Rapid Estimation of Evolutionary Distances between Bacterial Genomes

Fabian Klötzl & Bernhard Haubold

kloetzl@evolbio.mpg.de

Max Planck Institute for Evolutionary Biology, Plön

MAX-PLANCK-GESELLSCHAFT

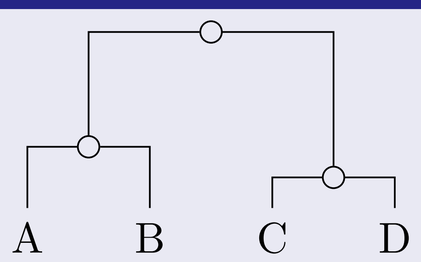
1. Sequence to Phylogeny

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

$\Rightarrow$

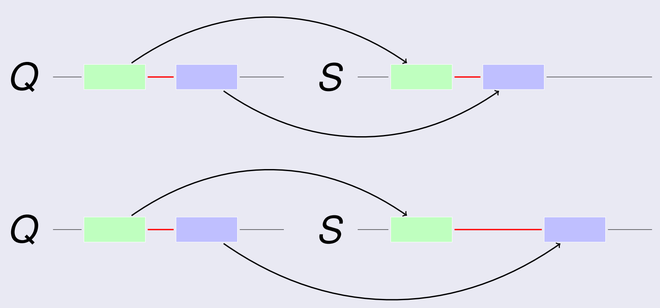
$$\begin{pmatrix} 0 & 0.1 & 0.25 & 0.3 \\ 0.1 & 0 & 0.3 & 0.3 \\ 0.25 & 0.3 & 0 & 0.05 \\ 0.3 & 0.3 & 0.05 & 0 \end{pmatrix}$$

$\Rightarrow$

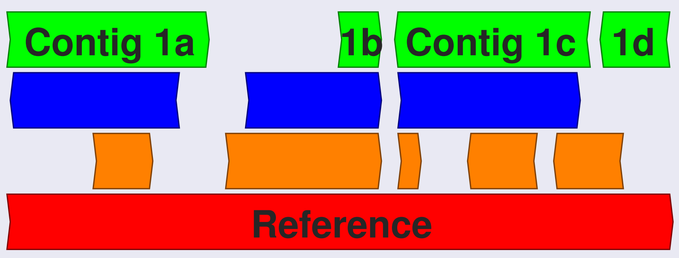


2. Anchor Distances

- The number of SNPs per nucleotide is one of the most widely used measures of evolutionary distances.
- To find homologous sequences, we first look for long and unique matches, termed *anchors*. Two equidistant anchors form a *pair*, creating a homologous region, in which SNPs can be counted.

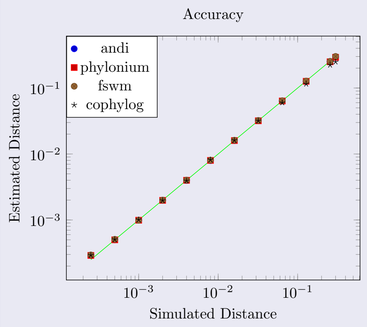


- We use an enhanced suffix array to find anchors.
- The runtime of *andi* is dominated by the pairwise search for exact matches that underlies the construction of anchor distances. To improve its runtime, we are working on a method where all the input sequences are stacked onto a single reference sequence.



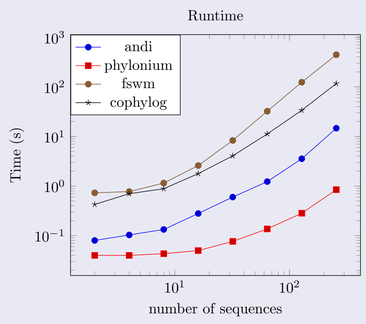
- This method requires only a single index and a linear amount of matches to be computed, making it potentially much faster than the pairwise computations underlying *andi*.

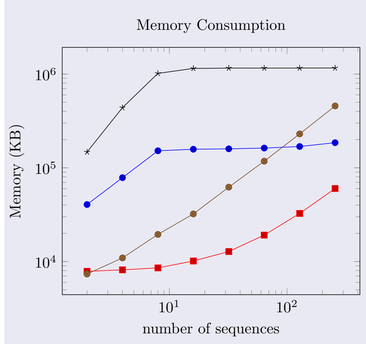
3. Performance



We implemented this idea in our program *phylonium*. To assess its performance, we applied it to simulated data and compared it to the alternative distance measures *andi*, *fswm*, and *cophylog*. The diagram on the left shows that all four methods accurately estimate the number of substitutions per site when applied to simulated and hence "perfect" data.

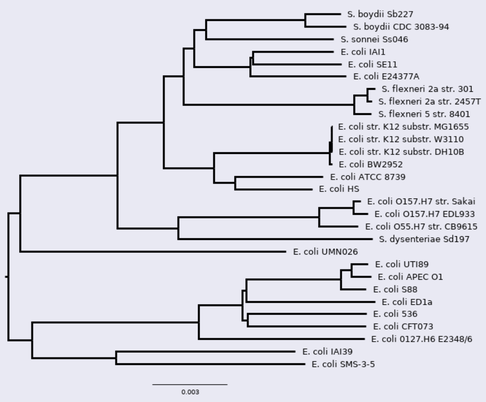
All methods investigated have a runtime roughly linear in the sequence length. However, when speed is measured as a function of the number of sequences, *phylonium* is fastest with the smallest rate of increase. This makes it well suited for the analysis of data sets consisting of many closely related genomes.





Memory usage as a function of number of sequences. Again, *phylonium* outperforms the other methods by using the least amount of memory.

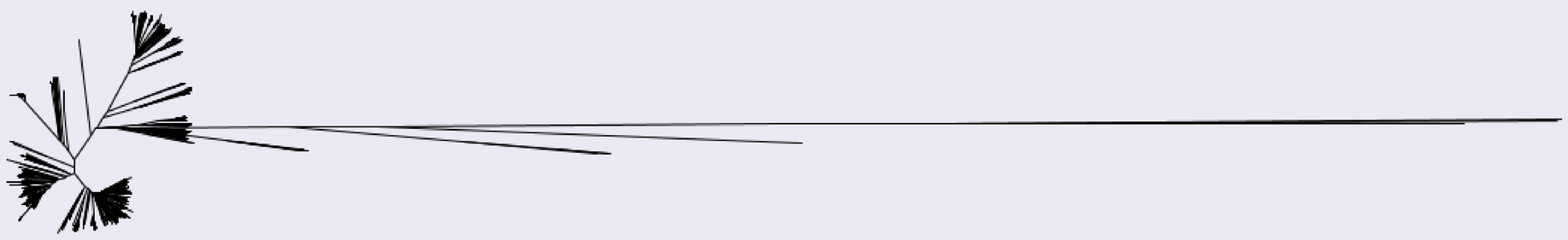
4. Small Bacterial Datasets



To asses the performance of the four methods investigated when applied to real genomes, we clustered 29 *E. coli* and *Shigella* genomes. These genomes are on average 4.9 Mb long, totaling 128 Mb. As our reference phylogeny we used an alignment-based tree computed using *Mugsy* for aligning and *PhyML* for phylogeny estimation.

Method	RF-distance	time (s)	mem (GB)
mugsy	0	10164	2.9
andi	2	22	1.3
cophylog	4	71	1.2
fswm	6	426	2.5
phylonium	6	4	0.5

5. Database-Scale Bacterial Datasets



- The *Ensembl Bacteria* database contains 2677 complete genomes labeled *E. coli*. The FASTA files amount to 14 Gb in total.
- *phylonium* took 68 minutes and 25 GB to compute the phylogeny shown above.
- To perform the same computation, *andi* used 3.5 hours and also 25 GB of RAM.
- The four most divergent sequences are not *E. coli*, but two *Klebsiella pneumoniae*, one *Enterobacter cloacae*, and one *Citrobacter freundii*.

GitHub: @kloetzl
Twitter: @kloetzl
Web: kloetzl.info

References

- Samuel V. Angiuoli and Steven L. Salzberg. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*, 27(3):334–342, 2011.
- Bernhard Haubold, Fabian Klötzl, and Peter Pfaffelhuber. andi: Fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics*, 31(8):1169–1175, 2015.
- Chris-André Leimeister, Salma Sohrabi-Jahromi, and Burkhard Morgenstern. Fast and accurate phylogeny reconstruction using filtered spaced-word matches. *Bioinformatics*, 33(7):971–979, 2017.
- Andrew Yates, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, et al. Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710–D716, 2016.
- Huiguang Yi and Li Jin. Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Research*, 41(7), 2013.