter. For a given problem size, clade sampling always yields the maximum density (there are no unsampled taxa for that group) and minimises the diameter. By comparison, random sampling will usually yield samples of greater diameter and less density (and therefore contain more long branches) for the same problem size, especially when the sample size is only a small percentage of the overall problem size. As the sampling size increases, the difference between clade and random sampling decreases, with the two strategies obviously converging when the sampling size equals the overall problem size. This explains why the performance differences between the strategies decreased as the sampling size increased.

More important, however, was the generally beneficial effect of clade sampling on running time (Figure 6.4A and Figure 6.4B), an effect that was only absent at the largest subproblem sizes and for ratchet-based MP searches (which will have a more rigid running time due to their more deterministic nature). To our knowledge, although this general finding might have been suggested informally, our results are the first to document it. The speed advantages to performing an analysis in a divide-and-conquer framework have usually been ascribed to the smaller problem sizes, with clade sampling being held to improve accuracy[28].

As such, our findings in this regard provide another important reason (in addition to increased accuracy) to ensure that the sampling is as complete as possible for a phylogenetic problem of a given size. Moreover, this recommendation applies equally to conventional phylogenetic analyses and to those performed in a divide-and-conquer framework. With respect to the latter, methods such as the disk-covering family[9,23,24] or IQPNNI[49] that intelligently subsample the data matrix are therefore to be preferred for reasons of both speed and accuracy over those such as classic quartet puzzling[22] that employ random sampling.

That said, technical considerations can occasionally override this general recommendation. A primary example here includes the parallel implementation of a divide-and-conquer algorithm for large scale phylogeny reconstruction, where the selected strategy has important practical implications. Recent work on a parallel version of Rec-I-DCM3(RAxML) revealed significant problems of processor load imbalance due to the great variation in subproblem sizes yielded by the intelligent decomposition method of Rec-I-DCM3[50]. However, such load imbalance problems could be

resolved based on our findings that the sampling strategy has a decreasing effect on accuracy and inference times for proportionately larger sampling sizes, and that the choice of sampling strategy does not significantly affect the accuracy of the ML analyses, which are known to be more immune to the adverse effects of taxon sampling and long-branch attraction. As such, it should be possible in a parallel implementation of Rec-I-DCM3(RAxML) to initially split the alignment naïvely into relatively large subsets of approximately equal size (comprising approximately 12.5–50% of the original dataset) based on the guide tree. This strategy should improve load balance without any undue loss of performance. In turn, these large initial subproblems would then be optimised using the more intelligent subdivision method employed by Rec-I-DCM3, where the benefits of clade sampling take on greater importance. The potential utility of a similar naïve division method has been observed with a proprietary divide-and-conquer algorithm implemented in RAxML (Stamatakis unpublished data).

6.4.3 Implications for the Divide-and-Conquer Framework

Although the ML-DCM3 strategy employed here showed slightly less accuracy (~1–2% less) than a ML search using RAxML alone, it showed tremendous savings in terms of running time, running faster by a factor of about 1.5 or greater on average. Admittedly, the fast ML heuristic was even faster than the ML-DCM3 strategy for the time-unlimited analyses of 4,096 taxa; however, the latter could be easily adapted to use the fast heuristic throughout and so also benefit from the speed improvement. Similar, if not even greater, performance gains with respect to both running time and especially accuracy for a MP analysis performed within a divide-and-conquer framework have also been reported[9,28].

Thus, it seems clear that a divide-and-conquer based strategy will form a key component for studying very large phylogenetic problems. In this sense, it is instructive to compare the cumulative running times for multiple analyses of a given subproblem size such that the total number of taxa examined equals the global problem size of 4,096 taxa (Figure 6.5). For instance, for a subproblem

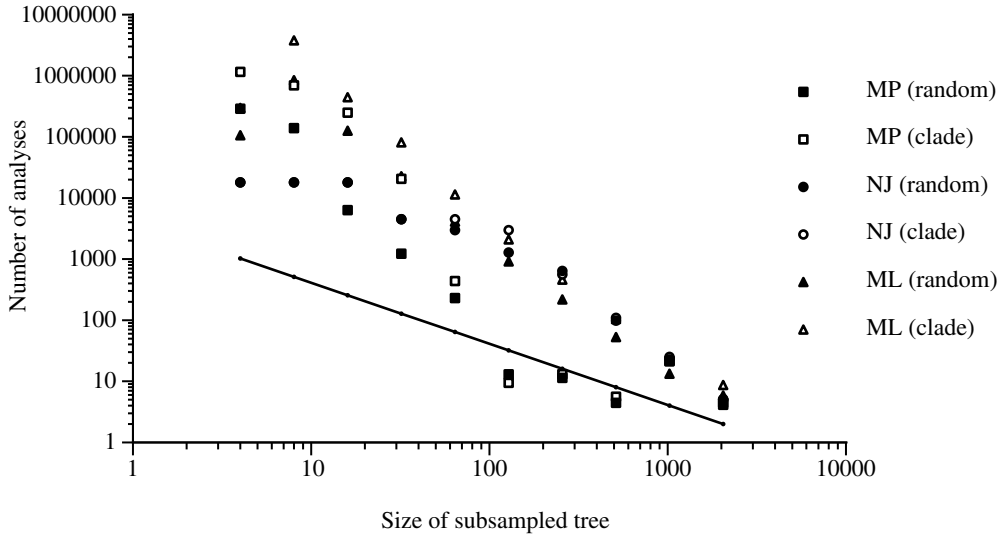

FIGURE 6.5 Number of analyses for a given tree size that could be completed in the same time needed for an analysis of 4,096 taxa. For each tree size, the average running time over the 50 runs (see Figure 6.4) was compared against the time required to analyse 4,096 taxa from the time-unlimited analyses (see Table 6.1). The black line represents the number of analyses required such that 4,096 taxa are analysed in total.

size of 32 taxa, this amounts to 128 individual analyses (although a global tree of all 4,096 taxa could not be derived from these analyses because the trees do not overlap). However, Figure 6.5 reveals that over 20,000 MP analyses of 32 taxa selected using clade sampling could be conducted in the time taken for a single MP analysis of 4,096 taxa, or 176 sets of 128 analyses. Note also that these particular numbers are underestimates, given that the MP search strategy used for 4,096 taxa was considerably less robust, and therefore comparatively faster, than that used for 32 taxa. For all optimisation criteria (including NJ) and at virtually all subproblem sizes, the time savings are similarly enormous; only MP analyses at tree sizes of 128 to 512 taxa show a decrease in time (Figure 6.5). Thus, there is tremendous scope in a divide-and-conquer framework for many individual analyses to ensure high overlap between the trees, a factor that has been shown to improve the accuracy of the merged supertree[51].

Although it is tempting to try and derive the optimal subproblem size from Figure 6.5 under the assumption that accuracy is approximately flat with respect to the size of the subproblem, it must be remembered that these numbers do not account for the initial accuracy of the merged solution and, therefore, the time required for any global optimisations of it. Because such optimisations are computationally expensive (which accounts for the proportionately longer analysis times of larger solutions), they represent an important performance bottleneck. For example, the global optimisation step, even with the use of the fast ML algorithm, consumed the most execution time for the ML-DCM3 analyses. Moreover, there is a general consensus among researchers involved in the development of divide-and-conquer algorithms that global optimisations must be applied at some point to obtain the most accurate trees possible[52]. As such, the role for divide-and-conquer strategies will be, as for NJ, to yield as good a starting tree in as little time as possible. Research should now focus, therefore, on determining the optimal subproblem size and merger method that maximise the accuracy of the merged tree (so as to minimise the global optimisation time) in as short a time as possible. To our knowledge, there has been little work in this area (although the Rec-I-DCM3 user guide suggests a maximum subproblem size of 25% of the global size), nor in examining the accuracy of the merged tree without any subsequent global optimisation. Additional benefits would derive from pursuing this course of action with an eye toward the development of efficient parallel optimisation methods, particularly for the computationally intensive global optimisation step.

6.5 CONCLUSIONS

Together with other similar findings[11–13], the results we present here are encouraging for the prospects of building ever larger phylogenetic trees in our efforts to reconstruct the tree of life. Continued developments in computer technology and algorithm development can only increase our feeling of optimism. Even so, it must be remembered that even 10,000 taxa, the approximate limit for all simulations performed to date, represent only a minute fraction of the entire tree of life. Larger problems have been analysed successfully, but without any real knowledge of how accurate the answers might be. We simply do not know at this point how far the scalability of acceptable accuracy extends. As such, it seems clear that a divide-and-conquer approach, whereby we can break the problem down into pieces where we are confident of achieving good accuracy (and in less time), must form a necessary part of our efforts to obtain the tree of life.

ACKNOWLEDGEMENTS

We thank Trevor Hodkinson and John Parnell for the invitation to contribute to this volume. We are also grateful to the Department of Informatics at the Technical University of Munich for providing access to their Infiniband computer cluster and to Usman Roshan for allowing us to

include the (at the time) unpublished Rec-I-DCM3(RAxML) procedure in our simulations. This work was funded as part of the NGFN-funded project Bioinformatics for the Functional Analysis of Mammalian Genomes (BFAM) (Olaf Bininda-Emonds).

REFERENCES

1. Rokas, A. and Carroll, S.B., More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy, *Mol. Biol. Evol.,* 22, 1337, 2005.
2. Bininda-Emonds, O.R.P., Supertree construction in the genomic age, in *Molecular Evolution: Producing the Biochemical Data, part B,* Zimmer, E.A. and Roalson, E., Eds., *Methods in Enzymology,* Vol. 395, Elsevier, Amsterdam, 2005, 745.
3. Sanderson, M.J. and Driskell, A.C., The challenge of constructing large phylogenetic trees, *Trends Pl. Sci.,* 8, 374, 2003.
4. Wilson, E.O., Taxonomy as a fundamental discipline, *Philos. Trans. R. Soc. Lond. B,* 359, 739, 2004.
5. Felsenstein, J., The number of evolutionary trees, *Syst. Zool.,* 27, 27, 1978.
6. Mailund, T. and Pedersen, C.N., QuickJoin: fast neighbour-joining tree reconstruction, *Bioinformatics,* 20, 3261, 2004.
7. Stamatakis, A., Parallel inference of a 10,000-taxon phylogeny with maximum likelihood, in *Proceedings of 10th International Euro-Par Conference (Euro-Par 2004),* Springer Verlag, 2004, 997.
8. Coarfa, C. et al., PRec-I-DCM3: a parallel framework for fast and accurate large scale phylogeny reconstruction, in *The First IEEE Workshop on High Performance Computing in Medicine and Biology (HiPCoMP 2005),* Fukuoka, Japan, 2005.
9. Roshan, U. et al., Rec-I-DCM3: a fast algorithmic technique for reconstructing large phylogenetic trees, in *Proceedings of the IEEE Computational Systems Bioinformatics conference (CSB),* IEEE Computer Society Press, Stanford, California, 2004.
10. Garey, M.R. and Johnson, D.S., *Computers and Intractability: a Guide to the Theory of NP-Completeness,* W.H. Freeman, San Francisco, 1979.
11. Salamin, N., Hodkinson, T.R., and Savolainen, V., Towards building the tree of life: a simulation study for all angiosperm taxa, *Syst. Biol.,* 54, 183, 2005.
12. Hillis, D.M., Inferring complex phylogenies, *Nature,* 383, 130, 1996.
13. Bininda-Emonds, O.R.P. et al., Scaling of accuracy in extremely large phylogenetic trees, in *Pacific Symposium on Biocomputing 2001,* Altman, R.B. et al., Eds., World Scientific Publishing Company, River Edge, NJ, 2000, 547.
14. Erdös, P.L. et al., A few logs suffice to build (almost) all trees (I), *Random Struc. Alg.,* 14, 153, 1999.
15. Erdös, P.L. et al., A few logs suffice to build (almost) all trees: part II, *Theoret. Comput. Sci.,* 221, 77, 1999.
16. Huelsenbeck, J.P. and Hillis, D.M., Success of phylogenetic methods in the four-taxon case, *Syst. Biol.,* 42, 247, 1993.
17. Bergsten, J., A review of long-branch attraction, *Cladistics,* 21, 163, 2005.
18. Hillis, D.M., Taxonomic sampling, phylogenetic accuracy, and investigator bias, *Syst. Biol.,* 47, 3, 1998.
19. Pollock, D.D. et al., Increased taxon sampling is advantageous for phylogenetic inference, *Syst. Biol.,* 51, 664, 2002.
20. Graybeal, A., Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.,* 47, 9, 1998.
21. Rosenberg, M.S. and Kumar, S., Incomplete taxon sampling is not a problem for phylogenetic inference, *Proc. Natl. Acad. Sci. USA,* 98, 10751, 2001.
22. Strimmer, K. and von Haeseler, A., Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies, *Mol. Biol. Evol.,* 13, 964, 1996.
23. Huson, D.H., Nettles, S.M., and Warnow, T.J., Disk-covering, a fast-converging method for phylogenetic tree reconstruction, *J. Comput. Biol.,* 6, 369, 1999.
24. Huson, D.H., Vawter, L., and Warnow, T.J., Solving large scale phylogenetic problems using DCM2, in *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology,* Lengauer, T. et al., Eds., AAAI Press, Menlo Park, California, 1999, 118.

25. Sanderson, M.J., r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock, *Bioinformatics,* 19, 301, 2003.
26. Rambaut, A. and Grassly, N.C., Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees, *Comput. Appl. Biosci.,* 13, 235, 1997.
27. Swofford, D.L. et al., Phylogenetic inference, in *Molecular Systematics,* Hillis, D.M., Moritz, C., and Mable, B.K., Eds., Sinauer Associates, Sunderland, Massachusetts, 1996, 407.
28. Roshan, U. et al., Performance of supertree methods on various data set decompositions, in *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life,* Bininda-Emonds, O.R.P., Ed., Kluwer Academic, Dordrecht, Netherlands, 2004, 301.
29. Stamatakis, A., Ludwig, T., and Meier, H., New, fast and accurate heuristics for inference of large phylogenetic trees, in *Proceedings of 18th IEEE/ACM International Parallel and Distributed Processing Symposium (IPDPS2004), High Performance Computational Biology Workshop,* IEEE Computer Society, Santa Fe, New Mexico, 2004.
30. Swofford, D.L., *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4,* Sinauer Associates, Sunderland, MA, 2002.
31. Nixon, K.C., The parsimony ratchet, a new method for rapid parsimony analysis, *Cladistics,* 15, 407, 1999.
32. Howe, K., Bateman, A., and Durbin, R., QuickTree: building huge neighbour-joining trees of protein sequences, *Bioinformatics,* 18, 1546, 2002.
33. Stamatakis, A., Ludwig, T., and Meier, H., RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees, *Bioinformatics,* 21, 456, 2005.
34. Stamatakis, A.P. et al., AxML: a fast program for sequential and parallel phylogenetic tree calculations based on the maximum likelihood method, *Proc. IEEE Comput. Soc. Bioinform. Conf.,* 1, 21, 2002.
35. Colless, D.H., Congruence between morphometric and allozyme data for *Menidia* species: a reappraisal, *Syst. Zool.,* 29, 288, 1980.
36. Colless, D.H., Predictivity and stability in classifications: some comments on recent studies, *Syst. Zool.,* 30, 325, 1981.
37. Robinson, D.F. and Foulds, L.R., Comparison of phylogenetic trees, *Math. Biosci.,* 53, 131, 1981.
38. Steel, M., The complexity of reconstructing trees from qualitative characters and subtrees, *J. Classif.,* 9, 91, 1992.
39. Rokas, A. et al., Genome-scale approaches to resolving incongruence in molecular phylogenies, *Nature,* 425, 798, 2003.
40. Hillis, D.M., Huelsenbeck, J.P., and Cunningham, C.W., Application and accuracy of molecular phylogenies, *Science,* 264, 671, 1994.
41. Guindon, S. and Gascuel, O., A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, *Syst. Biol.,* 52, 696, 2003.
42. Rogers, J.S. and Swofford, D.L., Multiple local maxima for likelihoods of phylogenetic trees: a simulation study, *Mol. Biol. Evol.,* 16, 1079, 1999.
43. Goloboff, P.A., Analyzing large data sets in reasonable times: solutions for composite optima, *Cladistics,* 15, 415, 1999.
44. Salamin, N. et al., Assessing internal support with large phylogenetic DNA matrices, *Mol. Phylogenet. Evol.,* 27, 528, 2003.
45. Hasegawa, M. and Kishino, H., Accuracies of the simple methods for estimating the bootstrap probability of a maximum-likelihood tree, *Mol. Biol. Evol.,* 11, 142, 1994.
46. Kishino, H., Miyata, T., and Hasegawa, M., Maximum likelihood inference of protein phylogeny and the origin of chloroplasts, *J. Mol. Evol.,* 31, 151, 1990.
47. Waddell, P.J., Kishino, H., and Ota, R., Very fast algorithms for evaluating the stability of ML and Bayesian phylogenetic trees from sequence data, in *Genome Informatics 2002,* Lathrop, R. et al., Eds., Universal Academy Press, Tokyo, 2002, 82.
48. Kim, J., General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa, *Syst. Biol.,* 45, 363, 1996.
49. le Vinh, S. and Von Haeseler, A., IQPNNI: moving fast through tree space and stopping in time, *Mol. Biol. Evol.,* 21, 1565, 2004.

50. Du, Z. et al., Parallel divide-and-conquer phylogeny reconstruction by maximum likelihood, in *High Performance Computing and Communications: First International Conference, HPCC 2005, Sorrento, Italy, September, 21–23, 2005, Proceedings,* Dongarra, J. et al., Eds., Springer Verlag, Berlin, 2005, 776.
51. Bininda-Emonds, O.R.P. and Sanderson, M.J., Assessment of the accuracy of matrix representation with parsimony supertree construction, *Syst. Biol.,* 50, 565, 2001.
52. Roshan, U., *Fast Algorithmic Techniques for Large Scale Phylogenetic Reconstruction,* Ph.D. thesis, University of Texas at Austin, 2004.