Alignment-Free Phylogeny Reconstruction

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

MAX-PLANCK-GESELLSCHAFT

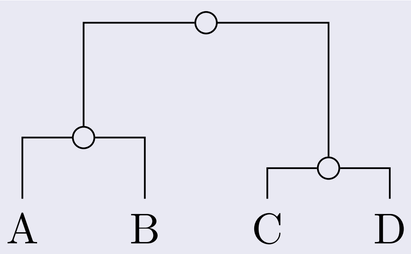
0. Sequence to Phylogeny

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

$\Rightarrow$

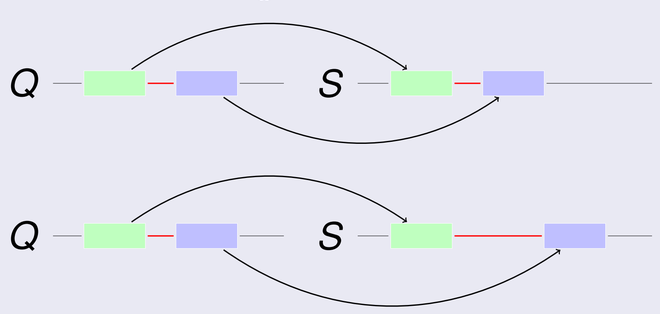
$$\begin{pmatrix} 0 & 0.1 & 0.25 & 0.3 \\ 0.1 & 0 & 0.3 & 0.3 \\ 0.25 & 0.3 & 0 & 0.05 \\ 0.3 & 0.3 & 0.05 & 0 \end{pmatrix}$$

$\Rightarrow$



1. Anchor Distances

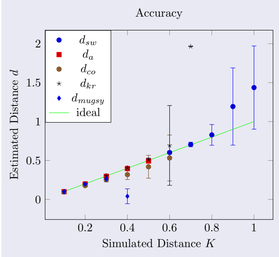
- Counting SNPs is the one of the most widely used measures of evolutionary distances available.
- To find homologous sequences, we first look for long and unique matches, termed *anchors*. Two equidistant anchors form a *pair*, surrounding a homologous region and thus allow counting SNPs.



- We use an enhanced suffix array to find anchors.

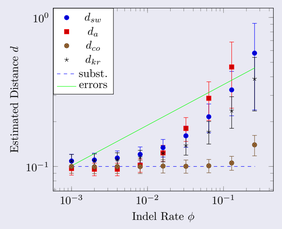
i	SA	LCP	$S^{SA[i]}$	lcp-intervals
0	4	−1	AAGG	
1	0	3	AAG**T**AAGG	3
2	5	1	A**G**G	1
3	1	2	AG**T**AAGG	2
4	7	0	**G**	0
5	6	1	G**G**	1
6	2	1	G**T**AAGG	
7	3	0	**T**AAGG	
8		−1		

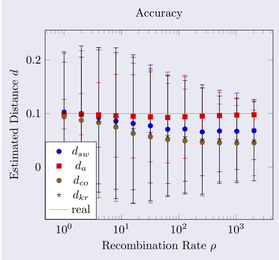
2. Performance



We evaluated 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 the substitution rate (100 runs).

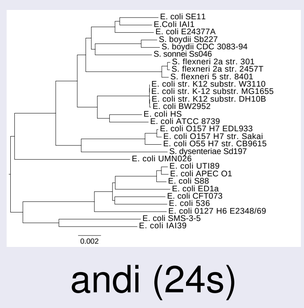
Here, the substitution rate was fixed at 0.1 but a variable number of indels were added. Substitution rate and total errors (SNPs + indels) are shown as solid lines.



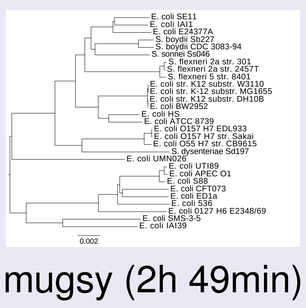


As a final test using simulated data we introduced recombination, as this leads to local variation in the substitution rate.

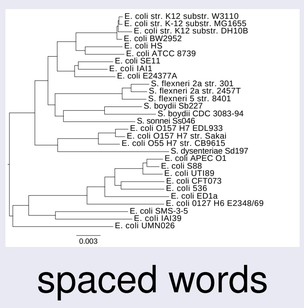
To evaluate the accuracy of the methods on real data, we chose a sample of 29 *E. Coli* and *Shigella* genomes (Eco29). On average the genomes have a length of 4.9 Mbp amounting to 128 MB of data.



andi (24s)



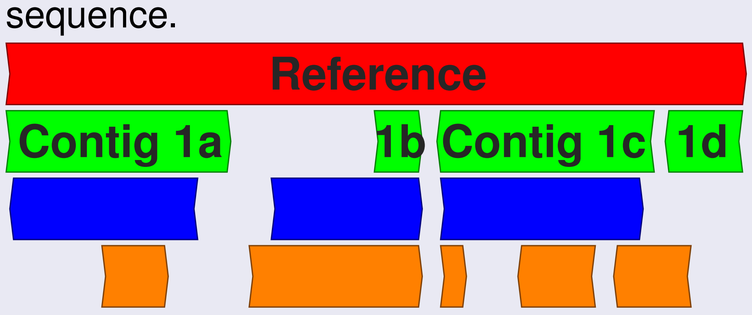
mugsy (2h 49min)
alignment-based



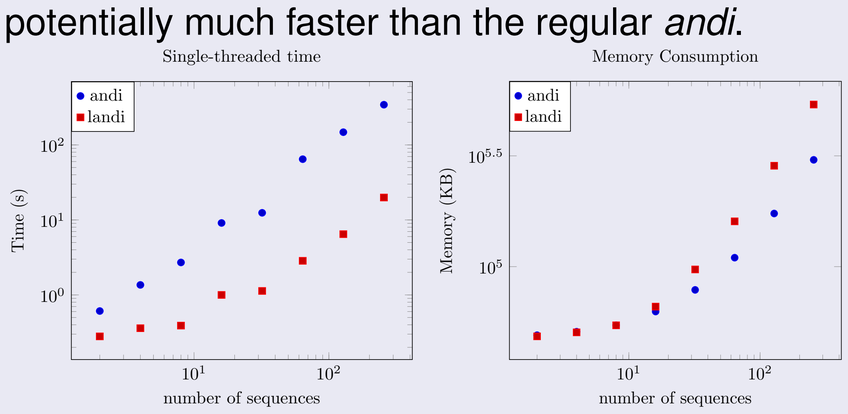
spaced words
(7min)

3. A Faster Algorithm

- 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.

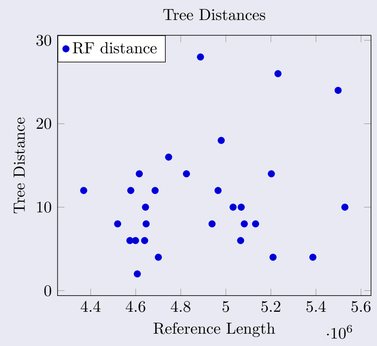


- This method requires only a single index and a linear amount of matches to be computed, making it potentially much faster than the regular *andi*.



4. Choice of the Reference

- The result is highly dependent on the reference. Currently our research is focused on finding an a-priori criterion for a good reference.
- One heuristic might be sequence length. In the aforementioned data set Eco29, the genomes vary by 20% in size. We used each sequence in turn as to the reference and compared the resulting trees to the one produced by *andi*.



There is no correlation between sequence length and the distance.

- Another candidate for the reference could be the union of input genomes, similar to the *pan-genome*. Conversely, the *core genome* (the intersection of the sampled genomes) can also be advantageous.

Getting andi

- Ubuntu, Debian: `apt-get install andi`
- Homebrew: `brew install science/andi`
- ArchLinux: `aura -A andi`
- Source Code: `github.com/evobioinf/andi`

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, 2015.
- 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.
- Hervé Tettelin and al. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: Implications for the microbial “pan-genome”. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13950–13955, 2005.