# Phylogenetic Bootstrapping under Resource Constraints: Higher Model Accuracy or more Replicates?

Alexandros Stamatakis[1*] and Vincent Rousset[2]

[1] Department of Computer Science, Technical University of Munich, Bolzmannstr. 3, 85747, Garching b. München, Germany
[2]Department of Biology, University of California, Riverside, Riverside, California, 92521, USA
*Correspondence to: stamatak@cs.tum.edu

**Motivation:** Bootstrapping is a standard method to infer confidence values on phylogenetic trees. Given the rapid growth of current input datasets that is driven by advances in wet-lab techniques as well as the high energy consumption and cost of computational resources we address the following question: How can a limited amount of computational resources best be used to infer the most accurate relative bootstrap support values under resource constraints. In particular, we address the question whether more computing time should be invested into optimizing per-bootstrap replicate Maximum Likelihood model parameters or if one should compute more replicates at the expense of lower model accuracy. Our computational experiments with the RAxML and GARLI algorithms indicate that it is better to invest more time into perreplicate model parameter optimization at the expense of computing less replicates, i.e., the computation of more, superficially optimized replicates, does not yield advantages.

**Introduction:** The significant progress in DNA sequencing technology over the last years poses new challenges for phylogenetic analyses, since it provides the possibility to build and analyze significantly larger data sets that incorporate more sequences and/or more taxa. An important observation within this context, is that the pace of molecular data accumulation is significantly higher than the pace at which hardware architectures become faster, i.e., advances in sequencing techniques have outpaced Moore's law. Figure 1 provides the relative growth of data in Genebank compared to the transistor count in hardware architectures from 1982-2005 (note the logscale on the y-axis). We call this phenomenon the “Bio-Gap”.

With the emerging discipline of phylogenomics (see Delsuc *et al.* (2005) for a review) and the growing popularity of Expressed Sequence Tags (ESTs) for phylogeny reconstruction (Jeffroy *et al.*, 2006), there is an increasing need to develop more efficient algorithms for tree-building, especially for time- as well as memory-intensive model-based methods such as Maximum Likelihood (Felsenstein, 1981) or closely related Bayesian methods. Because of the growing popularity of multi-gene alignments, multi-core and supercomputer architectures will be deployed more frequently to conduct large-scale real-world phylogenetic analyses (see for example Dunn *et al.*, 2008). Access to “classic” supercomputers such as the IBM BlueGene/L or BlueGene/P systems is typically granted in terms of CPU hours, i.e., a limited amount of time and resources is available to carry out an analysis. Moreover, many supercomputing centers currently face serious problems because of high energy costs, that in some cases even lead to the shutdown of relatively new and powerful facilities. This recent development has lead to a new research area in high performance computing called power-aware computing. Following this trend, the so-called *green500* list (http://www.green500.org/) that covers the 500 most energy efficient supercomputers in the world was introduced. This list only partially overlaps with the “classic” *top500* list (http://www.top500.org/) that comprises the most powerful supercomputers worldwide. Thus, taking into account this trend in computer architectures and energy prices, combined with the current molecular data explosion and rapidly growing alignment sizes we address the following question: Given a certain amount of time- or cost-constrained computational resources (CPU hours), how can those limited resources best be used for accurate phylogeny reconstruction? We intend to determine the trade-off, by means of the relative topological accuracy, between investing resources into optimization of per-bootstrap (bootstrapping: see below) replicate ML model parameters at the expense of computing less replicates versus superficial model parameter optimization for the sake of computing more replicates.

The general Bootstrapping procedure is a well-established computer-based statistical method to obtain non-parametric error estimates (Efron, 1979), by inferring the variability in an unknown distribution from which the data (bootstrap replicates) was drawn by re-sampling from the original data. The seminal paper by Joe Felsenstein (Felsenstein, 1985), introduced the application of the bootstrap method to phylogenetic inference. The non-parametric phylogenetic Bootstrap (BS) method proceeds by randomly re-sampling the characters (columns) from the original matrix with replacement by creating a respective pseudo-replicate matrix that contains as many columns as the original alignment but has a slightly different column composition. This re-sampled (bootstrapped) alignment is then used as input for a Maximum Likelihood (ML, Felsenstein, 1981) tree search algorithm like GARLI or RAxML. Once ML trees for all 100 or 1,000 replicates have been computed, the frequency of appearance of a particular group (bipartition) of taxa among all the bootstrapped trees corresponds to the bootstrap confidence limit or simply BS value and can then be used to assess the relative stability of the respective phylogenetic group and stability of the overall tree under slight alterations (re-samplings) of the original input alignment. To summarize and visualize the results of such a BS analysis the bootstrapped tree topologies are either used to compute various flavors of consensus trees: strict, majority rule, or extended (bifurcating) majority rule trees. Alternatively, the bootstrapped trees can be used to draw support values on the best-scoring (best-known) ML tree on the original alignment. The phylogenetic BS is probably the most widely and commonly used approach to assess confidence on phylogenetic trees.

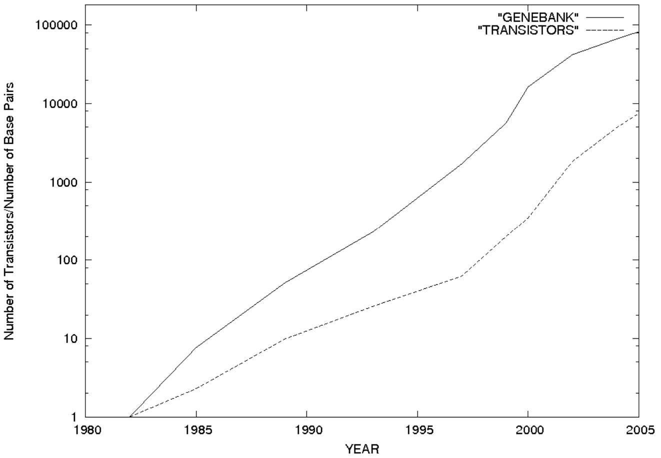

*Figure 1: The "Bio-Gap": Relative growth of processor speeds and sequence data in GeneBank 1982-2005; Molecular data growth has overtaken Moore's law. Note the logscale on the y-axis.*

To date, Joe Felsenstein's seminal paper on the phylogenetic BS (Felsenstein, 1985) has been cited over 11,000 times in the literature (*ISI Web of Knowledge*). Despite the significant progress in the development of heuristic ML search algorithms over the last years (see Morisson (2007) for a review), the bootstrap analyses still represent *the* major computational bottleneck in real-world phylogenomic studies under ML (Stamatakis *et al*., 2008; Morisson, 2007) and thus require a high amount of computational resources. These high resource requirements for inference of BS values, particularly on large memory-intensive phylogenomic datasets, have so far been a limiting factor in phylogeny reconstruction, in particular under the ML model (see McMahon and Sanderson 2006). Phylogenetic inference under ML for a single BS replicate/input alignment requires the estimation and optimization of:

1. the substitution model parameters
2. the branch lengths
3. the tree topology

In the standard phylogenetic BS procedure the above three computational steps have to be repeated for every BS replicate, i.e., the tree for every replicate is computed "de novo", without making use of any topological or model parameter information from ML searches on preceding BS replicates. Recently, Stamatakis *et al.* (2008) have partially solved the computational problem associated with bootstrapping by developing a novel rapid BS search algorithm (RBS algorithm) that uses the following two approximations to improve inference times by more than one order of magnitude while returning qualitatively comparable support values: *Firstly*, the replacement of the model parameter optimization procedure for each replicate by a *one-time* model parameter optimization on the original alignment and a reasonable, i.e., non-random starting tree, and, *secondly*, by designing a "quick & dirty" search algorithm to accelerate per-replicate BS tree searches. This "quick & dirty" algorithm also makes use of the topological information collected during preceding replicates (see Stamatakis *et al.*, 2008 for details). Based on this fast RBS algorithm, here, we assess the trade-offs between speed and accuracy with respect to the first approximation only. We compare support values obtained from standard RBS searches (henceforth denoted as 1-RBS) that use the aforementioned one-time only model parameter optimization on the original alignment to modified RBS searches that use a more compute-intensive per-replicate optimization of model parameters (denoted as N-RBS, where N is the number of replicates). Our objective is to experimentally determine if, given a certain CPU time limit T, more replicates with 1-RBS or less replicates using N-RBS yield better relative accuracy with respect to a large number of 500 N-RBS replicates. Our findings, that are based on a large benchmark set of 18 real-world alignments, indicate that investing more time into per-replicate model parameter optimization yields slightly (RAxML) to significantly (GARLI) more accurate results than the execution of more replicates without thorough model parameter re-optimization.

**Experimental Setup:** We conducted computational experiments on 18 single- and multi-gene real world alignments (17 DNA alignments and 1 Protein alignment) comprising 8 up to 2,000 taxa using an appropriately modified version of RAxML (Stamatakis, 2006b, see below). To ensure that our results are not biased by the specific RBS search algorithm and model parameter optimization strategy implemented in RAxML, we also performed 3 experiments with GARLI (Zwickl, 2006). GARLI allows to disable model parameter optimization by reading in user-specified model parameters instead. As model parameters for GARLI we used ML model parameters that were estimated with RAxML on the original alignment and an MP starting tree. Moreover, the GARLI search algorithm is equally powerful with respect to finding bestscoring trees as the RAxML search algorithm (Stamatakis, 2006b), but significantly slower by 1-2 orders of magnitude (Stamatakis 2006b, Stamatakis *et al.*, 2008). Therefore, we only conducted 3 GARLI reference runs on datasets d150, d218, d354 (see below) with 500 replicates each. We call the resulting collections of BS trees with fixed model parameters 1-GARLI and with on-the-fly model parameter optimization N-GARLI, accordingly.
The 18 test alignments are diverse in terms of organisms (green plants, acer, mammals, bacteria, archaea, fungi, Pappilomaviruses), genes (protein, mitochondrial, ribosomal, and non-coding genes), the number of taxa (from 8 to 2,000), and the number of concatenated genes in a single matrix (up to 106 genes for d8_M). Data set sizes are provided in Table 1 and are referred to in the text by dXYZ, where XYZ is the respective number of taxa. Multi-gene datasets for which we executed partitioned analyses are denoted by _M and the protein protein datasets by _AA.

| Dataset | #bp | Dataset | #bp |
|---|---|---|---|
| d8_M | 127,026 | d354 | 460 |
| d53 | 7,542 | d404_M | 13,158 |
| d59_M | 6,951 | d500 | 1,398 |
| d81 | 4,552 | d628 | 1,228 |
| d125_M | 29,149 | d714 | 1,241 |
| d140_AA_M | 1,104 | d855 | 1,436 |
| d150 | 1,269 | d1604 | 1,276 |
| d217_M | 3,665 | d1908 | 1,424 |
| d218 | 2,294 | d2000 | 1,251 |

*Table 1: Dataset sizes in the benchmark set.*

Computational experiments were conducted at the CIPRES (http://www.phylo.org) project cluster located at the San Diego Supercomputer Center that is equipped with 16 8-way AMD 2.4 GHz Opteron shared-memory nodes. For each data set, we executed two RBS analyses with 500 replicates each: One analysis using a per-replicate optimization of ML model parameters (N-RBS) and a second analysis using the one-time model parameter optimization on the original alignment that are optimized on a Maximum Parsimony starting tree (1-RBS). All 1-RBS and N-RBS analyses on DNA data sets were performed under the GTR+Γ model (General Time-Reversible model of nucleotide substitution (Tavare, 1986) with the Γ model of rate heterogeneity (Yang, 1994)), which is among the most widely used models for phylogenetic analyses of DNA (Ripplinger and Sullivan, 2008). Analyses on the 140 taxon protein dataset (d140_AA_M) were conducted under WAG+Γ (Whelan and Goldman, 2001), a widely used model of amino acid substitution. In addition, we conducted searches for the respective best-scoring ML trees on all original alignments.

### *Adaptation of RAxML RBS Algorithm*

The RAxML RBS algorithm was adapted as follows with respect to the standard publicly available code described in Stamatakis *et al.* (2008): The restriction that only the GTR+CAT approximation of rate heterogeneity (Stamatakis, 2006a) can be used in combination with the “quick & dirty” RBS tree search algorithm was removed to allow usage of GTR+Γ. In addition, we changed the command line interface parameters for invoking RBS to run 1-RBS (`-f x`) and N-RBS (`-f X`) analyses (for details please refer to the RAxML manual at http://icwww.epfl.ch/~stamatak/. The modified code as well as all test datasets and result files are available for download at the following address
http://wwwbode.in.tum.de/~stamatak/HSCBB2008.tar.bz2.

### *Result Analysis*

Experimental results were analyzed as follows: For each N-RBS run we determined the number of replicates that can be compute within the time taken by the significantly faster (average speedup 2.63) 500 1-RBS replicates, i.e., we used the execution time of the 500 1-RBS replicates on each dataset as a time constraint. We denote this reduced number of N-RBS replicates as N-RBS | T, i.e., N-RBS constrained by CPU time T. We then computed various statistics between the N-RBS | T trees, the 500 1-RBS trees, and the full N-RBS trees. We use the 500 N-RBS trees as reference because they represent the statistically more correct, but slower, standard approach to bootstrapping. For all sets of replicates we computed the relative Robinson-Foulds distance (RF, Robinson and Foulds, 1981) with treedist from PHYLIP as well as the weighted Robinson-Foulds distance (WRF, Robinson and Foulds, 1979) with a script by Olaf Bininda-Emonds called partitionMetric.pl (available at http://www.unioldenburg.de/molekularesystematik/33997.html) on the extended majority rule (bifurcating/binary) consensus trees obtained by applying the `consense` program from the PHYLIP package. The unweighted RF distance between two tree topologies counts the number of subtrees found in one tree or the other, but not both. The WRF distance uses the support value information on the respective subtrees instead, i.e., a subtree with a support of 0.5 counts 0.5 instead of 1 as for RF. The WRF distance thus allows to take support values on consensus trees into account and penalizes differently placed subtrees with low support to a lesser extent than differently placed subtrees with high support. If the RF distance for a pair of trees with support values is significantly larger than the respective WRF distances, this means that the topological differences are mainly due to differently placed subtrees with low support, whereas when RF ≈ WRF this means that subtrees with high support are placed differently in the two trees under comparison. In addition, we computed the Pearson correlation coefficient ρ between support values obtained via 1-RBS and N-RBS, as well as N-RBS | T and N-RBS drawn on the respective best-scoring ML trees. We also computed the intercept and slope of the respective linear regression function. We mostly focus on the results obtained via RF and WRF since those topological distances appear to be more sensitive than the Pearson coefficient to slight changes/differences in the collections of replicates and better allow to discriminate between the approaches (see also Stamatakis *et al.*, 2008). For the experiments with GARLI we computed the RF as well as WRF distances between 500 1-GARLI runs and 500 N-GARLI runs. We also computed the topological distances between 250 N-GARLI runs, i.e., N-GARLI | T and 500 N-GARLI runs. Here we chose a fixed value of 250, because unlike in RAxML, the run time differences between 1-GARLI and N-GARLI runs are insignificant, the run time variation lies between 0.97 and 1.17. This insignificant execution time improvement is due to the genetic search algorithm that is used in GARLI (Zwickl, 2006). The omission of model parameter optimization does not necessarily mean that the algorithm will execute less generations, i.e., converge faster, in particular because the search space becomes less smooth without model parameter optimization. In the RAxML RBS algorithm the run time improvements are more prevalent, because the number of iterations of the search algorithm is fixed a priori (Stamatakis *et al.*, 2008).

**Results:** We provide the speedup (denoted as acc) of 500 1-RBS replicates over 500 N-RBS replicates in Table 2. We also indicate the number of replicates (#reps) in the time-constrained N-RBS | T searches.

| Dataset | acc | #reps | Dataset | acc | #reps |
|---|---|---|---|---|---|
| d8_M | 13.18 | 38 | d354 | 2.34 | 213 |
| d53 | 3.29 | 152 | d404_M | 1.44 | 348 |
| d59_M | 3.06 | 163 | d500 | 1.81 | 277 |
| d81 | 1.88 | 266 | d628 | 1.65 | 302 |
| d125_M | 3.99 | 125 | d714 | 1.92 | 261 |
| d140_AA_M | 1.06 | 472 | d855 | 1.55 | 322 |
| d150 | 1.89 | 265 | d1604 | 1.61 | 310 |
| d217_M | 1.85 | 270 | d1908 | 1.65 | 302 |
| d218 | 1.84 | 272 | d2000 | 1.27 | 394 |

*Table 2: Speedup of 1-RBS over N-RBS and number of replicates that N-RBS can conduct within the time required for 500 1-RBS replicates.*

The average number of replicates is 264 while the average speedup amounts to 2.63. In Table 3 we indicate the topological RF as well as WRF distances in percent between 500 1-RBS replicates and the 500 N-RBS replicates. The average relative RF is 6.51% and the average WRF is 2.47%.

| Dataset | RF | WRF | Dataset | RF | WRF |
|---|---|---|---|---|---|
| d8_M | 0 | 0 | d354 | 23.9 | 6.9 |
| d53 | 0 | 0 | d404_M | 13.2 | 5.0 |
| d59_M | 3.5 | 2.6 | d500 | 9.6 | 4.7 |
| d81 | 7.6 | 0.7 | d628 | 6.7 | 3.4 |
| d125_M | 0 | 0 | d714 | 6.2 | 2.0 |
| d140_AA_M | 2.9 | 2.5 | d855 | 7.4 | 3.4 |
| d150 | 0 | 0 | d1604 | 11.6 | 4.6 |
| d217_M | 5.6 | 2.4 | d1908 | 6.5 | 2.4 |
| d218 | 2.7 | 1.4 | d2000 | 9.4 | 1.7 |

*Table 3: Relative RF and WRF distances in % between consensus trees induced by 500 1-RBS and 500 N-RBS replicates.*

Finally, in Table 4 we indicated relative RF and WRF topological distances in % between varying numbers of NRBS | T replicates (see Table 2) and 500 N-RBS replicates. The average RF amounts to 4.60% and the WRF to 1.98%. The average Pearson correlation coefficient (data not shown) between 500 1-RBS and 500 N-RBS support values drawn on the respective best-scoring ML trees is 0.999 (min: 0.998, max: 1.0), the average slope of the linear regression function is 0.998 (min: 0.984, max: 1.009) and the average absolute offset amounts to 0.34% (max: 1.48%) . The average Pearson correlation coefficient between time constrained N-RBS | T and 500 N-RBS replicates on the best-scoring ML tree is 0.998 (min: 0.995, max: 1.0) with an average slope of 1.001 (min: 0.97, max: 1.012) and an average absolute offset of 0.56% (max: 3.09%).

| Dataset | RF | WRF | Dataset | RF | WRF |
|---|---|---|---|---|---|
| d8_M | 0 | 0 | d354 | 9.1 | 4.2 |
| d53 | 0 | 0 | d404_M | 13.4 | 2.2 |
| d59_M | 0 | 0 | d500 | 5.6 | 3.0 |
| d81 | 0 | 0 | d628 | 3.8 | 2.2 |
| d125_M | 3.2 | 0.5 | d714 | 6.8 | 3.1 |
| d140_AA_M | 0 | 0 | d855 | 8.3 | 3.7 |
| d150 | 2.7 | 2.4 | d1604 | 4.7 | 3.2 |
| d217_M | 0.93 | 0.9 | d1908 | 11.0 | 5.0 |
| d218 | 6.5 | 3.2 | d2000 | 4.6 | 2.2 |

*Table 4: Relative RF and WRF distances in % between consensus trees induced by varying numbers of N-RBS | T and 500 N-RBS replicates.*

### *GARLI Results*

The results obtained by GARLI generally show a similar, though stronger tendency as the results obtained by RAxML. The relative RF/WRF topological distances between 1-GARLI and N-GARLI are 10.8%/4.4% (d150), 25.1%/12.8% (d218), and 55.8%/14.9% (d354). The RF/WRF distances between 250 N-GARLI | T replicates and 500 N-GARLI replicates are 5.4%/1.6% (d150), 8.3%/2.4% (d218), and 34.47%/5.5%(d354) respectively.

**Discussion:** In general our results indicate that the differences between the relative accuracy of 1-RBS bootstrap replicates and time-constrained NRBS | T replicates with respect to a reference set of 500 N-RBS replicates are not very large. This indicates that the model parameter approximation used in the original RBS algorithm, which uses the 1-RBS strategy only has an insignificant impact on the support value distribution and shape of the extended majority rule consensus trees. The average speedup achieved by omitting per-replicate model parameter optimization is 2.63. However, if the very large speedup on the 8-taxon data set is not included in the calculation, 1-RBS is only two times faster than N-RBS. The large speedup on the 8-taxon dataset is due to the low number of taxa and hence small search space in combination with the very long sequences (127,026 bp). Hence, in contrast to other datasets, the largest amount of CPU time is required to optimize model parameters and not for the tree search. The fact that virtually no speedup is observed on the protein alignment (d140_AA_M) is due to the fact that a fixed model of amino acid substitution is used and hence less parameters need to be optimized. Whereas the Pearson correlation on the best-scoring ML trees only yields insignificant differences between 1-RBS and N-RBS | T, the RF as well as WRF distances on the extended majority rule consensus trees exhibit larger discrepancies. As mentioned before, RF and WRF exhibit a higher degree of sensitivity for the comparison of sets of bootstrapped trees than the Pearson correlation. The N-RBSI T approach shows smaller average RF and WRF values than 1-RBS, in particular on partitioned multigene analyses (denoted by _M). The results also indicate that the effects of Bootstrapping with *one-time* model parameter optimization regarding the relative accuracy and the run time improvements are highly algorithm-specific, since the omission of model parameter optimization in GARLI does not yield any speedup. Besides, unlike for RBS, the relative accuracy of 500 1-GARLI bootstrap analyses is significantly worse than for 250 N-GARLI | T analyses. The main reason for this phenomenon is that the model parameter optimization, branch length optimization, and tree search mechanisms are more strongly interleaved and interdependent in GARLI than in RAxML.

**Conclusion:** Our results support that computing more replicates at the expense of a more superficial per-replicate model parameter optimization is costly in terms of relative accuracy and does not constitute a good alternative to the standard bootstrapping method. With respect to GARLI, omission of per-replicate model parameter optimization actually significantly decreases relative accuracy. We thus conclude that when only a limited amount of computational resources is available, it should be used to infer less replicates with higher per-replicate model accuracy. In addition, the biologically more meaningful average WRF distance is lower for this approach and at the same time it represents the statistically less debatable approach. Moreover, except for DNA datasets with few taxa and many genes (d8_M, d53, d59_M, d125_M) the inference times do not increase dramatically and the average WRF for these datasets is significantly lower for N-RBS | T. The results obtained with GARLI also indicate that speedups induced by omission of the model parameter optimization procedure are highly algorithm-specific. Hence, the computation of more superficially optimized replicates does not seem to yield any substantial advantages. Future work will focus on devising a fast per-replicate model optimization procedure in the RBS algorithm.

**Acknowledgments:** The authors wish to thank Markus Göker, Guido Grimm, Nicolas Salamin, Chuck Robertson, Nikos Poulakakis, and Marc Gottschling for providing real-world test datasets for this study. AS is funded under the auspices of the Emmy-Noether program by the German Science Foundation (DFG).

### References

Delsuc, F., Brinkmann, H., Philippe, H., *et al.* (2005) Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Gen.*, **6**(5):361-375.

Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M. , Edgecombe, G.D., *et al.* (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**(7188):745-749.

Efron, B. (1979) Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, **7**:1-25.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.*, **17**(6):368-376.

Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**(4):783-791.

Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H. (2006) Phylogenomics: the beginning of incongruence? *Trends in Genetics*, **22**(4):225-231.

Morrison, D.A. (2007) Increasing the Efficiency of Searches for the Maximum Likelihood Tree in a Phylogenetic Analysis of up to 150 Nucleotide Sequences. *Syst. Biol.*, **56**(6):988-1010.

McMahon, M.M., Sanderson M.J. (2006) Phylogenetic Supermatrix Analysis of GenBank Sequences from 2228 Papilionoid Legumes. *Syst. Biol.*, **55**(5):818-836.

Ripplinger, J., Sullivan, J. (2008) Does Choice In Model Selection Affect Maximum Likelihood Analyses? *Syst. Biol.*, **57**(1):76-85.

Robinson, D.F., Foulds, L.R. (1979) Comparison of weighted labelled trees. *Lecture Notes Math.*, **748**:119126.

Robinson, D.F, Foulds, L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**:1311-47.

Stamatakis, A. (2006a) Phylogenetic models of rate heterogeneity: A high performance computing perspective. *In Proceedings of IPDPS2006*.

Stamatakis, A. (2006b) RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. *Bioinformatics*, **22**(21):2688-2690.

Stamatakis, A., Hoover, P., Rougemont, J. (2008) A Rapid Bootsrap Algorithm for the RAxML WebServers. *Syst. Biol.* **, 75**(5): 758-771, 2008

Tavare, S. (1986) Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *DNA Sequence Analysis,* **17**:57-86.

Whelan, S., Goldman, N. (2001) A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a MaximumLikelihood Approach. *Mol. Biol. Evol.*, **18**:691-699

Yang, Z. (1994) Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites. *J. Mol. Evol.*, **39**:306-314

Zwickl, D. (2006) Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets Under the Maximum Likelihood Criterion. *PhD Thesis*, The University of Texas at Austin.