Picking Fruit from the Tree of Life

Comments on Taxonomic Sampling and Quartet Methods

Jonathan H. Badger
Bioinformatics Research Group
Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
jhbadger@math.uwaterloo.ca

Paul Kearney
Bioinformatics Research Group
Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
pkearney@math.uwaterloo.ca

Keywords
Taxonomic Sampling, Phylogenetics, Quartet Methods.

ABSTRACT

A topic of recent interest and controversy in the field of systematic biology is the value of "taxonomic sampling", the practice of adding additional sequences (taxa) to an analysis to improve the accuracy of the inferred evolutionary tree. In terms of tree inference algorithms that construct trees from four taxa subtrees (quartet topologies), the value of taxonomic sampling can be rephrased as the question "are quartet topologies more accurately estimated when embedded within a larger set of taxa?". Here we show that the answer to this question is negative, based on an analysis of nine 40 taxa trees with varying amounts of sequence divergence sampled from the Ribosomal Database Project. This result complements and contrasts previous research that examined the effects of taxonomic sampling on a single pathological quartet topology using artificially generated data. Our result is based on an experimental study using real data and examines the effect of taxonomic sampling on all quartet topologies induced by an evolutionary tree.

1. INTRODUCTION

Large DNA sequence datasets available for evolutionary analyses have been generated in recent years through the use of DNA sequencing technology. Examples of such datasets are the Ribosomal Database Project's [13] prokaryotic and eukaryotic datasets and the Green Plant Phylogeny [2] dataset which include thousands of gene sequences. A critical component of evolutionary analyses of these datasets is the confident estimation of the evolutionary history of the sequences. Such histories are typically modeled by evolutionary trees. The accurate estimation of evolutionary trees is a challenging biological and computational problem. The biological challenges are numerous and include the fact that our current understanding of evolutionary processes, especially during the early stages of life on earth, is far from adequate. The computational challenges arise from the fact that the number of possible evolutionary trees on a given set of sequences is exponential in the number of sequences and for most objective functions that evaluate a given evolutionary tree hypothesis, the problem of finding an evolutionary tree that maximizes the objective function is intractable.

Many methods for estimating evolutionary trees have been proposed. These include maximum parsimony [7], maximum likelihood [6], distance methods such as neighbor joining [15] and quartet methods [17, 1, 4, 3]. In this paper we examine the performance of quartet methods in the context of taxonomic sampling.

1.1 Quartet Methods and Taxonomic Sampling

The topology of an evolutionary tree is uniquely defined by its set of quartet topologies. Let S be the set of sequences labeling the leaves of an evolutionary tree T. A *quartet* of S is a set of four sequences taken from S, or equivalently, four leaves taken from T. A *quartet topology* is an evolutionary tree on four sequences and can take one of the four forms depicted in figure 1.

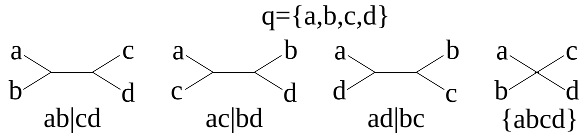


Figure 1: The four possible quartet topologies. We use the notation ab|cd to to denote the quartet topology where sequences a and b are separated from sequences c and d as depicted in the leftmost quartet topology.

We let Q_T denote the set of quartet topologies induced by the evolutionary tree T. Since Q_T defines the topology of T, a reasonable approach to estimating the topology of T from sequence data is the following two step approach:

1. Estimate Q_T by generating a set Q of quartet topologies inferred from the sequence data using a method such as maximum likelihood, maximum parsimony or neighbor joining.
2. Recombine quartet topologies in Q (like pieces of a puzzle) to form an estimate of the unknown evolutionary tree T.

For example, quartet puzzling [17] uses maximum likelihood to generate Q and then recombines these quartet topologies using a greedy iterative algorithm. When recombining quartet topologies

SAC 2001, Las Vegas, NV
© 2001 ACM 1-58113-287-5/01/02…$5.00

most quartet methods try to realize as many quartet topologies in Q as possible by obtaining a tree T' that maximizes the intersection of $Q_{T'}$ and Q.

The motivation for the quartet method is that, although large evolutionary trees cannot be estimated directly using computational intense methods such as maximum likelihood and maximum parsimony, these methods can be used to estimate all quartet topologies. That is, even if the entire evolutionary tree cannot be estimated, pieces of the evolutionary tree can be.

Critical to the feasibility of the quartet method is that quartet topologies can be estimated accurately. This is intimately connected to the concept of taxonomic sampling. Hendy and Penny [10] introduced the idea that adding taxa (in our case sequences) to the dataset so that long branches of the evolutionary tree are shortened may increase the accuracy of the resulting estimate. A controversy surrounding taxonomic sampling began when Hillis [11], motivated by anecdotal evidence, stated:

> *Including large numbers of taxa in an analysis may be the best way to ensure phylogenetic accuracy.*

This resulted in many papers presenting research validating, criticizing or clarifying Hillis' statement (e.g. [12, 8, 14, 16]). Returning to the quartet method, the relevant question is:

> *Are quartet topologies more accurately estimated when embedded within a larger set of taxa?*

That is, can the quartet topology for a given quartet be more accurately inferred by first estimating an evolutionary tree for a larger dataset that includes the quartet ("supersample the quartet") and then extracting the quartet topology from this evolutionary tree. We address the above question in this paper using an experimental study based on the Ribosomal Database Project [13].

1.2 Previous Work

Previous related work has focused on the effects of taxonomic sampling on a "Felsenstein zone" quartet topology (see [5]). The Felsenstein zone is an area of the parameter space where a method will converge upon the wrong topology as the amount of sequence data increases. We describe this work in more detail.

Graybeal [8] used a simulation study to examine the effect of adding more taxa to a Felsenstein zone quartet topology. In her study, Graybeal inserted more and more taxa into the quartet topology at prespecified locations and then evolved artificial sequences on the resulting evolutionary tree according to the Kimura 2 parameter model with rate variation among sites. She then applied maximum parsimony, and in some cases maximum likelihood, to the resulting sequences and assessed the accuracy of the quartet topology. She found that as the number of taxa is increased while the amount of sequence data remains constant, the accuracy of the quartet topology increases. However, if the number of taxa is increased too much, accuracy begins to decline. Graybeal hypothesized that the eventual loss of accuracy was due to the fact that the amount of sequence data was kept constant, and so, as the number of taxa is increased it is expected that overall accuracy will naturally decrease due to low sequence to taxa ratios. Although the study indicated that there was an advantage to supersampling a quartet topology when using maximum parsimony, the study did not indicate an advantage when using maximum likelihood.

Smith and Warnow [16] also utilized a simulation study that examined the effects of adding more taxa to a Felsenstein zone quartet topology. However, in their study they examined maximum parsimony and neighbor joining and used the Jukes Cantor model of evolution to artificially evolve sequences on a variety of model trees. The authors found that maximum parsimony and neighbor joining can both benefit from supersampling the quartet topology when sequence length is sufficiently long. The authors also observed instances where accuracy was decreased by supersampling and they hypothesized that this could be the result of new Felsenstein zone quartet topologies being created by the addition of new taxa.

Hillis observed that quartet topologies can be difficult to estimate when evolutionary rates are high [11]. In contrast to the above two simulation studies, Hillis offered anecdotal evidence indicating that supersampling quartet topologies can improve accuracy. Hillis evolved sequences using the Kimura 2 parameter model of evolution with rate variance and a high rate of evolution. Hillis observed that almost all quartet topologies across a particular edge of the model tree are incorrectly estimated by maximum parsimony but that maximum parsimony can be used to correctly estimate these quartet topologies when supersampled.

The above research suggests that supersampling a quartet topology increases accuracy. However, this conclusion must be tempered by the following observations:

- The above research focuses on the effect of supersampling a Felsenstein zone quartet topology. Such quartet topologies are not representative of the entire distribution of quartet topologies induced by an evolutionary tree. From the perspective of the quartet method it is essential to understand how *all* quartet topologies induced by an evolutionary tree are affected by taxonomic sampling. We observe that Felsenstein zone quartet topologies are difficult to estimate to begin with, and so, increased accuracy by supersampling is not surprising.

- It is not clear how much the observed effects of supersampling are the result of the chosen method of estimation. Graybeal, Smith and Warnow, and Hillis all observe improved accuracy when using maximum parsimony. Graybeal examined maximum likelihood which yielded no improvement due to supersampling. However, Graybeal acknowledge that maximum likelihood perfectly matched the Kimura 2 parameter model of evolution used in her simulation study. In this sense the simulation study was not a reasonable evaluation of maximum likelihood. Smith and Warnow examined the effects of supersampling when using neighbor joining. However, they used the Jukes Cantor model of evolution without rate variance on which neighbor joining is known to be consistent.

- The Graybeal study supersampled a quartet topology by inserting new taxa at prespecified locations. In practice a biologist does not have the option of selecting insertion points for taxa on a long branch in the tree, nor does a biologist even know *a priori* which taxa would insert on a long branch. In contrast, Smith and Warnow begin with a model tree from which a supersample of the quartet topology in question is selected. Smith and Warnow use artificially generated trees with uniform branch lengths along with a single 35 taxon sampled from the rbcL dataset. Hillis examined a single branch of a single tree. Of the three studies, Smith and Warnow use the most robust set of model trees yet it is unclear how the chosen uniform branch lengths might affect their conclusions.

2. METHODOLOGY

The experimental study presented here examines the simultaneous effect of supersampling quartet topologies over all quartets of a given dataset. In the study various methods are used including maximum parsimony, quartet puzzling and neighbor joining (maximum likelihood trials are still in progress). The datasets (sequences and associated evolutionary trees) are extracted from the Ribosomal Database Project (RDP) [13] and represent sequence sets with various degrees of divergence.

Our experimental study is designed to address the issues not addressed by previous work:

- We examine the effect of supersampling on all quartet topologies induced by an evolutionary tree, not just Felsenstein zone topologies. This is necessary in order to access the impact of taxonomic sampling on quartet methods as well as to access the effects of taxonomic sampling on quartet topologies in general.
- A broad range of methods are used to determine the relative impact of taxonomic sampling on quartet methods.
- Instead of generating artificial sequences on artificial evolutionary trees we have chosen to use real sequences and real evolutionary trees[1]. This allows us to avoid some of the pitfalls of simulation studies such as simplistic models of evolution and cladogenesis and perhaps makes the conclusions more practically relevant.

In all, we extracted nine trees of 40 taxa each from the RDP 16S rRNA tree. In order to determine what effect different levels of sequence divergence have on supersampling, three trees were extracted from each of three divergence levels (shallow, medium, and deep). The details of the extracted trees are found below. 1000 random quartets from each tree, uniformly distributed across the edges of the tree were chosen for tracking. This was necessary in order to make computation time feasible. For each quartet, 10 supersamples of 5, 10, 15, 20, and 30 sequences were analyzed by maximum parsimony, neighbor joining and quartet puzzling using the RDP 16S rRNA alignment. These supersamples were created by adding randomly chosen taxa from the 40 taxa tree to the quartet to obtain a tree with the desired number of taxa. Additionally the quartet itself and the complete 40 taxa tree were also estimated by the methods. For each set of ten trees the following statistics were computed:

- **Quartet Accuracy:** This statistic is the percentage of times that the RDP quartet being supersampled was correctly inferred in the trees supersampled from it.
- **Consensus Accuracy:** For each set of supersamples, the consensus (majority) version of the quartet containing the species from the RDP quartet was computed. The percentage of times in which the analogous quartet from each of the supersampled trees agrees with the consensus quartet is termed “consensus accuracy”. This statistic provides a measure of consistency independent of the RDP tree.
- **Tree Accuracy:** This statistic is the percentage of edges in the supersampled trees that were correctly inferred in regard to the RDP.

The 40 taxa trees were randomly sampled from monophyletic groups of the RDP tree. To avoid unreasonably short branches, only one strain per species was considered for inclusion. The basis of the trees were the following groups: shallow tree #1 (Enteric Bacteria), medium tree #1 (Cyanobacteria), deep tree #1 (Archaea + Eubacteria), shallow tree #2 (Pseudomonas), medium tree #2 (Methanogenic Archaea), deep tree #2 (Archaea + Eubacteria), shallow tree #3 (Arthrobacter), medium tree #3 (Lactobacilli), deep tree #3 (Archaea + Eubacteria). Table 1 presents a summary of the branch lengths of these trees, and supports our assertion that our trees represent a broad range of sequence divergence.

The phylogenetic methods used in the analysis were the following: neighbor-joining [15] as implemented in PAUP* 4.0b4a [18], using maximum likelihood derived distances based on the Hasegawa-Kishino-Yano (HKY) model of evolution [9], heuristic maximum parsimony [7] also as implemented in PAUP* 4.0b4a, and finally quartet puzzling [17], on maximum likelihood inferred quartets, as implemented in PUZZLE 4.0.2.

3. RESULTS AND CONCLUSIONS

The following graphs present our results. In each graph, the x-axis presents the supersample sizes and the y-axis represents the percent accuracy. Figures 2 through 4 depict the how supersampling affected the accuracy of the tracked quartets as compared to the RDP quartet topology, while figures 5 through 7 depict how supersampling affected the consensus accuracy of the tracked quartets. Finally, figures 8 through 10 depict how edge accuracy is affected as larger and larger datasets are analyzed.

[1] The RDP evolutionary trees used are themselves estimates. However, the RDP is a carefully assembled and maintained evolutionary tree with some level of confidence.

Table 1: Summary of the branch lengths of the sampled trees

	1S	1M	1D	2S	2M	2D	3S	3M	3D
Min.	0.0	0.0025	0.0020	0.0001	0.0	0.0033	0.0	0.0005	0.0028
Max.	0.0361	0.0743	0.5231	0.1088	0.1010	0.8893	0.0288	0.0859	0.7504
Ave.	0.0032	0.0101	0.0521	0.0055	0.0094	0.0697	0.0035	0.0064	0.0630
Std. Dev.	0.0059	0.0170	0.1195	0.0156	0.0181	0.1359	0.0060	0.0129	0.1257

(1S means shallow tree #1, 1M means medium tree #1, etc.)

Figure 2: Quartet Accuracy – Neighbor Joining

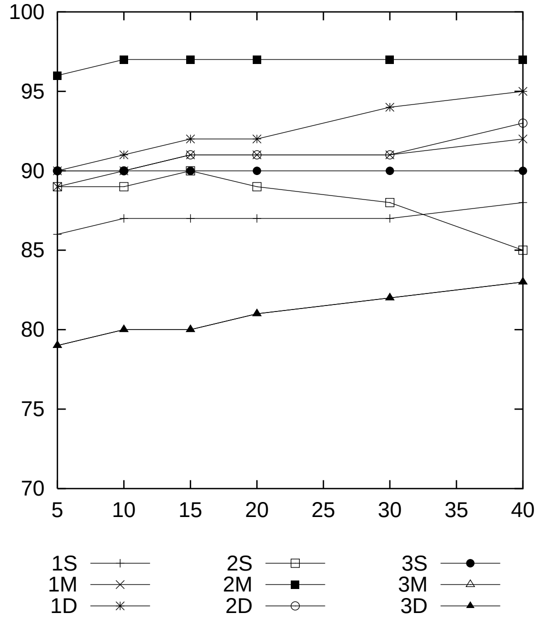


Figure 4: Quartet Accuracy – Quartet Puzzling

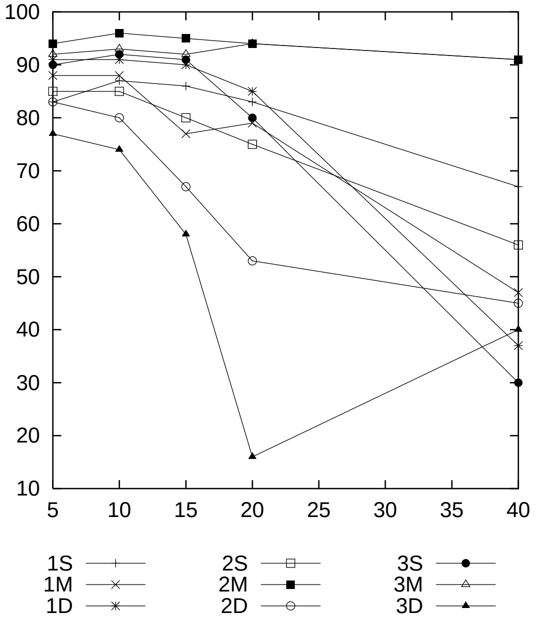


Figure 3: Quartet Accuracy – Maximum Parsimony

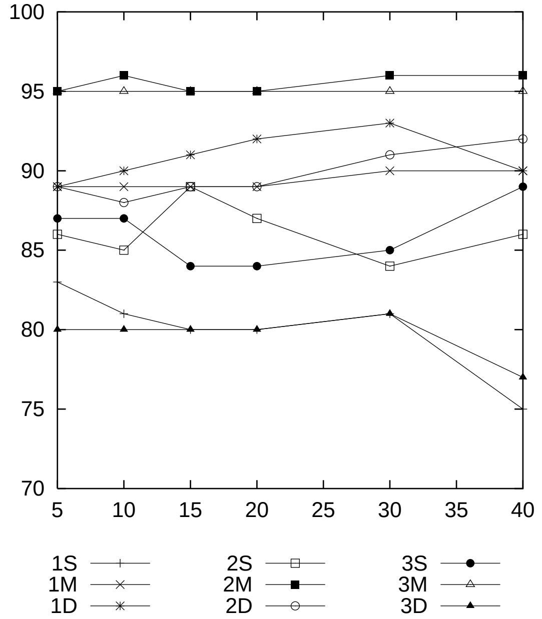


Figure 5: Consensus Accuracy – Neighbor Joining

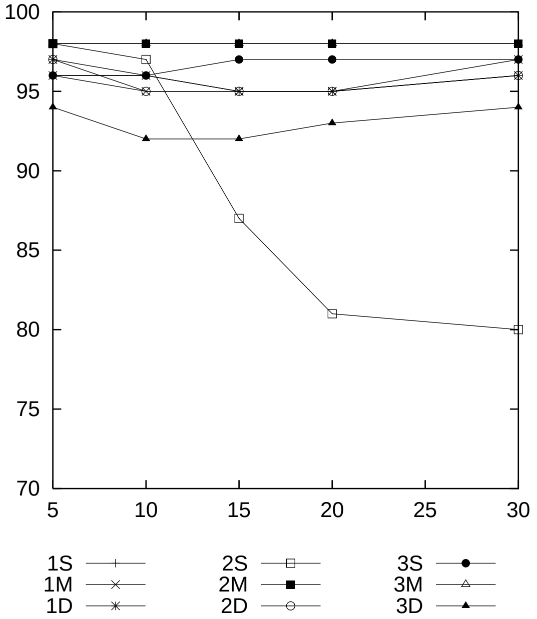


Figure 6: Consensus Accuracy – Maximum Parsimony

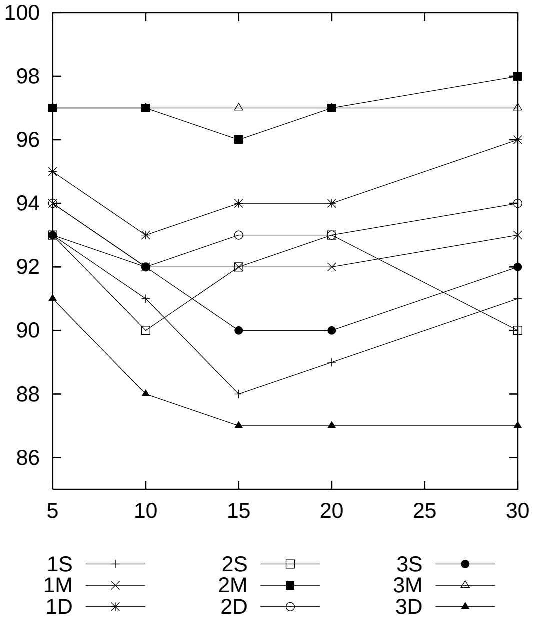


Figure 7: Consensus Accuracy – Quartet Puzzling

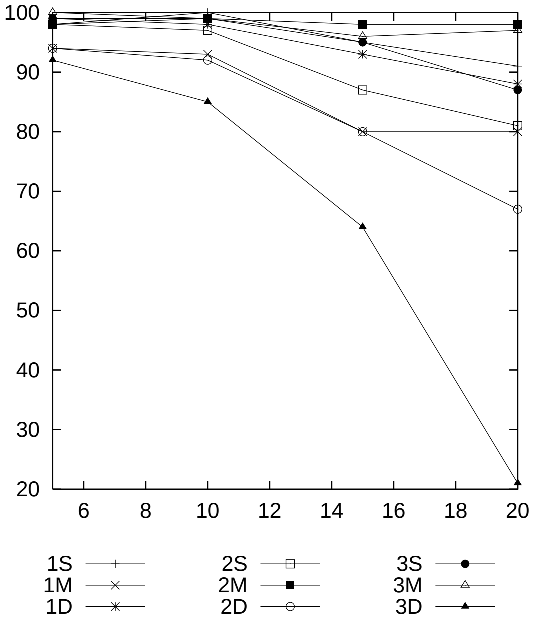


Figure 8: Tree Accuracy – Neighbor Joining

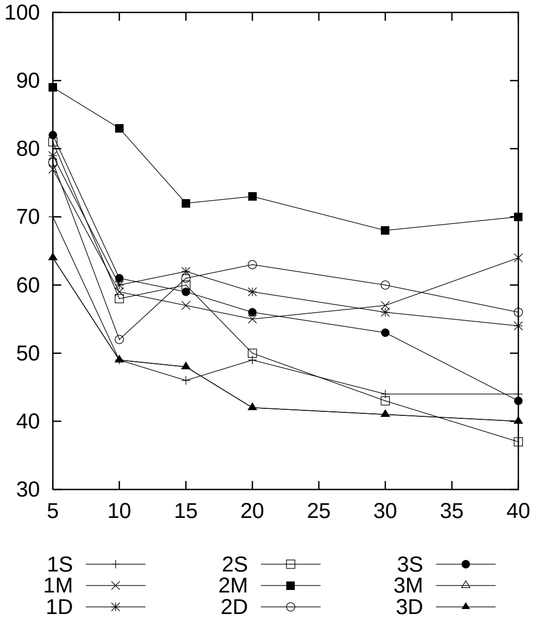
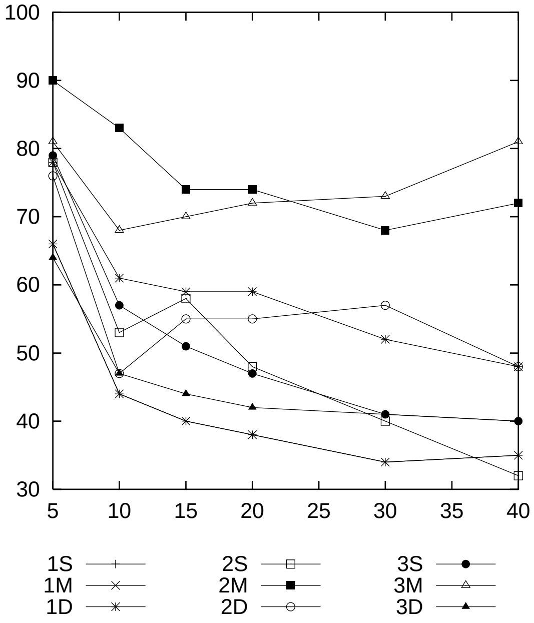


Figure 9: Tree Accuracy – Maximum Parsimony



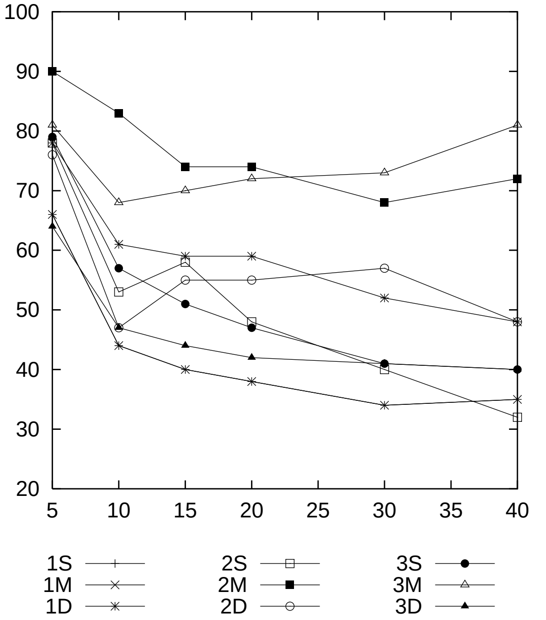


Figure 10: Tree Accuracy – Quartet Puzzling

The results of the experimental study strongly support the conclusion that supersampling does not increase quartet topology accuracy. This conclusion is independent of method and of dataset. The average increase in accuracy afforded by supersampling is less than 2.5% for neighbor joining, less than 1.2% for maximum parsimony and 0% for quartet puzzling. Furthermore, in many cases supersampling resulted in a decrease in accuracy. The consensus accuracy, which is independent of the RDP tree, also supports this conclusion as a similar decrease in accuracy is also observed. This result does not necessarily contradict the conclusions of Graybeal, Smith and Warnow, and Hillis as their conclusions were based on examining the effect of taxonomic sampling on a single pathological quartet topology. Our results indicate that the beneficial effects of taxonomic sampling on collections of quartet topologies is minimal. The reason for this is not entirely clear. One possibility is that the number of pathological quartet topologies in a dataset is small. Another possibility is that supersampling introduces additional pathological quartet topologies thereby negating any benefit of adding additional taxa.

The results of the experimental study also support the conclusion that smaller datasets are more accurately estimated than larger datasets. This is true regardless of inference method. In fact, the reduction in edge accuracy can be dramatic, for example, maximum parsimony and neighbor joining on the 2S dataset and maximum parsimony on the 1M dataset. This is consistent with the conclusions of the Smith and Warnow simulation study. The surprisingly poor performance of quartet puzzling in our study (sometimes going below the 33% accuracy expected by chance alone) can be explained by the fact that the trees obtained by this method are often highly unresolved. The edges that are resolved, however, are generally quite accurate.

We conclude that although certain (pathological) quartet topologies may benefit from supersampling, there is no evidence to suggest that a collection of quartet topologies drawn from the same dataset will benefit collectively from supersampling. The implication of this for quartet methods is that supersampling cannot be used to amplify the accuracy of a set of quartet topologies. Although we took efforts to sample broadly from the RDP, one must be cautious in extending the results of this study to other datasets. 16S rRNA is a highly conserved molecule, and it is possible that the types of pathological quartets that supersampling is thought to eliminate are rare in this data. Future work should include studies of other molecules, such as the protein rbcL. Additionally, investigating other phylogenetic methods, in particular maximum likelihood, to see if they too are immune to the benefits of supersampling would provide additional practical knowledge to biologists.

4. ACKNOWLEDGMENTS

Jonathan Badger was supported by CITO, and Paul Kearney by NSERC and CITO. Both authors gratefully acknowledge the culinary support of Mel's Diner, Waterloo.

5. BIOGRAPHICAL NOTE

Jonathan Badger received his doctorate in microbiology in 1999 from the University of Illinois at Urbana-Champaign. He is now a postdoctoral fellow in the Computer Science department of the University of Waterloo, and his research focuses on phylogenetic inference and algorithms for biological sequence analysis. Paul Kearney received his doctorate in Computer Science in 1997 from the University of Toronto. He is now an assistant professor in the Computer Science department of the University of Waterloo and director of UW's undergraduate program in bioinformatics. His research focuses on phylogenetic inference and algorithms for analyzing proteomic data.

6. REFERENCES

[1] V. Berry and O. Gascuel. Inferring evolutionary trees with strong combinatorial evidence. *Proceedings of the Third Annual International Computing and Combinatorics Conference*, pages 111–123, 1997.

[2] K. S. Brown. Deep Green rewrites evolutionary history of plants. *Science*, 285:990–991, 1999.

[3] D. Bryant, V. Berry, P. Kearney, M. Li, T. Jiang, T. Wareham, and H. Zhang. A practical algorithm for recovering the best supported edges of an evolutionary tree. In *Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms*, 2000.

[4] P. Erdös, M. Steel, L. Székely, and T. Warnow. Constructing big trees from short sequences. *Proceedings of the 24th International Colloquium on Automata, Languages, and Programming*, 1997.

[5] J. Felsenstein. Cases in which parsimony and compatibility will be positively misleading. *Sys. Zoo.*, 27:401–410, 1978.

[6] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[7] W. M. Fitch. Toward defining the course of evolution: Minimal change for a specific tree topology. *Sys. Zoo.*, 20:406–41, 1971.

[8] A. Graybeal. Is it better to add taxa or characters to a difficult phylogenetic problem? *Sys. Biol.*, 47:9–17, 1998.

[9] M. Hasegawa, H. Kishino, and K. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22:160–174, 1985.

[10] M. D. Hendy and D. Penny. A framework for the quantitative study of evolutionary trees. *Sys. Zoo.*, 38:297–309, 1989.

[11] D. M. Hillis. Inferring complex phylogenies. *Nature*, 383:130, 1996.

[12] J. Kim. General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa. *Sys. Biol.*, 45:363–374, 1996.

[13] B. L. Maidak, J. R. Cole, C. T. Parker, G. M. G. Jr, N. Larsen, B. Li, T. G. Lilburn, M. J. McCaughey, G. J. Olsen, R. Overbeek, S. Pramanik, T. M. Schmidt, J. M. Tiedje, and C. R. Woese. A new version of the RDP (Ribosomal Database Project). *Nucleic Acids Research*, 27:171–173, 1999.

[14] S. Poe. Sensitivity of phylogeny estimation to taxonomic sampling. *Sys. Biol.*, 47:18–31, 1998.

[15] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.

[16] K. Smith and T. Warnow. Taxon sampling and accuracy of evolutionary tree reconstruction. In *DIMACS Symposium on Large Phylogenetic Tree Reconstruction*, 1998.

[17] K. Strimmer and A. von Haeseler. Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.*, 13:964–969, 1996.

[18] D. L. Swofford. *PAUP* (Phylogenetic Analysis Using Parsimony and Other Methods) version 4.0b4a.* Sinauer Associates, Sunderland, Massachusetts., 2000.