Comparing Distance Matrices

Fabian Klötzl & Bernhard Haubold

kloetzl@evolbio.mpg.de

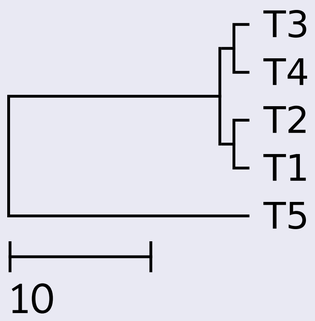
Max Planck Institute for Evolutionary Biology, Plön

MAX-PLANCK-GESELLSCHAFT

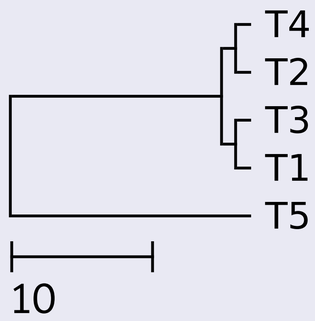
1. Three Phylogenies

In recent years, a number of methods have been published for estimating evolutionary distances between genomes. In order to compare the quality of the various methods, the distances are typically converted to phylogenies and the distance between them measured using the Robinson-Foulds metric. Consider two trees, T_1 and T_2, the RF-distance between them is the number of clades present in T_1 but absent from T_2 plus the number of clades present in T_2 but absent from T_1. To see that this might be a problematic measure, consider the following three trees.

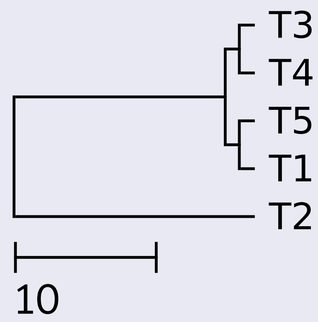
A



B



C



Between trees **A** and **B**, the two taxa T_2 and T_3 were swapped. This leads to a RF distance of 4. In the pair **A**/**C** T_2 instead was swapped with T_5. Again, this leads to a rooted RF distance of 4, however with quite a different biological interpretation.

2. Comparing Distance Matrices

Alignment-free programs for genome comparison produce distance matrices, rather than trees. It thus makes sense to compare these matrices directly. We propose the Δ measure, a variant of the Hausdorff metric.

$$\Delta(D, d) = \max\{|D_{ij} - d_{ij}| : 1 \leq i, j \leq n\}$$

The more similar the two distance matrices D_{ij} and d_{ij} are, the the lower Δ will be. The above trees were computed from the following distance matrices.

A

	T_1	T_2	T_3	T_4	T_5
T_1	0	2	4	4	34
T_2	2	0	4	4	34
T_3	4	4	0	2	34
T_4	4	4	2	0	34
T_5	34	34	34	34	0

B

	T_1	T_2	T_3	T_4	T_5
T_1	0	4	2	4	34
T_2	4	0	4	2	34
T_3	2	4	0	4	34
T_4	4	2	4	0	34
T_5	34	34	34	34	0

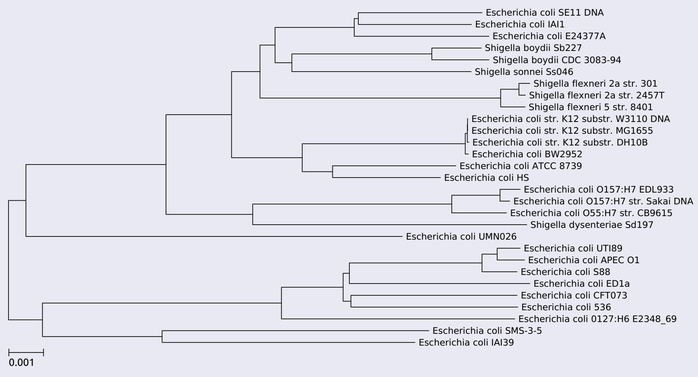
C

	T_1	T_2	T_3	T_4	T_5
T_1	0	34	4	4	2
T_2	34	0	34	34	34
T_3	4	34	0	2	4
T_4	4	34	2	0	4
T_5	2	34	4	4	0

It is easy to see, that $\Delta(\mathbf{A}, \mathbf{B}) = 2$ and $\Delta(\mathbf{A}, \mathbf{C}) = 32$. Thus, **A** and **B** are considered much more similar than **A** and **C**, reflecting our intuition about the trees.

3. Twenty-nine *E. coli/Shigella* Genomes

In a recent analysis Zielezinski et al. (2019) compared various alignment-free tools. One of their data set consisted of 29 whole *Escherichia coli/Shigella* genomes. The reference phylogeny was computed via the multiple sequence alignment tool mugsy. A difficulty for phylogeny estimation here is the significant amount of horizontal gene transfer.

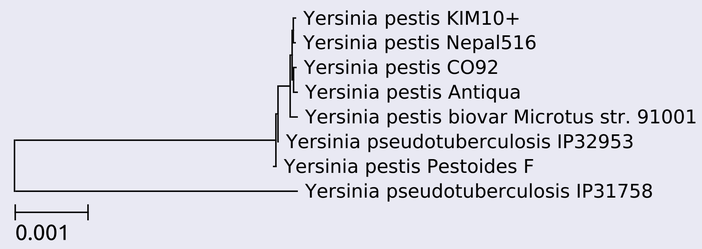


The table gives the accuracy on this dataset measured using Robinson-Foulds and the Hausdorff distance.

Tool	RF	Δ
andi	2	$4.67 \cdot 10^{-3}$
phylonium	6	$4.94 \cdot 10^{-3}$
fswm	6	$6.59 \cdot 10^{-3}$
mash	8	$8.28 \cdot 10^{-3}$
spaced	8	$9.14 \cdot 10^{-3}$
alfpy-euclid	26	$2.96 \cdot 10^{-2}$

4. Eight *Yersinia* Genomes

This dataset consists of eight full *Yersinia* genomes. The short branches are particularly challenging as a small difference in the estimated distances can easily change the topology.



The table gives the accuracy on this dataset measured using Robinson-Foulds and the Hausdorff distance.

Tool	RF	Δ
andi	6	$1.191 \cdot 10^{-3}$
phylonium	8	$1.449 \cdot 10^{-3}$
spaced	6	$2.122 \cdot 10^{-3}$
mash	12	$2.307 \cdot 10^{-3}$
fswm	2	$2.317 \cdot 10^{-3}$
alfpy-euclid	5	$7.331 \cdot 10^{-3}$

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

References

- 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.
- Brian D. Ondov, Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. Mash: fast genome and metagenome distance estimation using minhash. *Genome Biology*, 17(1):132, 6 2016.
- Andrzej Zielezinski, Hani Z. Girgis, Guillaume Bernard, Chris-Andre Leimeister, Kujin Tang, Thomas Dencker, Anna Katharina Lau, Sophie Röhling, Jae Jin Choi, Michael S. Waterman, Matteo Comin, Sung-Hou Kim, Susana Vinga, Jonas S. Almeida, Cheong Xin Chan, Benjamin T. James, Fengzhu Sun, Burkhard Morgenstern, and Wojciech M. Karlowski. Benchmarking of alignment-free sequence comparison methods. *Genome Biology*, 20, Jul 2019.