Fast Multiple Sequence Alignment

Fabian Klötzl

MPI for Evolutionary Biology, Plön

Aquavit, 2016-05-27

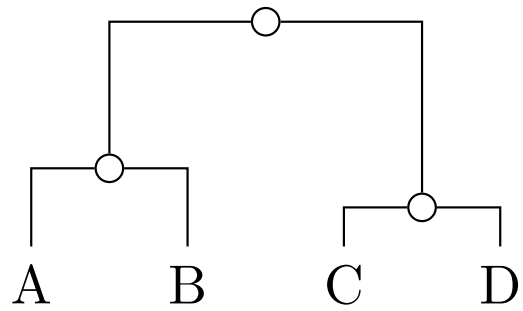
You Had One Job

```
>A
AACGTTGTGCA
>B
CACGTTTTG
>C
AACGATGCGC
>D
ACCGGTGTGCT
```

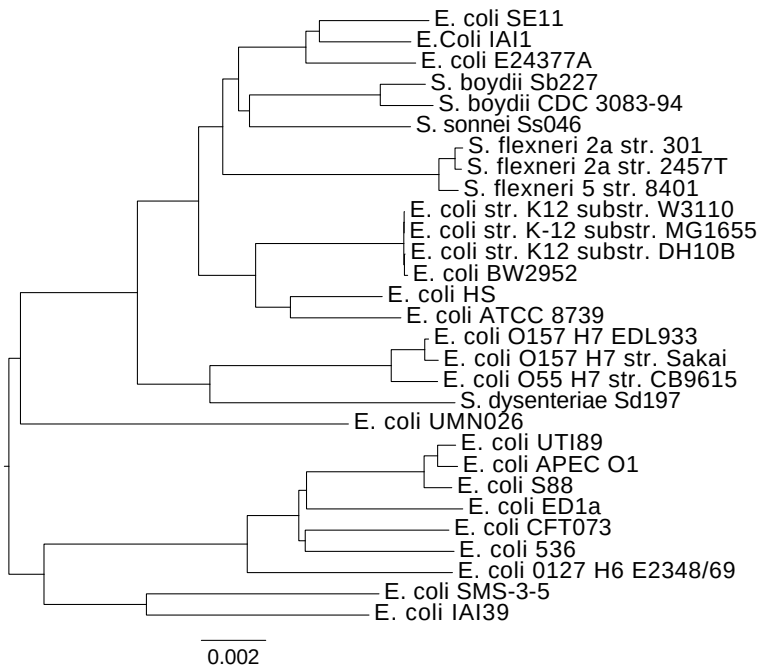
$\Rightarrow$

```
A   AACGTTGTGCA
B   CACGTT-TTG-
C   AACGATGCGC-
D   ACCGGTGTGCT
```

$\Rightarrow$



Alignment
Y U SO SLOW?

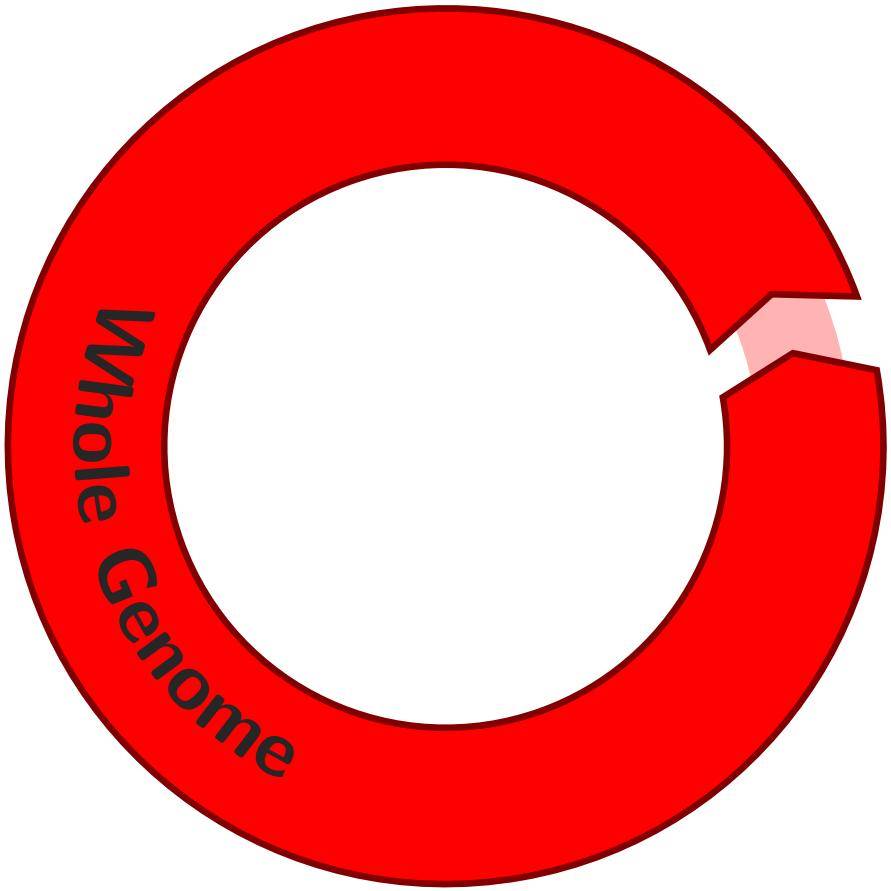


Eco29

- 29 *E. coli/ Shigella* genomes
- file size: 138 MByte

- andi: 45 s
- Mugsy: 2 h 49 min
- Multiz & TBA (UCSC Genome Browser): 1 d 2 h
- Muscle, MAFFT, HAlign exceed memory limit

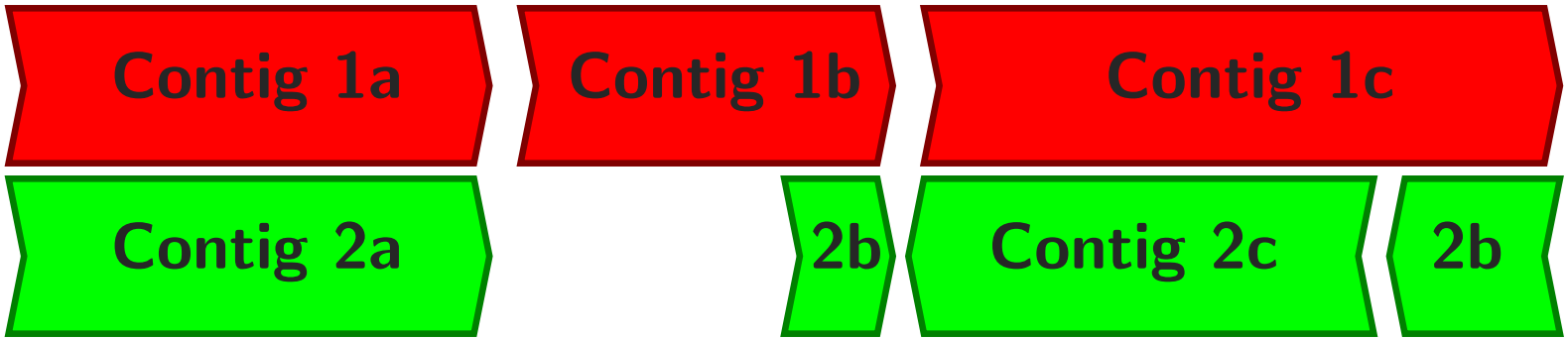
Assembly
What you won't get



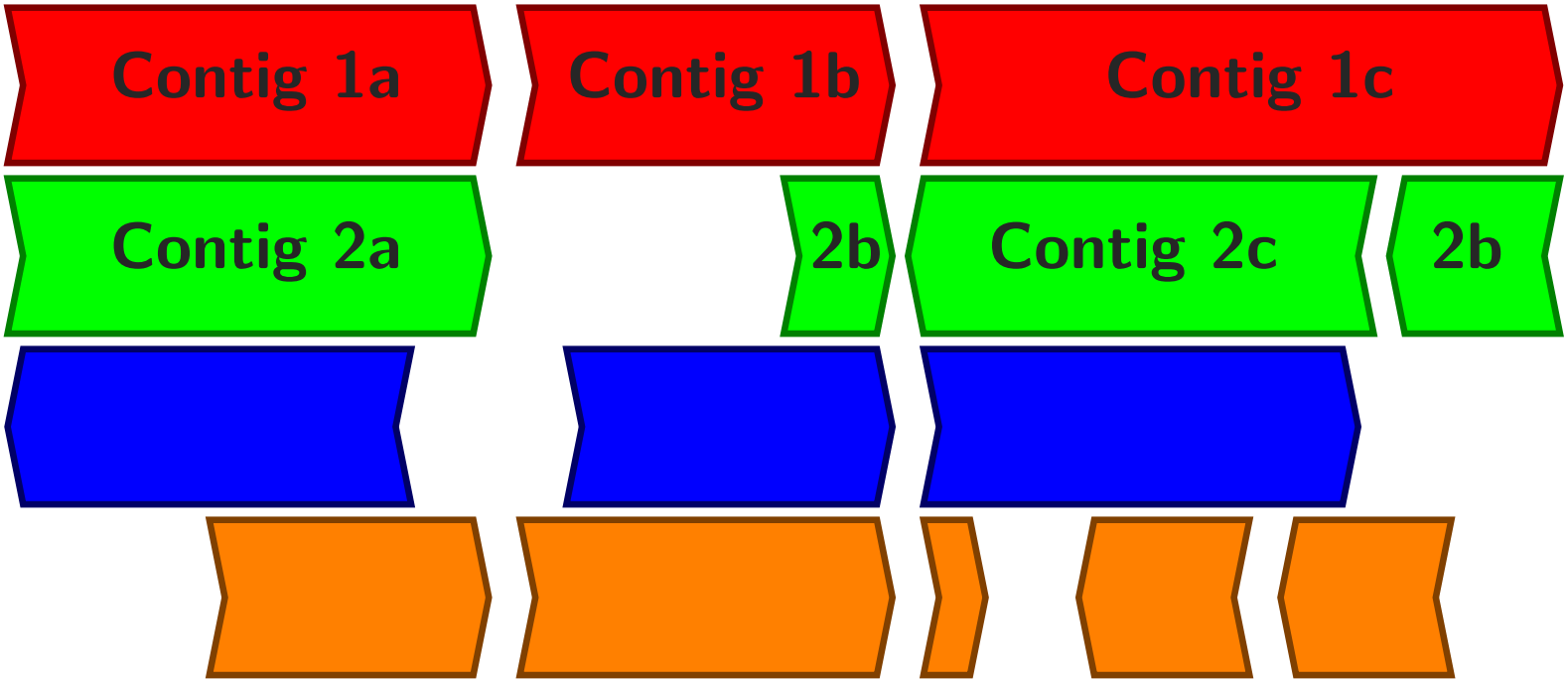
Contigs

Contigs

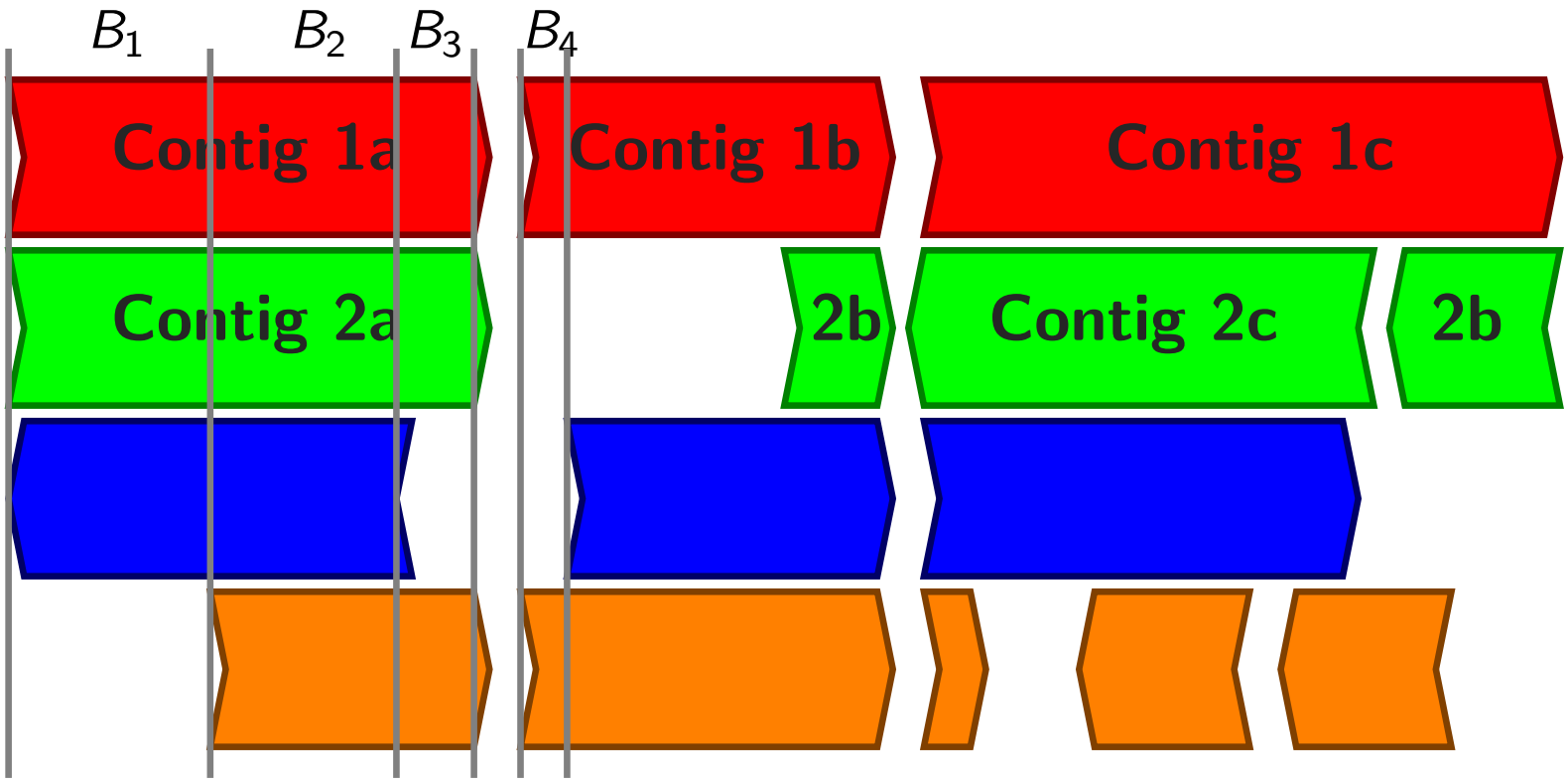
Reference Alignment



More Reference Alignment



Blockwise Reference Alignment



Blocking

Lessons learned from andi

Gaps are hard to compute. What happens if we leave them out?

Gapped Block

S1	A	C	C	G	T	T
S2	A	C	-	G	T	A
S3	A	T	C	C	T	T

$\Rightarrow$

Block 1

S1	A	C
S2	A	C
S3	A	T

Block 2

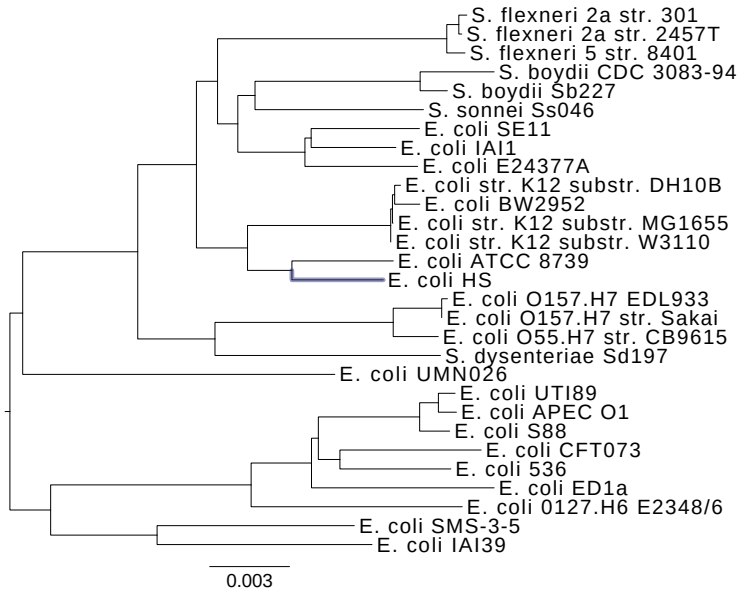
S1	C
S3	C

Block 3

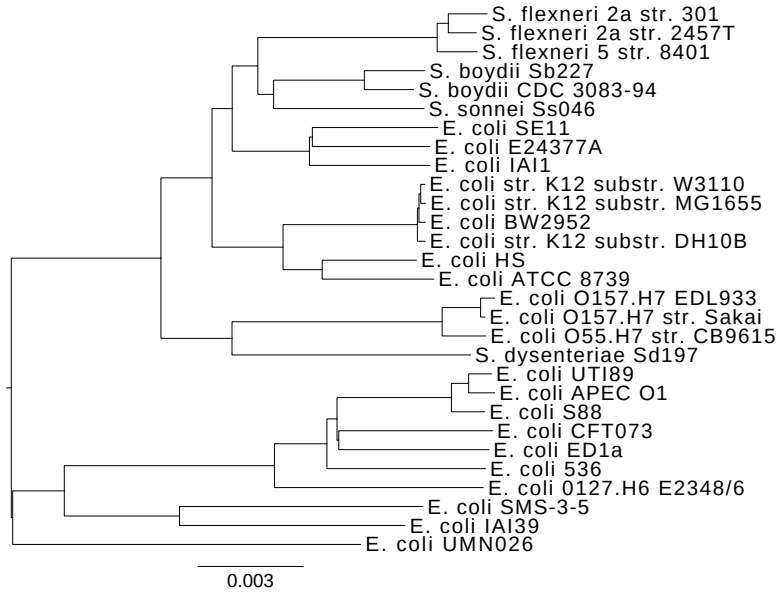
S1	G	T	T
S2	G	T	A
S3	C	T	T

Eco29 again

Multiz & TBA

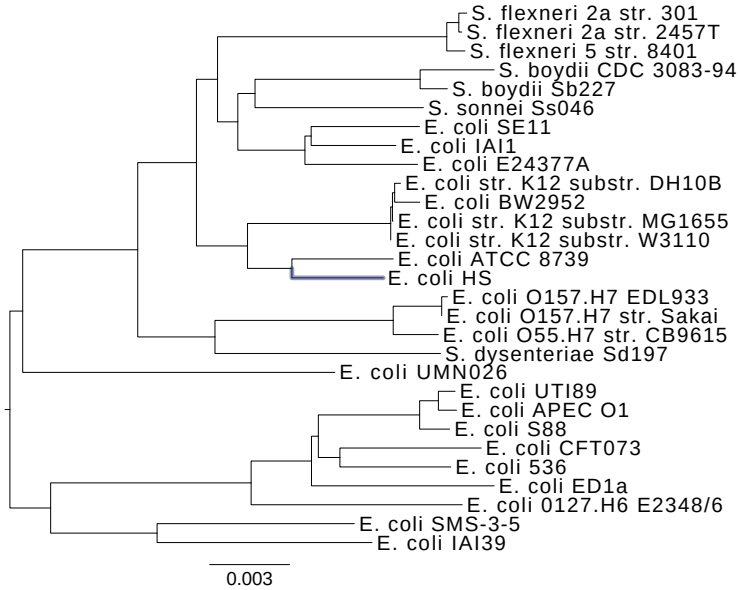


UGAlign



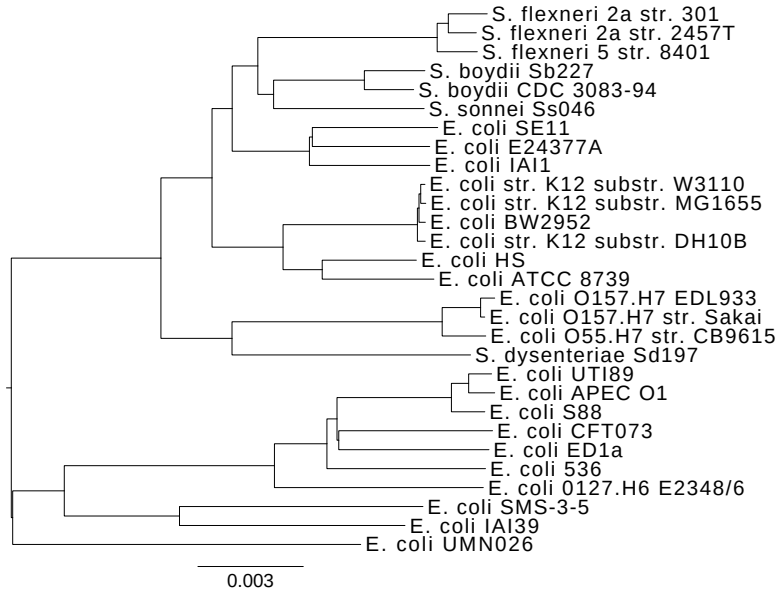
Eco29 again

Multiz & TBA



1d 2h

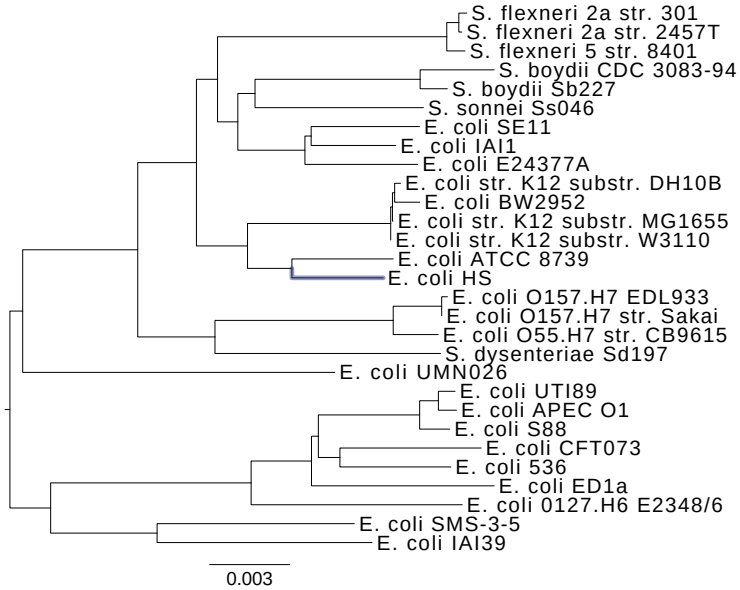
UGAlign



3.5 s

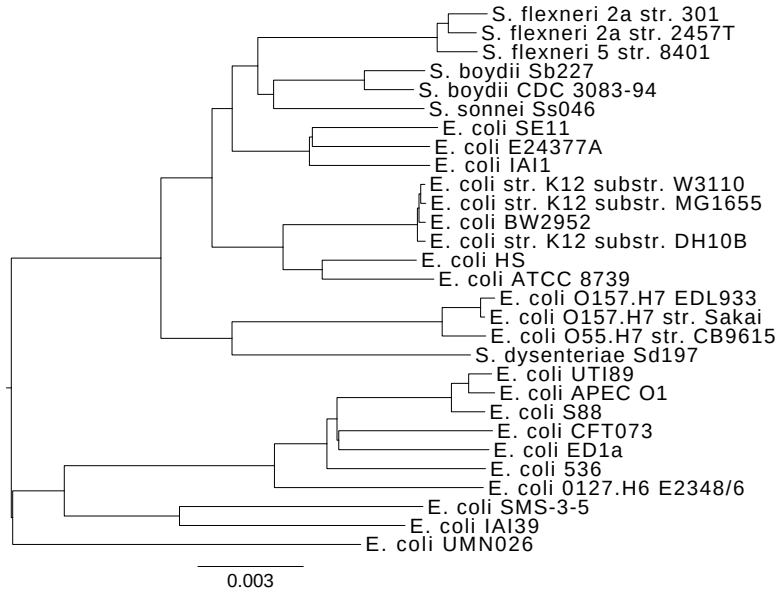
Eco29 again

Multiz & TBA



1d 2h

UGAlign



3.5 s

$\times$25,000 faster

Twenty-five thousand

$\times 25{,}000$ faster

Twenty-five thousand

$\times 25{,}000$ faster

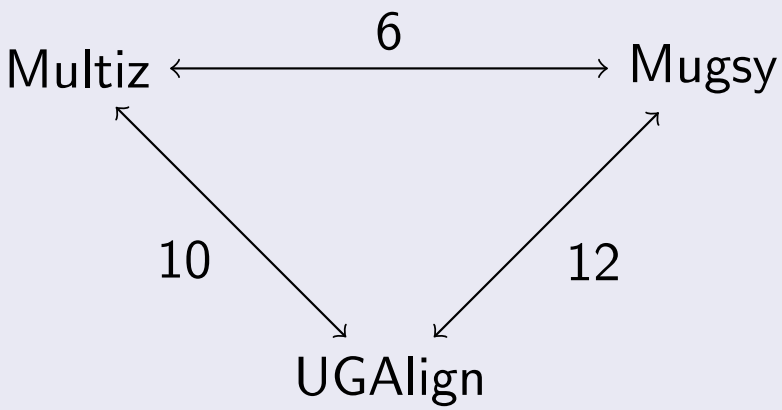
Twenty-five thousand

$\times 25{,}000$ faster

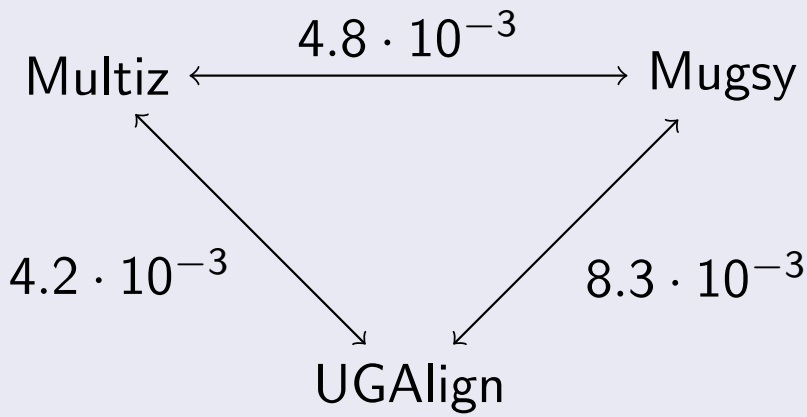
$\times 3{,}000$ w.r.t Mugsy

Tree Distances on Eco29

Symmetric (Robinson-Foulds)



Branch Score

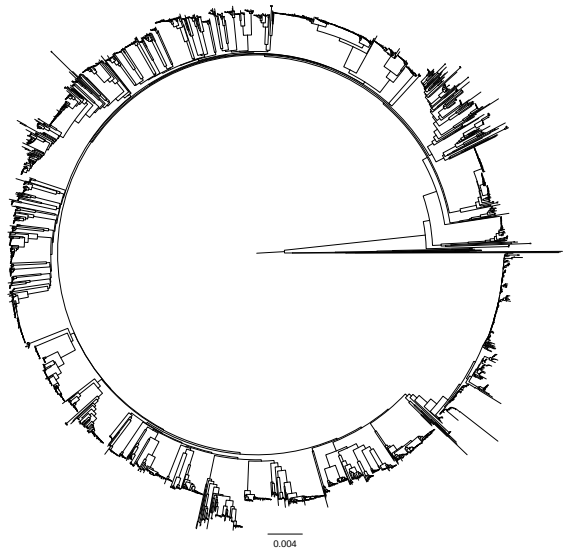


New Frontiers

Pneu3085

- 3085 *Streptococcus pneumoniae* genomes
- multiple contigs per genome
- 2 million nucleotides per genome
- a total filesize of 6.8 GB

- andi: 4 h 59 min, 9.8 GB RAM
- UGAlign: 15 min, 20 GB RAM, alignment file: 36 GB



UGAlign

Pros

- super-duper fast
- reasonably accurate
- $O(n \cdot l \cdot \log(n \cdot l))$
- handles massive data

Cons

- ungapped
- reference-based
- sensitive w.r.t reference-choice
- heuristic: longest

UGAlign

Pros

- super-duper fast
- reasonably accurate
- $O(n \cdot l \cdot \log(n \cdot l))$
- handles massive data

Cons

- ungapped
- reference-based
- sensitive w.r.t reference-choice
- heuristic: longest

Speed Up

3,000 – 25,000

Thank you for your attention

Bernhard Haubold

- Available in a repository near you (soon)
- Did I mention the 25,000 speed up?
- Previous projects: `github.com/evolbioinf`
- If you use Multiz, you are contributing to global warming
- Random information: `klötzl.info`
- TODO: Don't forget the 25,000x improvement