Benchmark datasets and software for developing and testing methods for large-scale multiple sequence alignment and phylogenetic inference

November 18, 2010 · Tree of Life
C. Randal Linder, Rahul Suri, Kevin Liu, Tandy Warnow

Abstract
We have assembled a collection of web pages that contain benchmark datasets and software tools to enable the evaluation of the accuracy and scalability of computational methods for estimating evolutionary relationships. They provide a resource to the scientific community for development of new alignment and tree inference methods on very difficult datasets. The datasets are intended to help address three problems: multiple sequence alignment, phylogeny estimation given aligned sequences, and supertree estimation. Datasets from our work include empirical datasets with carefully curated alignments suitable for testing alignment and phylogenetic methods for large-scale systematics studies. Links to other empirical datasets, lacking curated alignments, are also provided. We also include simulated datasets with properties typical of large-scale systematics studies, including high rates of substitutions and indels, and we include the true alignment and tree for each simulated dataset. Finally, we provide links to software tools for generating simulated datasets, and for evaluating the accuracy of alignments and trees estimated on these datasets. We welcome contributions to the benchmark datasets from other researchers.

Funding Statement
The research was supported by the US National Science Foundation DEB 0733029, and by Microsoft Research through support to TW.
One of the principal goals of the National Science Foundation’s Assembling the Tree of Life (AToL) initiative is “[a]ssembly of a framework phylogeny, or Tree of Life, for all major lineages of life.” [1] Much of that effort has focused on accumulating and analyzing data for the major taxonomic groups. However, because of the scale of the problems (numbers of species and amount of sequence information), the initiative has also required development of methods for sequence alignment, phylogenetic inference and supertree estimation that can handle hundreds, thousands or even tens of thousands of sequences. In the last decade, many new methods have been developed to address these challenging computational problems, including RAxML[2], GARLI [3], POY [4], SATé [5], and MrBayes [6]. However, evaluations of the efficacy of these methods for large-scale alignment and tree estimation—required for highly accurate estimations of the Tree of Life—have lagged behind method development.

To facilitate testing of large-scale alignment and phylogeny estimation methods, we have assembled a collection of web pages of (1) benchmark datasets and (2) software appropriate for creating new simulated benchmark datasets (http://www.cs.utexas.edu/users/phylo/datasets/). Because these datasets have been assembled with an eye to their usefulness for Tree of Life-scale projects, only datasets that have large numbers of taxa and/or present other difficulties for phylogenetic reconstruction and alignment (e.g., high rates of substitution and insertions and deletions) are included. The datasets we provide range in numbers of taxa from a few hundred to more than 300,000 sequences. The datasets are broken down into sets most appropriate for three types of phylogenetic problems: phylogenetic estimation given aligned sequences, supertree estimation, and multiple sequence alignment. Some datasets are appropriate for more than one type of problem and therefore are referenced more than once. Reference information and links are provided for all published datasets.

Benchmarks for phylogenetic estimation
The benchmark datasets for phylogenetic estimation are both empirical and simulated. They have been used in large-scale systematics studies, and so present challenges for maximum likelihood, maximum parsimony and Bayesian estimation. A subset of the empirical datasets (Table 1) include curated alignments and reference trees (generated using RAxML version 7.0.4 [2]). Reference trees have been assessed by bootstrapping, with edges having less than 75% support contracted. The remaining empirical datasets lack curated alignments and reference trees, but are appropriate for assessing the ability of alignment and phylogenetic software to operate on large and/or difficult datasets. They can also be used to compare how well algorithms solve particular optimality criteria, e.g., maximum parsimony or maximum likelihood. The empirical datasets include
Most of the benchmark datasets for multiple sequence alignment are the same as those for phylogenetic estimation. Both
benchmarks for multiple sequence alignment vary in the number of sequences, rates of substitution, and sizes and rates of indels.

Simulated datasets (Table 2) were taken from three sources and include both amino acid and nucleic acid sequences of widely
varying numbers of sequences, rates of substitution and sizes and rates of indels.

### Table 1. Empirical datasets and their properties.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Gene</th>
<th>Taxonomic Range</th>
<th>Number of Taxa</th>
<th>Number of Characters</th>
<th>Percen Indels</th>
</tr>
</thead>
<tbody>
<tr>
<td>16S.B.ALL</td>
<td>16S rRNA</td>
<td>Bacteria</td>
<td>27,643</td>
<td>6,857</td>
<td>80.0</td>
</tr>
<tr>
<td>16S.T</td>
<td>16S rRNA</td>
<td>The three domains of life plus mitochondria and chloroplasts</td>
<td>7,350</td>
<td>11,856</td>
<td>87.4</td>
</tr>
<tr>
<td>16S.3</td>
<td>16S rRNA</td>
<td>The three domains of life</td>
<td>6,323</td>
<td>8,716</td>
<td>82.1</td>
</tr>
<tr>
<td>16S.M.aa_ag</td>
<td>16S rRNA</td>
<td>Mitochondria</td>
<td>1,028</td>
<td>4,907</td>
<td>82.6</td>
</tr>
<tr>
<td>16S.M</td>
<td>16S rRNA</td>
<td>Mitochondria</td>
<td>901</td>
<td>4,722</td>
<td>78.1</td>
</tr>
<tr>
<td>23S.M</td>
<td>23S rRNA</td>
<td>Mitochondria</td>
<td>278</td>
<td>10,738</td>
<td>83.7</td>
</tr>
<tr>
<td>23S.M.aa_ag</td>
<td>23S rRNA</td>
<td>Mitochondria</td>
<td>263</td>
<td>10,305</td>
<td>83.5</td>
</tr>
<tr>
<td>23S.E.aa_ag</td>
<td>23S rRNA</td>
<td>Eukaryotes nuclear</td>
<td>144</td>
<td>8,619</td>
<td>61.1</td>
</tr>
<tr>
<td>23S.E</td>
<td>23S rRNA</td>
<td>Eukaryotes</td>
<td>117</td>
<td>9,079</td>
<td>59.7</td>
</tr>
</tbody>
</table>

a Unless otherwise noted, all datasets in this table are taken from Cannone et al. [7]. Curated alignments were produced by Cani and secondary structure. The reference trees reported on our web site were generated using RAxML version 7.0.4. Complete run parameters and program commands are listed on the web site.
b The number of columns in the aligned dataset.
c [8]

### Table 2. Simulated sequence datasets and some of their properties.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source</th>
<th>Data Typea</th>
<th>Number of Taxa</th>
<th>Num Chare</th>
</tr>
</thead>
<tbody>
<tr>
<td>FastTree</td>
<td>Price et al. [9]</td>
<td>AA NA</td>
<td>250; 1,250; 5,000; 78,132</td>
<td>N/A</td>
</tr>
<tr>
<td>SATé</td>
<td>Liu et al. [5]</td>
<td>NA</td>
<td>100; 500; 1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>RNASim</td>
<td>kim.bio.upenn.edu/software/csd.shtml</td>
<td>NA (SSU rRNA)</td>
<td>128; 256; 512; 1,024; 2,048; 4,096; 8,192; 16,384; 1,000,000</td>
<td>1,542</td>
</tr>
</tbody>
</table>

a AA = amino acid; NA = nucleic acid
b The number of characters in the root sequence
c The software used to generate the datasets

### Benchmarks for multiple sequence alignment

Most of the benchmark datasets for multiple sequence alignment are the same as those for phylogenetic estimation. Both
empirical and simulated datasets are provided. Taken as a whole, these datasets have properties that are typical both of markers currently used in large-scale phylogeny estimation and of markers that have evolved under high rates of indels, and are thus extremely difficult to align.

In addition to the simulated datasets for phylogeny estimation, a simulated amino acid dataset is included [14], which has sequences that were generated using Rose [11] and ranging from 20 to 100 taxa. Several additional empirical amino acid datasets are also included (BAliBASE [15], OXBench [16], PREFAB [17], SABmark [18]).

The empirical benchmark datasets for testing multiple sequence alignment include datasets with highly reliable curated sequence alignments that have been carefully validated by the community. The gold standard for this sort of dataset is Robin Gutell’s Comparative RNA Website (CRW) [19]. The curated alignments provided in CRW are based upon secondary structural information, which is particularly helpful where the mature rRNA is double stranded due to sequence complementarity.

**Benchmarks for supertree methods**

Finally, we provide benchmarks for testing supertree methods. As with the other benchmark collections, we provide both empirical (Table 3) and simulated [20] supertree datasets, and include datasets with different properties (such as the number of source trees, and the taxon sampling strategies used to produce the source trees).

**Table 3.** Empirical supertree datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Taxonomic Range</th>
<th>Total Taxa</th>
<th>Number of Source Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>McMahon and Sanderson [21]</td>
<td>Comprehensive papilionoid legumes</td>
<td>2,228</td>
<td>39</td>
</tr>
<tr>
<td>Cardillo et al. [22]</td>
<td>Marsupials</td>
<td>267</td>
<td>158</td>
</tr>
<tr>
<td>Beck et al. [23]</td>
<td>Placental mammals</td>
<td>116</td>
<td>726</td>
</tr>
<tr>
<td>Wojciechowski et al. [25]</td>
<td>Temperate herbaceous papilionoid legumes</td>
<td>558</td>
<td>19</td>
</tr>
</tbody>
</table>

**Software for generating datasets**

The collection of simulation software is useful for three aspects of producing simulated datasets: generating simulated phylogenetic trees (e.g., r8s [26] and Mesquite [27]); evolving sequences on phylogenetic trees (particularly tools that evolve sequences with both substitution and indel events) and tools for post-processing model trees in order to deviate from the model assumptions (such as the molecular clock).

**Conclusions**

We hope that having these benchmark datasets in a single location will facilitate research on large-scale phylogenetic methods and enhance the reproducibility of work based upon the datasets. We also hope that posting simulation tools that are capable of generating large-scale phylogenetic datasets will promote the generation of new benchmark datasets for use by the community.

We conceive of these pages as an evolving resource for the community and welcome input that would improve or expand them. We invite readers to contact us if they wish to contribute benchmark datasets or software to this resource or if they know of additional datasets or software that should be added to the pages. We will be happy to work with the laboratories providing them, and will either store them locally or provide links to their sites.

**Competing Interests**

The authors have declared that no competing interests exist.

**Acknowledgements**

The authors thank the editors and two anonymous reviewers for their very helpful suggestions on the manuscript. Correspondence should be sent to CRL. The research was supported by the US National Science Foundation DEB 0733029, and by Microsoft Research through support to TW.
References

1. National Science Foundation, Available from: REFERENCE LINK


3. Zwickl, D.J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. In Section of Integrative Biology, School of Biological Sciences, University of Texas at Austin: Austin.


REFERENCE LINK