Using Parsimony to Guide Maximum Likelihood Searches

Kenneth Sundberg*,Timothy O'Connor†,Hyrum Carroll*, Mark Clement*, and Quinn Snell*
*Computer Science Department
Brigham Young University
Provo, Utah 84606
Email: kasundberg@gmail.com, hdc@cs.byu.edu, clement@cs.byu.edu, snell@cs.byu.edu
†Integrative Biology Department
Brigham Young University
Provo, Utah 84606
Email: timothydoconnor@gmail.com

***Abstract*—The performance of maximum likelihood searches can be boosted by using the most parsimonious tree as a starting point for the search. The time spent in performing the parsimony search to find this starting tree is insignificant compared to the time spent in the maximum likelihood search, leading to an overall gain in search time. These parsimony boosted maximum likelihood searches lead to topologies with scores statistically similar to the unboosted searches, but in less time.**

I. Introduction

There are two common methods for inferring phylogenies from multiple alignment data, maximum parsimony [2] (MP) and maximum likelihood [10] (ML). The search problem for both methods is known to be NP-Hard [6] with parsimony known to also be NP-Complete [7], however the scoring of individual topologies during the search is quite different. The parsimony score of a topology can be computed in linear time with respect to the number of nodes in the topology, whereas the best known ML computations run in exponential time with respect to the number of nodes. As a result MP searches are much faster than ML searches.

Unfortunately parsimony has a strong bias toward long branch attraction, and can lead to positively misleading topologies [9]. Maximum likelihood also models more of the biology, including probabilities and estimates of branch lengths, information that can be required by other methods of phylogenetic analysis. As a result of the flaws of parsimony and the additional information provided by maximum likelihood, maximum likelihood is preferred by many researchers.

II. Data Sets

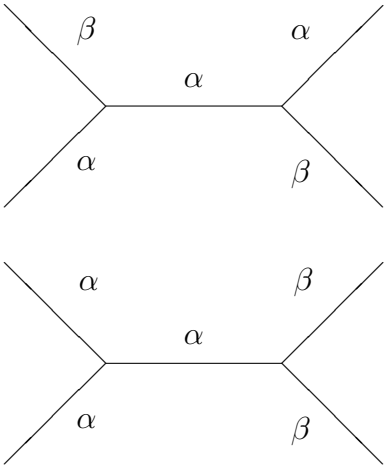
For this work we used three types of alignment data sets: four taxa synthetic data, small real data sets and large data sets.

A. Synthetic Alignments

The four taxa synthetic data sets were used to establish areas of general concordance between ML and MP. To produce them we used the program Dawg [5] under a General Time Reversible (GTR) [26], [17], [22] model of evolution. We modeled the parameters according to the examples included with the program and explored a range of branch lengths as seen in Figure 1. The lambda value of 0.1 was used for the indel evolution rate and can be interpreted as one indel for every ten substitutions. The sequence length was increased to 2000 as this gives a reasonably sized sequence to allow for the expected value of any simulation to be seen. We ran two types of trees, a Felsenstein topology and a Farris topology, see Figure 2. Dawg generated data sets for trees under both topologies where the α and β branch lengths were varied from a branch length of 0.1 to 4.05 incremented by 0.05. A branch length of one is interpreted to mean that each site is expected to have one substitution from the internal node under the GTR definition of branch length. For each permutation of α and β branch lengths we ran a hundred replicates to get a percentage of matches between the two methods. This created a total of 1,280,000 data sets. In addition, this entire procedure was repeated with the the model changed to Jukes-Cantor (JC) [15], creating an additional 1,280,000 data sets.

```
Tree =
((a:0.1,b:0.25):0.05,
 (c:0.25,d:0.1):0.05);
Length = 2000
Model = "GTR"
#Rates of Substitution:
#AC, AG, AT, CG, CT, GT
Params = {1.5, 3.0, 0.9, 1.2, 2.5, 1.0}
#Frequencies of Nucleotides: A,C,T,G
Freqs  = {0.20, 0.30, 0.30, 0.20}
Format = "Phylip"
Lambda = 0.1
GapModel = "NB"
GapParams = {1,0.5}
```

Fig. 1. Typical file given to Dawg (Felsenstein zone)



Felsenstein Topology

Farris Topology

Fig. 2. The Felsenstein and Farris Trees

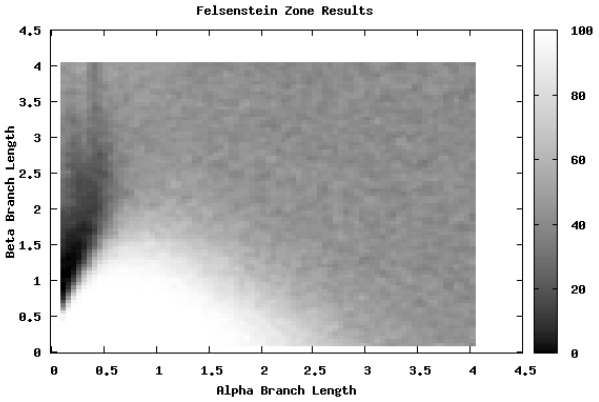


Fig. 3. Chart showing the percentage of datasets of a given true Felsenstein Topology where Maximum Parsimony and Maximum Likelihood predicted the same optimal topology.

B. Small Real Alignment Data Sets

We used all the data sets with nine, ten, and eleven taxa from the `mdsa_100s` version of the benchmarks published by Carroll *et al.* [3]. These 279 data sets are protein-coding DNA alignments, derived from reference amino acid alignments [27], [20], [8], [19]. The small real alignments were used to perform exhaustive MP and ML searches.

C. Large Data Sets

We also used several large data sets provided by Stamatakis (http://icwww.epfl.ch/~stamatak/index-Dateien/Page443.htm). These data sets include seven real alignments ranging from 101 to 500 taxa and 10 simulated data sets, each with 4000 taxa and 2000 base pairs. These data sets are used to evaluate our heuristic on large data sets using RAxML [24].

III. Correlations

Both parsimony and maximum likelihood methods try to extract the evolutionary signal out of a multiple sequence alignment. Though they do so in very different ways it is reasonable to expect a fair correspondence between the two methods.

A. Areas of General Concordance of ML and MP

There has been a large inquiry into the relationship of ML and MP from the perspective of accuracy and strength of methods [23], [13], [14], [16] but the comparison has never taken the view of how well one estimates the other. Here we have used simulated data to elucidate this relationship. To do so we focused on the standard four taxa case and compared the performance of both methods under a Felsenstein topology and a Farris topology.

This topology as seen in Figure 2 is a four taxa tree with non-sibling branches that have branch length β and all other branches with length α. The second tree in Figure 2 shows a Farris topology where the two β length branches are siblings.

The Felsenstein zone [9] can be loosely described as the subspace of all Felsenstein topology branch lengths where parsimony will predict a tree topology that is incorrect. Later work went on to show that ML performs well on Felsenstein topologies, including those that lie in the Felsenstein zone and reasonably well on Farris topologies. Also shown was that MP failed in certain areas of branch lengths in the Felsenstein topology but excelled at the Farris topology.[23], [13], [14], [16]

Parsimony, has a strong bias toward placing longer branches as siblings. This bias gives it an advantage when longer branches are 'truly' siblings and a disadvantage when the 'true' tree has non-sibling long branches.

To illustrate this we scored simulated alignments, as described in section II-A and compared the topological result produced from each method, the results are shown in Figures 3 and 4. These graphs show the percentage of time each method chose the same tree of the three possible trees.

As is clearly shown, one area in the Felsenstein topology space (the Felsenstein zone), as predicted by previous studies, does not correlate well forming the dark parabolic structure in Figure 3. This is due to the long branch attraction problem (LBA). From the perspective of parsimony as an estimator of likelihood topology, this should not be a problem because the number of long branches in a large data set are usually few and would only vary the topology by a few branches that are incorrect according to likelihood.

By running a likelihood search based on the best parsimony tree we can save time by getting the majority of branches correct and leaving likelihood to figure out the incorrect branches. This saves on the cost of using likelihood to search the whole tree space but lets it fine tune after the majority of bad trees are ruled out by parsimony. We are not trying to use parsimony as a function to predict ML but as a method to boost the search pass the large number of trees that both procedures predict are incorrect.

Most trees are assumed to be within the area where both

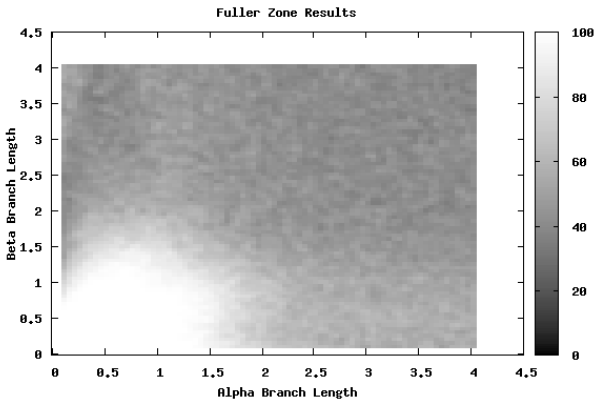


Fig. 4. Chart showing the percentage of datasets of a given true Farris Topology where Maximum Parsimony and Maximum Likelihood predicted the same optimal topology.

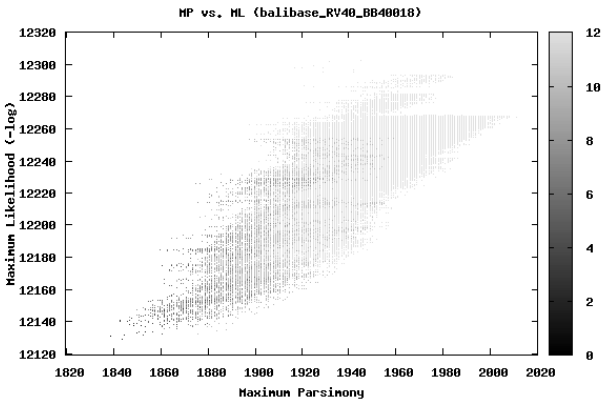


Fig. 5. All possible trees for BALiBASE dataset RV40 BB40010 scored with parsimony (GapMode = missing) and maximum likelihood (HKY under PAUP* defaults). Color indicates RF distance from optimal parsimony tree. Note that the optimal tree under likelihood is a small RF distance from the optimal parsimony tree.

methods are completely consistent, the light colored areas of Figures 3 and 4. In the gray areas of the figures the methods are doing no better than randomly guessing. This is to be expected as this occurs where there is an expected rate of 2 or more substitutions per site, which causes a large loss of phylogenetic signal. As expected, the major limitation to the concordance between ML and MP are those topologies that lie within the Felsenstein zone.

Each tree was scored using the phylogenetics program PAUP* [25] under both MP and ML. We derived the results shown in Figures 3 and 4 using the defaults of PAUP* for parsimony and likelihood. The model for likelihood used was HKY [12].

This is an under-parametrization of the base model, GTR, used by Dawg to generate the data. The use of this under-parametrization, as HKY is a more simplified model compared to GTR, was used because it allowed for a universal model for all data sets without having to calculate the optimum parameters or model for each data set. To further explore the effects of this assumption we reran the analysis using a base model of JC in Dawg to see if there was a change in the results. In this case the use of HKY in the likelihood evaluation would be an over-parametrization of the the data sets model. We found no particularly different results (data not shown).

B. Gap State

When using parsimony as a heuristic for likelihood the gap mode used for the parsimony score is critical for good correspondence. Treating the gaps in an alignment as a new state creates large variances in the parsimony scores, depending on the extent to which the gaps line up. One solution when using parsimony as a precursor to a maximum likelihood search to set the gap mode to missing to avoid this difficulty. Figures 5 and 6 show the difference in correspondence between the two methods. Both figures are a plot of an exhaustive search

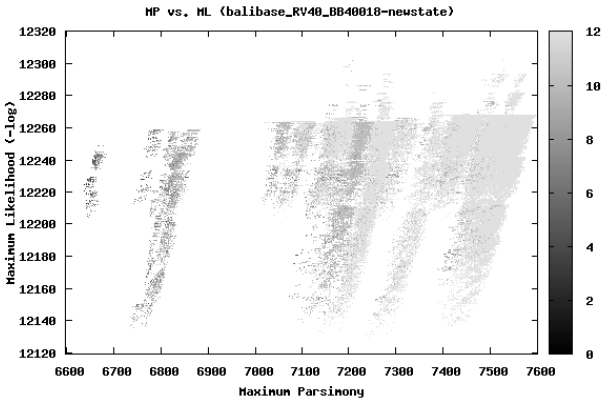


Fig. 6. All possible trees for BALiBASE dataset RV40 BB40010 scored with parsimony (GapMode = new state) and maximum likelihood (HKY under PAUP* defaults). Color indicates RF distance from optimal parsimony tree. Note that the optimal tree under likelihood is a large RF distance from the optimal parsimony tree.

through one of the BAliBASE [27] data sets (RV40 BB40010). Every possible topology was scored by both parsimony and maximum likelihood, and then plotted. The color of each point indicates the topological distance, (measured using Robinson Foulds [21]), from the most parsimonious topology. In both cases maximum likelihood was scored using the HKY method. These results are representative of the other data sets.

C. Correlation Between The Most Likely and The Most Parsimonious Topology

While the correspondence between MP and ML methods exists it does not preserve a partial order between topologies.

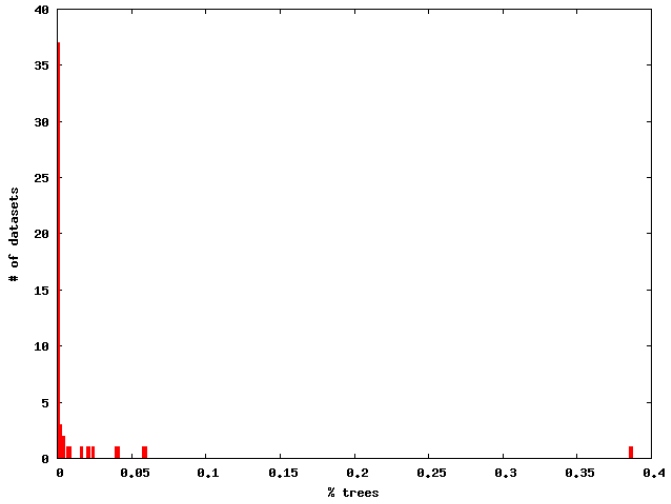


Fig. 7. Percentage of trees with a better likelihood score than the most parsimonious tree

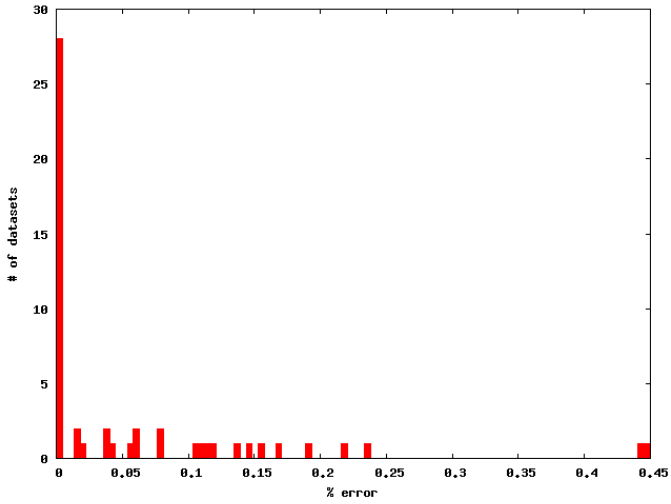


Fig. 8. Relative error in ML Score of MP tree

It is therefore not possible to directly substitute one method for the other. However it is possible to use parsimony to aid a maximum likelihood search. Figure 5 shows that as parsimony score decreases, the negative log maximum likelihood score generally decreases also. This trend also generalizes to the other data sets studied.

In all but two of the nearly 300 real data sets, less than 70 possible topologies had a better likelihood score than the most parsimonious topology for that dataset. It is clear that in the majority of cases, parsimony eliminates most of the incorrect topologies. Figure 7 is a histogram of data sets grouped by the percentage of trees with better likelihood scores than the best parsimony tree. Note that 37 datasets fell into the 0% trees bucket. For the 9 taxa case, from which these data sets have been drawn 0.05% corresponds to 67 trees.

Furthermore the likelihood score of the best parsimony tree is often very comparable to the likelihood score of the best maximum likelihood tree. Figure 8 shows the relative error of the best parsimony tree, taking the best maximum likelihood score as the true value.

IV. Parsimony Heuristic

Not only do likelihood and parsimony generally improve together, the most likely and the most parsimonious trees are often very close topologically. The major difficulty with parsimony methods is long branch attraction [1]. This problem causes two long branches to be incorrectly made siblings in the final topology. This topological change is however correctable with a very small number of TBR swaps. A simple heuristic that can be applied to many existing phylogenetic programs is to perform a parsimony search and use the result as the seed tree for a maximum likelihood search. This maximum likelihood search can be further limited to trees within a small TBR distance of the most parsimonious tree.

The idea of using parsimony to improve likelihood searches is not entirely new. RAxML uses a step-wise maximum parsimony tree as its starting tree [24]. This is a step in the right direction but performing a complete parsimony search further improves performance of this search.

```
begin paup;
  set criterion = parsimony;
  hsearch
  set criterion = likelihood;
  hsearch start=current swap=TBR
    rearrlimit=x;
  lscores all scorefile=<filename>;
  quit;
end;
```

Fig. 9. PAUP* block to perform heuristic search

A. Improving PAUP*

Our heuristic is very easily implemented in PAUP*. PAUP* already has the machinery to perform both parsimony and maximum likelihood searches. The heuristic can be expressed in a PAUP* block, as shown in Figure 9. As parsimony gives us a tree that is topologically close to the desired tree, we limit the length of the maximum likelihood search, gaining further savings in time. The x used as the $rearrlimit$ was set to the number of trees within one TBR swap.

We achieved identical scores on the majority of data sets using this method. Figure 10 shows a histogram of the differences in the scores of the final tree. It is not statistically significant that occasionally the heuristic outperforms the standard PAUP* search, nor is it significant that the standard PAUP* search occasionally produces a better score than the heuristic.

The use of the heuristic does improve search times significantly. As shown in Figure 11, the run time was on average improved by 25%. In a handful of the data sets, parsimony was not a good indicator of likelihood and the search took longer using the heuristic than it would have otherwise. All

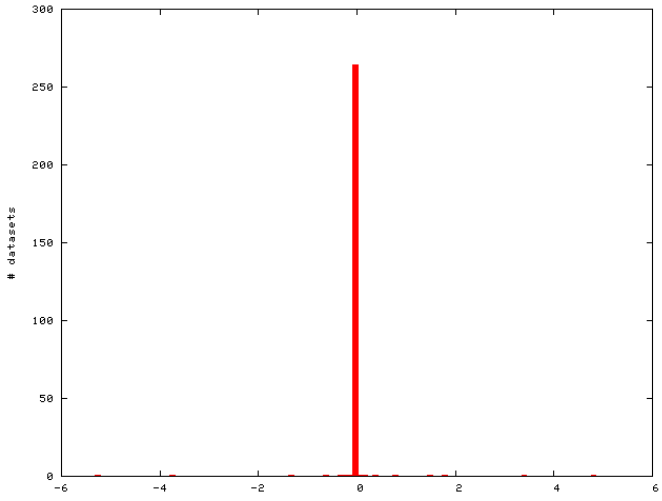


Fig. 10. Score differences between PAUP* with and without the Parsimony Heuristic

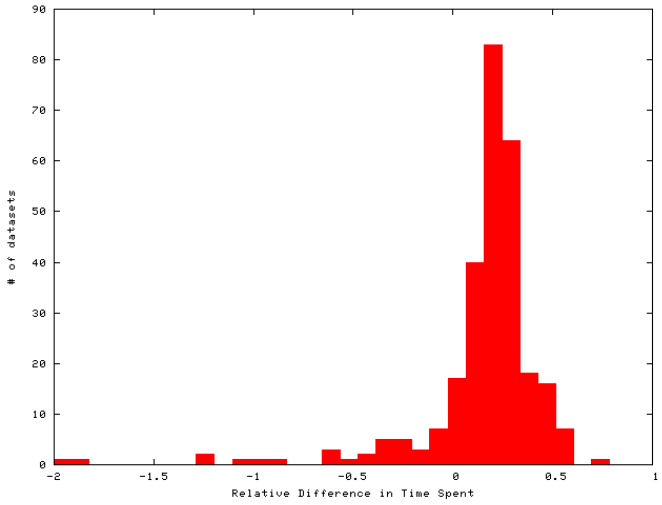


Fig. 11. Running time improvement from starting with a parsimony search in PAUP*

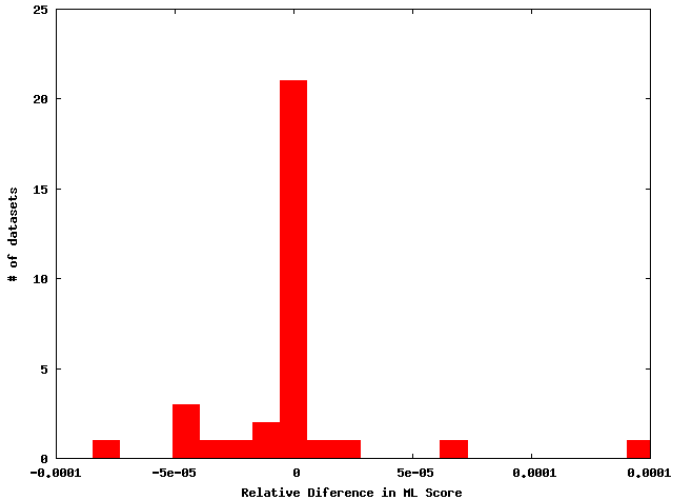


Fig. 12. Score differences between RAxML with and without using TNT

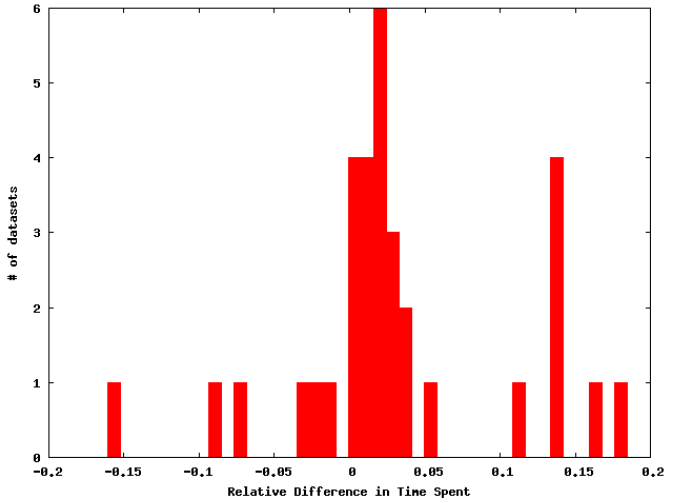


Fig. 13. Execute time improvements using RAxML and TNT versus just RAxML

times are the average of five runs.

B. Improving RAxML

Due to the size of the large data sets, exhaustive searching is not feasible. RAxML is a phylogenetic search program that can perform maximum likelihood searches on large data sets [24]. One of the heuristics it uses is to start the ML searching with a stepwise maximum parsimony tree. We took a slightly different approach to applying our heuristic to RAxML. First, we heuristically searched under MP, and then started the ML search in RAxML with the most parsimonious tree found. To search with MP, we used TNT, Tree Analysis Using New Technology [11], [18]. In other analysis that we've performed, TNT out performs PAUP* in terms of time and parsimony score [4]. We used two TNT search strategies per each date set: 1) sectorial-search and xmult and 2) TBR. The most parsimonious tree found with TNT was used to start a GTR-GAMMA search with RAxML. The relative difference in scores is shown in Figure 12. Again the scores are not statistically different, but there is more variability than with the PAUP* scores.

The execution time improvements are shown in Figure 13. There are two circumstances which are probably causing this reduction in the performance of the heuristic. First, RAxML by default starts its search with a step-wise maximum parsimony tree. This tree already has a reasonable parsimony score, so there is not as much benefit to be gained by finding a more parsimonious tree. Another difference is that RAxML does not allow us to limit the time spent in the likelihood search, which does not allow us to take full advantage of the small topological differences between the maximum parsimony and maximum likelihood trees.

V. Conclusions

Parsimony is correlated with likelihood in phylogenetic inference. In many data sets the predicted topologies are very close, if not identical. Maximum likelihood searches can be boosted using our parsimony heuristic in a manner that avoids the problems associated with the Felsenstein Zone and parsimony while maintaining some of the speed of a parsimony search.

References

[1] J. Bergsten, "A review of long-branch attraction," *Cladistics*, vol. 21, no. 2, pp. 163–193, 2005.

[2] J. Camin and R. Sokal, "A Method for Deducing Branching Sequences in Phylogeny," *Evolution*, vol. 19, no. 3, pp. 311–326, 1965.

[3] H. Carroll, W. Beckstead, T. O'Connor, M. Ebbert, M. Clement, Q. Snell, and D. McClellan, "DNA Reference Alignment Benchmarks Based on Tertiary Structure of Encoded Proteins," in Press with Bioinformatics.

[4] H. Carroll, M. Clement, Q. Snell, and K. Crandall, "Phylogenetic analysis of large sequence data sets," in *Proceedings of the Second Biotechnology and Bioinformatics Symposium*, 2005, pp. 20–24.

[5] R. Cartwright, "DNA assembly with gaps (Dawg): simulating sequence evolution," *Bioinformatics*, vol. 21, no. 3, pp. 31–38, 2005.

[6] B. Chor and T. Tuller, "Maximum likelihood of evolutionary trees is hard," *Proceedings of the 9th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2005*, vol. 3500, pp. 296–310, 2005.

[7] W. Day, D. Johnson, and D. Sankoff, "The computational complexity of inferring rooted phylogenies by parsimony," *Mathematical Biosciences*, vol. 81, no. 33-42, p. 299, 1986.

[8] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, mar 2004.

[9] J. Felsenstein, "Cases in which Parsimony or Compatibility Methods Will be Positively Misleading," *Systematic Zoology*, vol. 27, no. 4, pp. 401–410, 1978.

[10] ——, "Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach," *Journal of Molecular Evolution*, vol. 17, no. 6, pp. 368–376, 1981.

[11] P. Goloboff, S. Farris, and K. Nixon, "TNT: Tree analysis using new technology," http://www.cladistics.com/webtnt.html, 2001.

[12] M. Hasegawa, H. Kishino, and T. Yano, "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA," *Journal of Molecular Evolution*, vol. 22, no. 2, pp. 160–174, 1985.

[13] J. Huelsenbeck, "Performance of Phylogenetic Methods in Simulation," *Systematic Biology*, vol. 44, no. 1, pp. 17–48, 1995.

[14] J. Huelsenbeck and D. Hillis, "Success of Phylogenetic Methods in the Four-Taxon Case," *Systematic Biology*, vol. 42, no. 3, pp. 247–264, 1993.

[15] T. Jukes and C. Cantor, "Evolution of protein molecules," *Mammalian Protein Metabolism*, vol. 3, pp. 21–132, 1969.

[16] M. Kuhner and J. Felsenstein, "A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates [published erratum appears in Mol Biol Evol 1995 May; 12 (3): 525]," *Mol Biol Evol*, vol. 11, no. 3, pp. 459–468, 1994.

[17] C. Lanave, G. Preparata, C. Sacone, and G. Serio, "A new method for calculating evolutionary substitution rates," *Journal of Molecular Evolution*, vol. 20, no. 1, pp. 86–93, 1984.

[18] R. Meier and F. Ali, "Software Review. The newest kid on the parsimony block: TNT (Tree analysis using new technology) ," *Systematic Entomology*, vol. 30, no. 1, p. 179, 2005.

[19] C. Ponting, J. Schultz, F. Milpetz, and P. Bork, "SMART: identification and annotation of domains from signalling and extracellular protein sequences," *Nucleic Acids Research*, vol. 27, no. 1, pp. 229–232, 1999.

[20] G. P. S. Raghava, S. M. J. Searle, P. C. Audley, J. D. Barber, and G. J. Barton, "OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy," *BMC Bioinformatics*, vol. 4, p. 47, 2003.

[21] D. Robinson and L. Foulds, "Comparison of Phylogenetic Trees," *Mathematical Biosciences*, vol. 53, no. 1, pp. 131–147, 1981.

[22] F. Rodriguez, J. Oliver, A. Marin, and J. Medina, "The general stochastic model of nucleotide substitution." *J Theor Biol*, vol. 142, no. 4, pp. 485–501, 1990.

[23] M. Siddall, "Success of Parsimony in the Four-Taxon Case: Long-Branch Repulsion by Likelihood in the Farris Zone," *Cladistics*, vol. 14, no. 3, pp. 209–220, 1998.

[24] A. Stamatakis, "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models," *Bioinformatics*, vol. 22, no. 21, p. 2688, 2006.

[25] D. L. Swofford, *PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4*. Sunderland, Massachusetts: Sinauer Associates, 2003.

[26] S. Tavare, "Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math," *Life Sci*, vol. 17, pp. 57–86, 1986.

[27] J. D. Thompson, P. Koehl, R. Ripp, and O. Poch, "BALiBASE 3.0 latest developments of the multiple sequence alignmnet benchmark," *Proteins: Structure, Function, and Bioinformatics*, vol. 61, no. 1, pp. 127–136, 2005.