Fast and Accurate Distance Computation from Unaligned Genomes

Fabian Klötzl & Bernhard Haubold

GCB 2018

MPI for Evolutionary Biology, Plön

Alignment-Based Phylogeny Reconstruction

Unaligned Sequences

```
>A
AACGTTGTGCA
>B
CACGTTGGT
>C
AACGATGCGT
>D
ACCGGTGTGCT
```

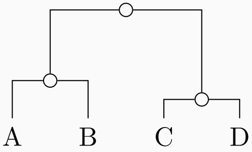
⇒

Alignment

```
>A
AACGTTGTGCA
>B
CACGTT--GGT
>C
AACGATGCG-T
>D
ACCGGTGTGCT
```

⇒

Phylogeny



Alignment-Free Phylogeny Reconstruction

Unaligned Sequences

```
>A
AACGTTGTGCA
>B
CACGTTGGT
>C
AACGATGCGT
>D
ACCGGTGTGCT
```

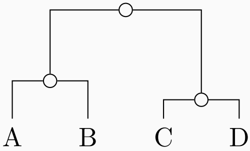
$\Rightarrow$

Distance Matrix

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

$\Rightarrow$

Phylogeny

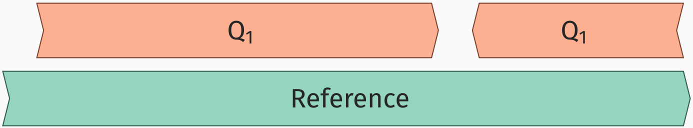


Phylonium

1. Use one sequence as the common coordinate system.

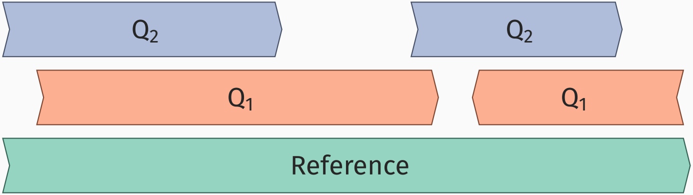
Phylonium

1. Use one sequence as the common coordinate system.
2. Align all other sequences against this reference.



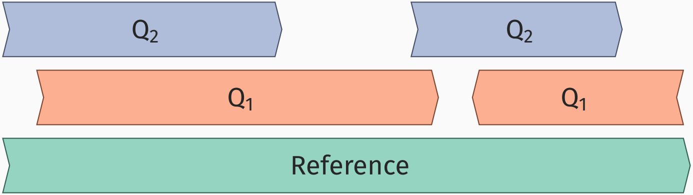
Phylonium

1. Use one sequence as the common coordinate system.
2. Align all other sequences against this reference.



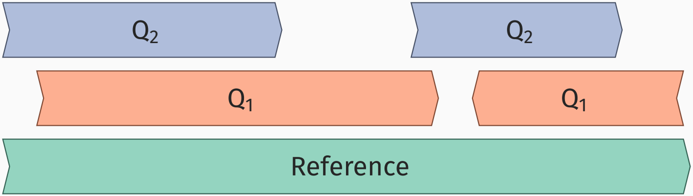
Phylonium

1. Use one sequence as the common coordinate system.
2. Align all other sequences against this reference.
3. For all pairs inspect the overlapping regions.

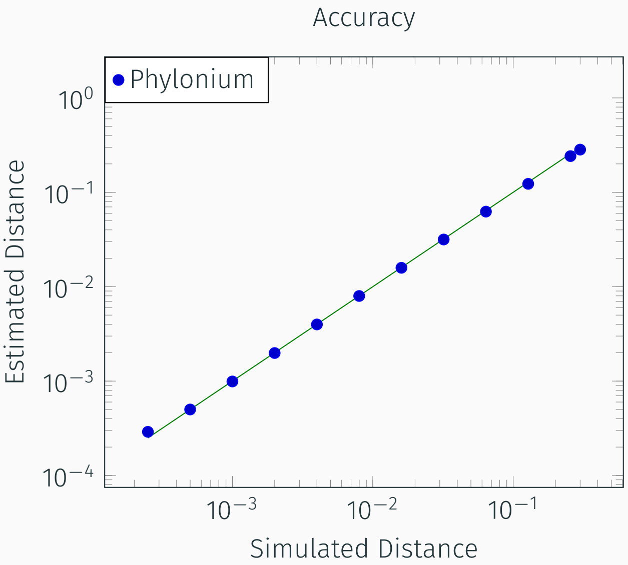


Phylonium

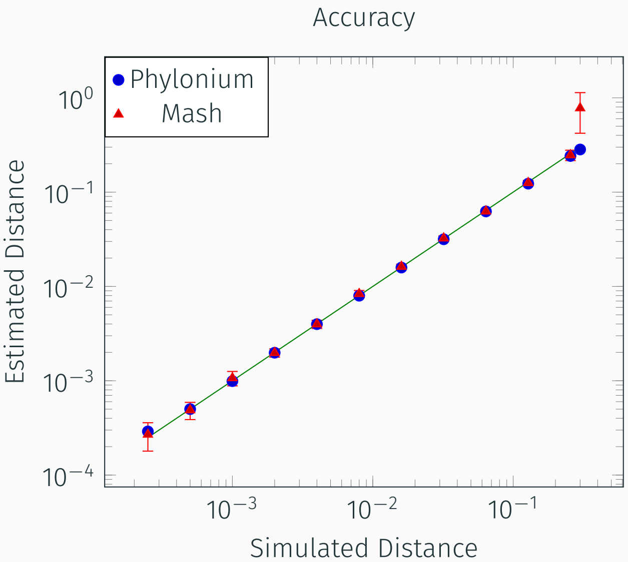
1. Use one sequence as the common coordinate system.
2. Align all other sequences against this reference.
3. For all pairs inspect the overlapping regions.
4. Estimate evolutionary distance from substitution rate.



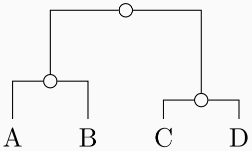
Quality

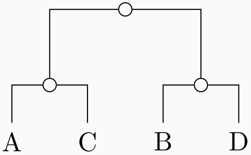

Accuracy
Phylonium
Estimated Distance
Simulated Distance
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-3}
10^{-2}
10^{-1}

Quality



Accuracy
Phylonium
Mash
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
Estimated Distance
10^{-3}
10^{-2}
10^{-1}
Simulated Distance





Robinson-Foulds Distance

The RF distance measures the number of partitions in the first tree, but not in the other. Thus, it only considers the topology. For above trees the RF distance is 2.

Phylogenetic Quality — Relative Matrix Dissimilarity

$$\mathsf{A} = \begin{pmatrix} 0 & 0.1 & 0.2 \\ 0.1 & 0 & 0.3 \\ 0.2 & 0.3 & 0 \end{pmatrix} \qquad \mathsf{B} = \begin{pmatrix} 0 & 0.11 & 0.22 \\ 0.11 & 0 & 0.33 \\ 0.22 & 0.33 & 0 \end{pmatrix}$$

Matrix Dissimilarity

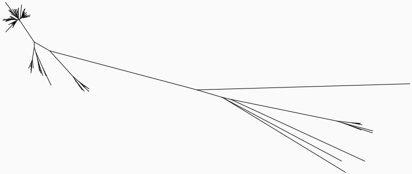
Compute the average relative dissimilarity of the entries.

$$d(A, B) = \frac{4}{n(n-1)} \sum_{i} \sum_{j<i} \frac{|a_{ij} - b_{ij}|}{a_{ij} + b_{ij}}$$

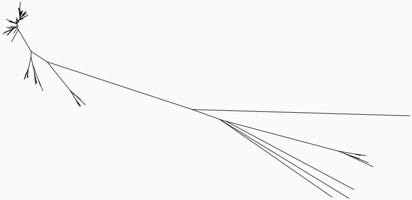
For above examples, $d(\mathsf{A}, \mathsf{B}) = 0.095$ approximately 10 %.

109 *E. coli* Genomes

Mugsy: 2 days
(alignment-based)

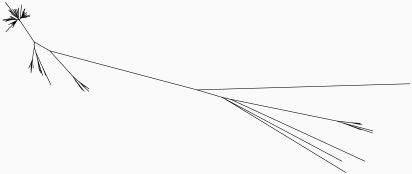


Phylonium: 23 s
RF distance: 130
relative dissimilarity: 20 %

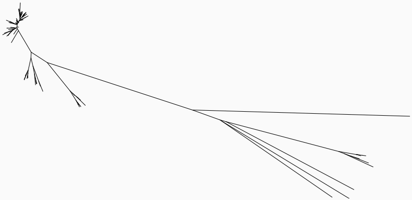


109 *E. coli* Genomes

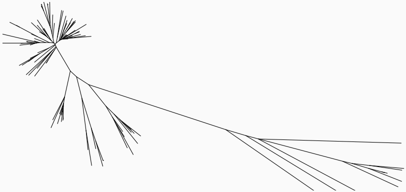
Mugsy: 2 days
(alignment-based)



Phylonium: 23 s
RF distance: 130
relative dissimilarity: 20 %



Mash: 20 s
RF distance: 161
relative dissimilarity: 84 %



2681 *E. coli* from Ensembl Genomes

Phylonium: 378 s

Mash: 49 s

Summary

- Goal: Phylogeny reconstruction from whole genomes.
- Alignment-free distance methods are fast and accurate.
- Work best on data from pathogen outbreaks.
- Scale up to massive data sets.
- Paper on Phylonium in prep.

`kloetzl@evolbio.mpg.de`

MAX-PLANCK-GESELLSCHAFT