Rapid Phylogeny Reconstruction with Support Values

Fabian Klötzl and Bernhard Haubold
kloetzl@evolbio.mpg.de
Max Planck Institute for Evolutionary Biology, Plön

MAX-PLANCK-GESELLSCHAFT

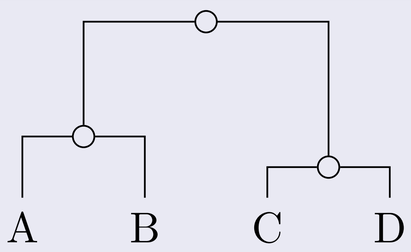
0. Sequence to Phylogeny

```
>Genome_A      >Genome_C
AAGGAAGTCT     CAGGAAGTCT
TGCCCTGGAA     TGCCCTGGAAA
>Genome_B      >Genome_D
AGGACGTCTT     AATGATGTCT
GCCCTCGGAA     GGCTCTGGAAA
```

$\Rightarrow$

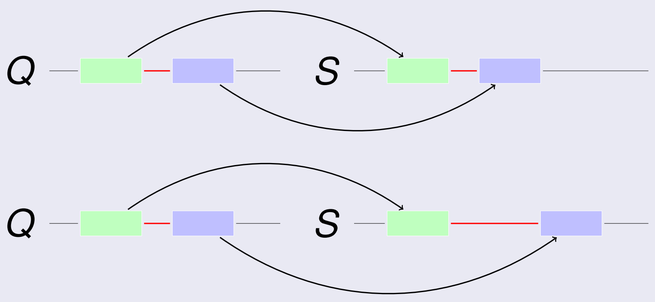
$$\begin{pmatrix} 0 & 0.1 & 0.25 & 0.3 \\ 0.1 & 0 & 0.3 & 0.3 \\ 0.25 & 0.3 & 0 & 0.05 \\ 0.3 & 0.3 & 0.05 & 0 \end{pmatrix}$$

$\Rightarrow$



1. Anchor Distances

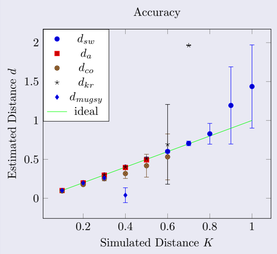
- Counting SNPs is the one of the most widely used measure of evolutionary distances available. Unfortunately, when comparing whole genomes, one has to first locate homologous sequences.
- To find homologous sequences, we first look for long and unique matches, termed *anchors*. Two equidistant anchors form a *pair*, surrounding a homologous sequence and thus allow counting SNPs.



- We use an enhanced suffix array to find anchors.

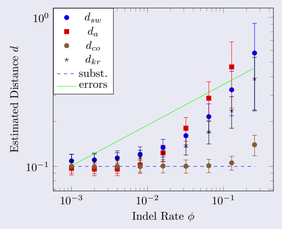
i	SA	LCP	$S^{SA[i]}$	lcp-intervals
0	4	−1	**A**AGG	
1	0	3	AAG**T**AAGG	3
2	5	1	A**G**G	
3	1	2	AG**T**AAGG	2
4	7	0	**G**	
5	6	1	G**G**	1
6	2	1	G**T**AAGG	
7	3	0	**T**AAGG	0
8		−1		

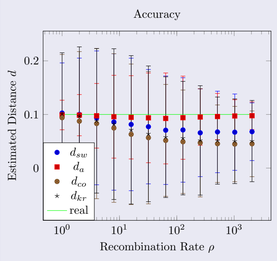
2. Results



We evaluate our implementation *andi* against other distance measures using simulated sequences (100 kbp). This first diagram shows the performance of each method as a function of different substitution rates (100 runs).

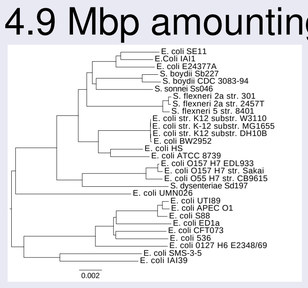
Here, the substitution rate was fixed at 0.1 but a variable number of indels were added. Substitution rate and total errors (SNPs + indels) are shown as lines.



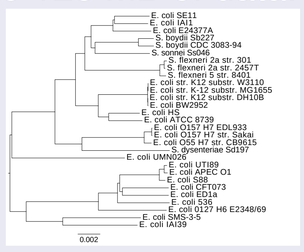


As a final test using simulated data we introduced recombination, as this leads to local variation in the substitution rate.

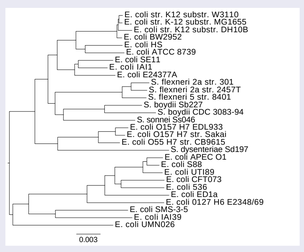
To evaluate the accuracy of the methods on real data, we chose a sample of 29 E. Coli and Shigella genomes (Eco29). On average the genomes have a length of 4.9 Mbp amounting to 128 MB of data.



andi (24s)



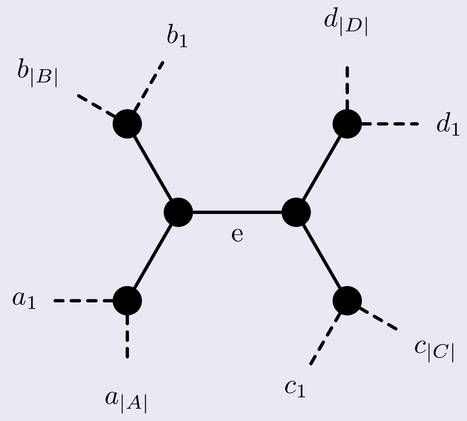
mugsy (2h 49min) alignment-based



spaced words (7min)

3. Alignment-Free Support Values

- Alignment-free methods can yield distances and hence distance-based phylogenies with surprising speed and accuracy. But without a multiple sequence alignment (MSA), bootstrapping cannot be applied to obtain confidence values. Instead, one can verify to what extend the tree fits the computed distance matrix.
- An alternative to bootstrapping proposed by Guénoche and Garreta (2001) is the proportion of *supporting* quadruples. A quadruple a, b, c, d is called *supporting* an edge e, if $D(a,b) + D(c,d) < \min\{D(a,c) + D(b,d), D(b,c) + D(a,d)\}$.

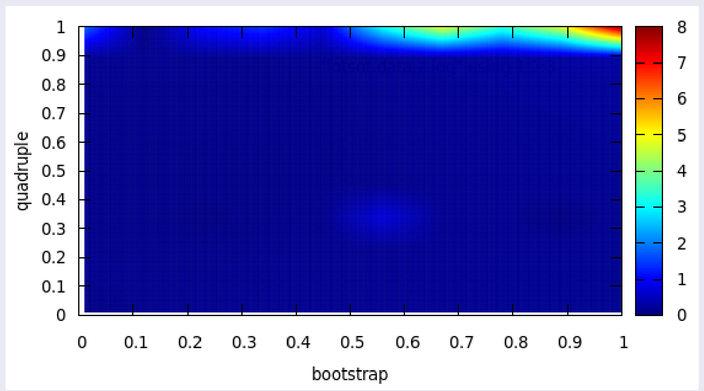


- The relative amount of supporting quadruples can thus be interpreted as confidence value for a given edge.

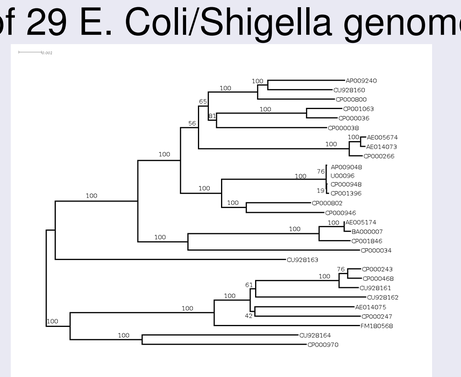
$$SV_e = \frac{\sum_{a,b,c,d} \mathbf{1}(a,b,c,d \text{ support } e)}{|A||B||C||D|}$$

4. Evaluation

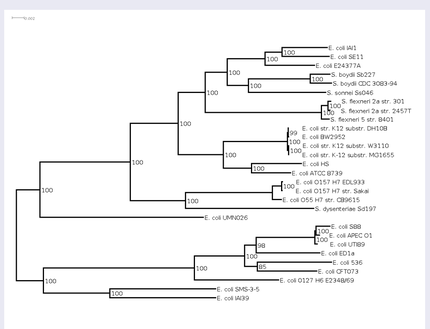
To asses the quality of these new support values, we simulated trees of 100 taxa with genomes of 100 kbp. For these genomes we computed their bootstrapped trees (100 replicates) and their quadruple distances. It is then possible to compare the support of a clade with its corresponding edge. For 100 trees we found a correlation of 0.6.The logarithmic heatmap shows clusters of corresponding support values.



For a comparison on real data, we again used the sample of 29 E. Coli/Shigella genomes.



andi



RAxML

Get your free copy, today.

github.com/evobioinf/andi

Fabian Klötzl
GitHub: @kloetzl
Twitter: @kloetzl

References

- Samuel V. Angiuoli and Steven L. Salzberg. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*, 27(3):334–342, 2011.
- Alain Guénoche and Henri Garreta. Can we have confidence in a tree representation? In Olivier Gascuel and Marie-France Sagot, editors, *Computational Biology*, volume 2066 of *Lecture Notes in Computer Science*, pages 45–56. Springer Berlin Heidelberg, 2001.
- Bernhard Haubold, Fabian Klötzl, and Peter Pfaffelhuber. andi: Fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics*, 2014.
- Bernhard Haubold, Peter Pfaffelhuber, and Mirjana Domazet-Lošo. Estimating mutation distances from unaligned genomes. *Journal of Computational Biology*, 16(10):1487–1500, 2009.
- Chris-Andre Leimeister, Marcus Boden, Sebastian Horwege, Sebastian Lindner, and Burkhard Morgenstern. Fast alignment-free sequence comparison using spaced-word frequencies. *Bioinformatics*, 30(14):1991–1999, 2014.
- Alexandros Stamatakis. Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 2014.
- Huiguang Yi and Li Jin. Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Research*, 41(7), 2013.