CS 598AGB
What simulations can tell us
Questions that simulations cannot answer

• Simulations are on finite data.
• Some questions (e.g., whether a method is statistically consistent) are about what happens as amount of data increases to infinity.
• Therefore, simulations cannot prove or disprove statistical consistency. (You need a mathematical proof for those.)
Questions simulations can address

• Every simulation is by necessity limited to its model condition(s). Therefore, try to avoid over-enthusiastic interpretations of results. Even so...

• Simulations can address:
  – Relative performance of methods
  – Factors that impact performance

• Simulations can also indicate gaps in theoretical understanding, and suggest theoretical results that you can then try to prove (or disprove).
Examples

• Many studies on four-leaf trees:
  – Success of Phylogenetic Methods in the Four-Taxon Case (Huelsenbeck & Hillis, Systematic Biology 1993)

• Main objectives: comparing methods, determining model conditions under which methods such as parsimony seem to be consistent.
Phylogeny Estimation Methods

- Distance-based methods
  - Naïve Quartet Method and other quartet-based methods
  - Neighbor joining (NJ), BioNJ, FastME, Weightor, etc.
  - $DCM_{NJ}$, HGT+FP, etc. ("afc" methods)
- Maximum Parsimony (MP) and Maximum Compatibility (MC)
- Maximum Likelihood (ML)
- Bayesian MCMC (e.g., MrBayes)

What do you know about these methods, and what would you predict about the relative accuracy of these methods, based on what you know?
Papers

• Performance study of phylogenetic methods: (unweighted) quartet methods and neighbor joining, St. John et al., J. Algorithms, 2003 (paper #61)

• Designing Fast Converging Phylogenetic Methods, Nakhleh et al., ISMB 2001 and Bioinformatics 2001 (paper #44) – see also paper #40 for background theory

• The Performance of Phylogenetic Methods on Trees of Bounded Diameter, Nakhleh et al., WABI 2001 (paper #45)

• The Accuracy of Fast Phylogenetic Methods for Large Datasets, Nakhleh et al., PSB 2002 (paper #48)
Quartet-based methods vs. NJ

• Paper: St. John et al., J. Algorithms, 2003 (paper #61)
• Basic question: how accurate are trees computed using quartet-based phylogeny estimation methods, compared to Neighbor Joining (NJ)?
• Quartet-based methods (all better than NQM):
  – Q* (aka “the Buneman tree”)
  – Quartet-cleaning (Berry and Gascuel 2000, Berry et al., SODA 2000)
  – Quartet Puzzling (Strimmer and von Haeseler, MBE 1996)
Fig. 13. Accuracy of various methods as a function of sequence length for 40 taxa.

Fig. 14. The sequence length required to accurately reconstruct all of the edges 90% of the time. For low rates of evolution (left), QCNJ and NJ perform comparably. At higher rates of evolution (right), NJ consistently requires shorter sequences to reconstruct the true tree.
Observations and Conclusions

• Quartet-based methods are generally less accurate than NJ, but the differences depend on the rate of evolution (and may depend on other model parameters as well)
• The theoretical guarantees for NJ and NQM (and hence for the quartet-based methods studied) are identical, so theory does not explain differences in performance.
• Main hypothesis: independently estimating quartet trees is less powerful than estimating a large tree – “taxon sampling” improves tree estimation.

• “NJ should be regarded as a universal lowest common denominator in phylogenetics...We suggest that a proposed method should be compared with NJ and abandoned if it does not offer a demonstrable advantage over NJ for substantial subproblem families.” (St. John et al., J. Alg., 2001)
Absolute Fast Converging Methods

• NJ, Maximum Likelihood, and many other phylogeny estimation methods are statistically consistent.

• But how much data (sequence length) do they need to reconstruct the true tree with high probability?

• This can be studied theoretically or empirically
Statistical consistency, exponential convergence, and absolute fast convergence (afc)
Absolute Fast Converging Methods

**Definition 1.** Let $f, g \geq 0$. Define $\mathcal{M}_{f,g} = \{(T, \{\lambda(e)\}) : \forall e \in E(T), f \leq \lambda(e) \leq g\}$.

We now define absolute fast convergence:

**Definition 2.** A phylogenetic reconstruction method $\Phi$ is (absolute) fast-converging (afc) for the model $\mathcal{M}$ if, for all positive $f, g, \varepsilon$, there is a polynomial $p$ such that, for all $(T, \{\lambda(e)\}) \in \mathcal{M}_{f,g}$, on a set $S$ of $n$ sequences of length at least $p(n)$ generated on $T$, we have $Pr[\Phi(S) = T] > 1 - \varepsilon$.

From Nakhleh et al., ISMB 2001 and Bioinformatics 2001; see also Warnow et al., SODA 2001 (paper #40)
Neighbor joining has poor performance on large diameter trees [Nakhleh et al. ISMB 2001]

Theorem (Atteson): NJ is accurate with high probability given exponential sequence lengths
DCM1-boosting distance-based methods

[Nakhleh et al. ISMB 2001]

Theoretical guarantees that DCM1-NJ converges to the true tree from polynomial length sequences.
Fig. 9. DCM-NJ+MP vs. DCM*-NJ vs. NJ vs. HGT+FP on random trees under the K2P+Gamma model. Sequence length is 1000. Average branch length is 0.05.

Figure 9 from Nakhleh et al., ISMB 2001 and Bioinformatics 2001
**Fig. 3.** DCM*-NJ vs. NJ vs. HGT+FP on the rbcL 500-taxon tree, under the JC model. Average branch length is 0.264
Observations

• Several AFC methods (HGT-FP, DCM*-NJ, short quartet methods), but they are not all equally accurate – and may not even be more accurate than NJ (or other non-AFC methods) under some model conditions.

• AFC is a theoretical statement about asymptotic performance, but like statements about being polynomial time – the constants are hidden.

• Even so – the theory predicts difficulties for NJ given *high evolutionary distances*, and less error for AFC methods under those conditions.
Other questions

• Impact of rate of evolution?
• Impact of sequence length?
• Impact of number of taxa?
• Impact of substitution model (Jukes-Cantor, K2P, GTR, General Markov Model, etc.)?
• Comparison to MP, ML, Bayesian MCMC, etc?
These experiments suggest strongly that except for low diameter situations, the DCM-NJ+MP method (and probably the other "fast-converging" methods) will outperform the neighbor joining method, especially for large numbers of taxa and high evolutionary rates.

**Fig. 3.** NJ (left graph) and DCM-NJ+MP (right graph) error rates on random birth-death trees as the diameter (x-axis) grows. Sequence length fixed at 500, and deviation factor fixed at 4.

**Figure 3 from** Nakhleh et al., WABI 2001
5.4 The Influence of the Model of Sequence Evolution

We reported all results so far under the K2P+Gamma model only. However, we explored performance under the JC (Jukes-Cantor) model as well. The relative performance of the methods we studied was the same under the JC model as under the K2P+Gamma model. However, throughout the experiments, the error rate of the methods was lower under the JC model (using the JC distance-correction formulas) than under the K2P+Gamma model of evolution (using the K2P+Gamma distance-correction formulas). This might be expected for the Weighbor method, which is optimized for the JC model, but is not as easily explained for the other methods. Figure 6 shows the error rate of NJ on trees of diameter 0.4 under the two models of evolution. NJ clearly does better under the JC model than under the K2P+Gamma model; other methods result in similar curves. Correlating the decrease in performance with specific features in the model is a challenge, but the results clearly indicate that experimentation with various models of evolution (beyond the simple JC model) is an important requirement in any study.

6 Conclusion

In earlier studies we presented the DCM-NJ+MP method and showed that it outperformed the NJ method for random trees drawn from the uniform distribution on tree topologies and branch lengths as well as for trees drawn from an more biologically realistic distribution, in which the trees are birth-death trees with a moderate deviation from ultrametricity. Here we have extended our result to include the Weighbor and
Figure 3: Accuracy as a function of the number of taxa under the K2P+Gamma model for expected diameter (2.0) and two sequence lengths (a) 200 taxa (b) 400 taxa.

Figure 4: Accuracy as a function of the sequence length under the K2P+Gamma model for expected diameter (2.0) and two numbers of taxa.

The dominance of DCM-NJ+MP is once again evident. Comparing NJ and Weighbor, we can see that NJ is actually marginally better than Weighbor at lower diameters, but Weighbor clearly dominates it at higher diameters—the two slopes are quite distinct.

Figure 4 from Nakhleh et al., PSB 2002
The dominance of DCM-NJ+MP is once again evident. Comparing NJ and Weighbor, we can see that NJ is actually marginally better than Weighbor at low diameters, but Weighbor clearly dominates it at higher diameters—the two slopes are quite distinct.

Figure 3(a), Nakhleh et al., PSB 2002
5.2 Sequence-Length Requirements

We can sort our experimental data in terms of accuracy and, for all datasets on which an accuracy threshold is met, count, for each fixed number of taxa, the number of datasets with a given sequence length, thereby enabling us to plot the average sequence length needed to guarantee a given maximal error rate. We show such plots for two accuracy values in Figure 1: 70% and 85%. Larger values of accuracy cannot be plotted reliably, since they are rarely reached under our challenging experimental conditions. The striking feature in these plots is the difference between the two NJ-based methods (NJ and Weighbor) and the methods using parsimony (DCM-NJ+MP and Greedy Parsimony): as the number of taxa increases, the former require longer and longer sequences, growing linearly or worse, while the latter exhibit only modest growth. The divide-and-conquer strategy of DCM-NJ+MP pays off by letting its NJ component work only on significantly smaller subsets of taxa—effectively shifting the graph to the left—and completing the work with a method (parsimony) that is evidently much less demanding in terms of sequence lengths. Note that the curves are steeper for the higher accuracy requirement: as the accuracy keeps increasing, we expect to see supralinear, indeed possibly exponential, scaling.

Figure 1: Sequence length requirements under the K2P+Gamma model as a function of the number of taxa

(a) 70% accuracy
   (b) 85% accuracy

Figure 1 from Nakhleh et al., PSB 2002
Greedy Parsimony methods. Our results confirm that the accuracy of the NJ method may suffer significantly on large datasets. They also indicate that Greedy Parsimony, while very fast, has mediocre to poor accuracy, while Weighbor and DCM-NJ+MP consistently return good trees, with Weighbor doing better on shorter sequences and DCM-NJ+MP doing better on longer sequences. Among interesting questions that arise are: (i) is there a way to conduct a partial parsimony search that scales no worse than quadratically (and might outperform DCM-NJ+MP)? (ii) would DCM-Weighbor+MP prove a worthwhile tradeoff? (iii) can we make quantitative statements about the accuracy achievable by any method (not just one of those under study) as a function of some of the model parameters?

Acknowledgments

This work was supported in part by the National Science Foundation with grants EIA 99-85991 to T. Warnow and ACI 00-81404 to B.M.E. Moret and a PIRE award to K. St. John; by the Texas Institute for Computational and Applied Mathematics and the Center for Computational Biology at UT-Austin (K. St. John); and by the David and Lucile Packard Foundation (T. Warnow).

References


Figure 6: Accuracy of NJ as a function of the number of taxa under JC and K2P+Gamma

Figure 6 from Nakhleh et al., PSB 2002
Things to think about

• What did you see about performance of MP and other methods that are not guaranteed to be statistically consistent?
• What was the best performing method so far?
• What would you suggest as a method to compute 4-taxon trees?
• What would you suggest as a method to compute a large tree with low evolutionary diameter?
• What would you suggest as a method to compute a large tree with high evolutionary diameter?
• Are there gaps between theoretical understanding and empirical performance for any methods?
Questions we did not address

• Performance of ML and Bayesian MCMC
• Impact of model misspecification
• What is easier – small trees or big trees?
• Impact of alignment error
• Species tree estimation from multiple genes
Are the big easy?

- Inferring complex phylogenies (Hillis, Nature 1996)
- Follow-up articles and letters, including