Fast Recovery of Evolutionary Trees with Thousands of Leaves

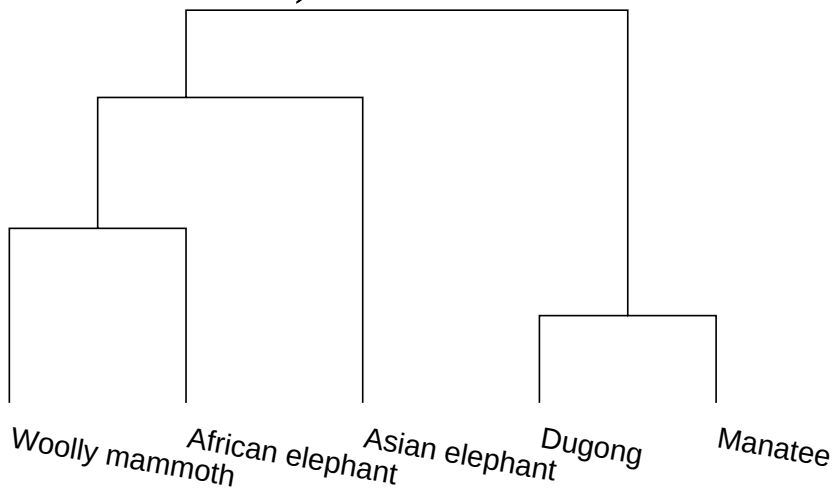
Miklós Csűrös

Department of Computer Science

Yale University

Molecular evolution

evolutionary tree (Noro et al. 1998)



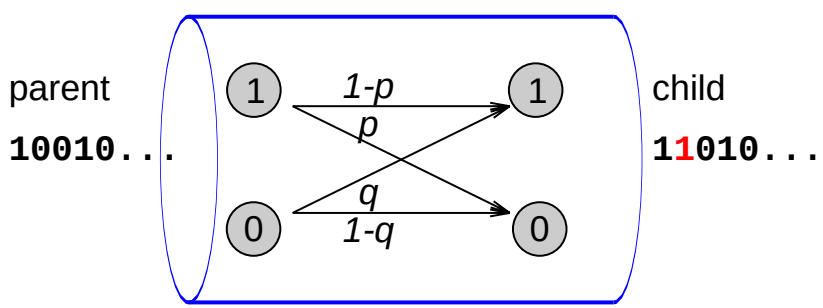
homologous gene sequences

```
Woolly mammoth    ...CTAAATCATCACTGATC--AAAGAGAGC...
African elephant  ...CTAAATCATCACCGATC--AAAGAGAGC...
Asian elephant    ...CTAAATCATCGCTGATC--AAAGAGAGC...
Dugong            ...TTAAATCACTCCCGATCATAAAG-GAGC...
Manatee           ...TCAAATCATTACTGACCATAAAG-GAGC...
```

differences between sequences grow with time

Markov model

- each character evolves independently
- root sequence characters are i.i.d.
- character transitions on edges



character at node u: $\xi(u)$
– random variables forming a Markov chain on each path

Distance based algorithms

Distance [coin-toss model: symmetric mutations]

$$D(u,v) = -\ln\big(\mathbb{P}\{\xi(u) = \xi(v)\} - \mathbb{P}\{\xi(u) \neq \xi(v)\}\big)$$

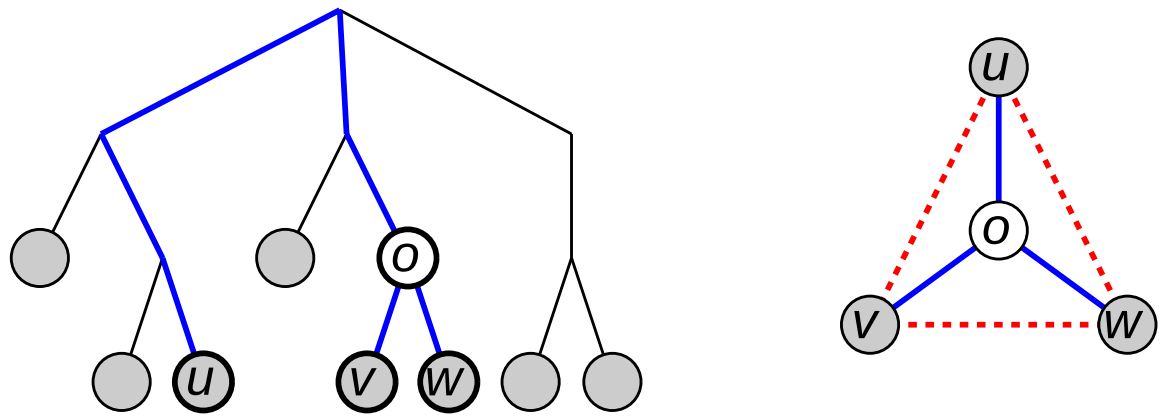
- symmetric
- additive along paths

Distance-based algorithm:

1. distance estimation between leaves $\hat{D}$
2. algorithm using pairwise distance matrix

Additive tree problem

build edge-weighted tree from sum-of-edge-weigths on paths between leaves – use triplets (eg., Waterman, Smith, Singh, Beyer 1977)



$$D(u,o) = \frac{D(u,v) + D(u,w) - D(v,w)}{2}$$

Estimated distances

Use relative frequencies in sample

$$\hat{D}(u,v) = -\ln\left(\hat{P}\{\xi(u) = \xi(v)\} - \hat{P}\{\xi(u) \neq \xi(v)\}\right)$$

- estimation error
- harder to recognize separate triplet centers
- estimation error grows with distance

Triplet center estimation

Similarity:

$$\begin{aligned} S(u,v) &= \exp\big(-D(u,v)\big) \\ &= \mathbb{P}\{\xi(u)=\xi(v)\} - \mathbb{P}\{\xi(u)\neq\xi(v)\} \end{aligned}$$

Distance estimation error: for $0<\varepsilon<1$,

$$\mathbb{P}\left\{\hat{D}(u,o)-D(u,o)\geq\frac{-\ln(1-\varepsilon)}{2}\right\}\leq a\exp\left(-b\ell\varepsilon^2 S^2(u,v,w)\right)$$

(with $a,b>0$ constants)

Average similarity:

$$S(u,v,w)=\frac{3}{\frac{1}{S(u,v)}+\frac{1}{S(u,w)}+\frac{1}{S(v,w)}}$$

Harmonic Greedy Triplets

Add one internal node and leaf at a time

- greedy selection of triplet by average similarity
- recognize separate inner nodes
  (four-point condition)
- restrict pool of triplets considered
  (relevant triplets)

Sample length

Bounded mutation probabilities on edges

$$0 < f \leq p_e \leq g < \frac{1}{2}$$

There exists

$$\ell = \mathcal{O}\left(\frac{\log\frac{1}{\delta} + \log n}{(1-2g)^{\mathcal{O}(d)} f^2}\right)$$

s.t. with probability $1-\delta$, topology is recovered correctly

- tree depth: $d \leq 1 + \log_2(n-1)$

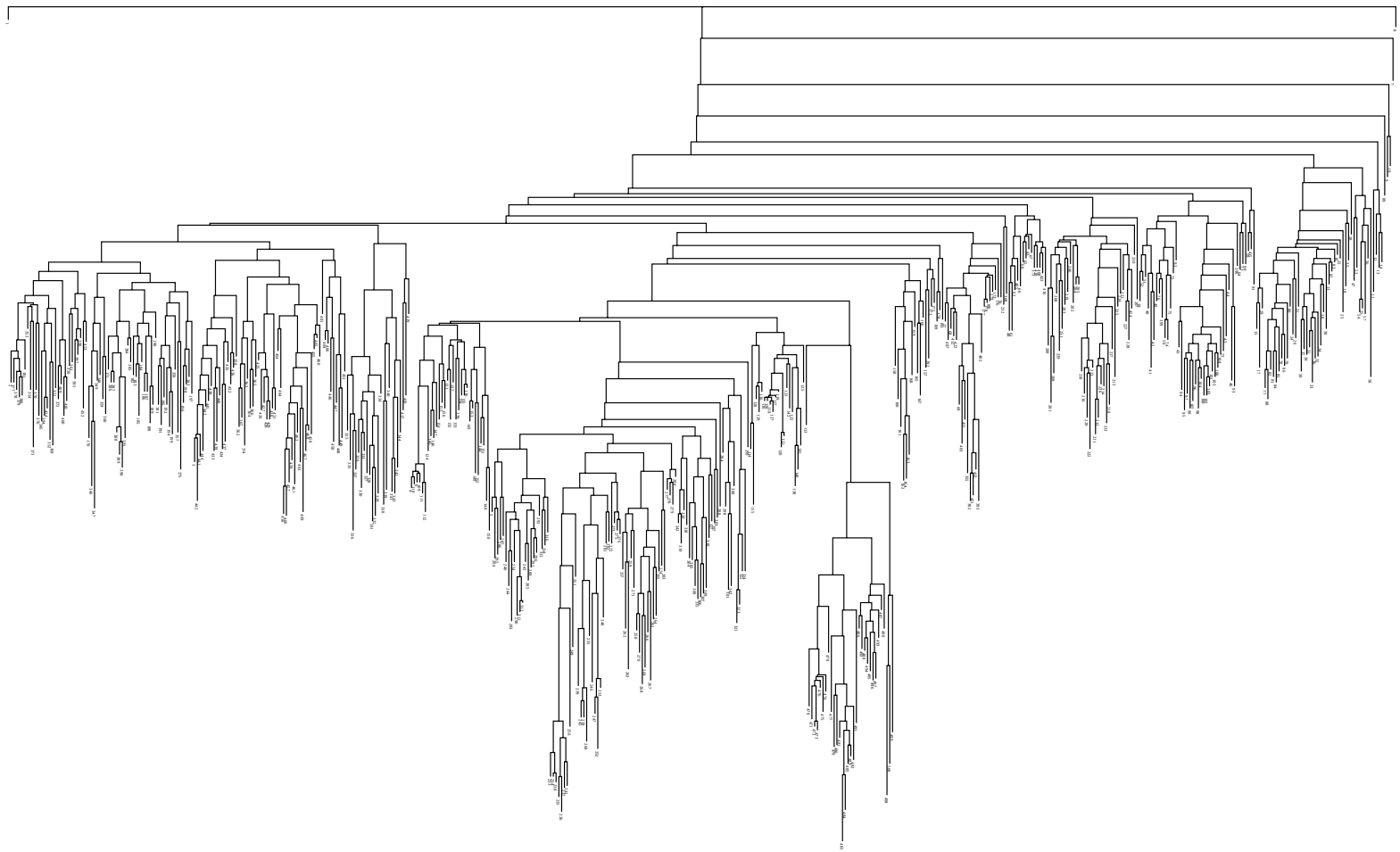
Simulated experiments

compare to Neighbor Joining (Saitou and Nei 1987) and other algorithms
simulate DNA sequence evolution (Jukes-Cantor & K2P+Γ)

- 500 leaf tree (Chase et al. 1993)
  tree of 500 seed plants from rbcL gene
- 1895 leaf tree (RDP 1999)
  tree of 1895 eukaryotes from ribosomal SSU
- 3135 leaf tree (RDP 1999)
  tree of 3135 Proteobacteria from ribosomal SSU

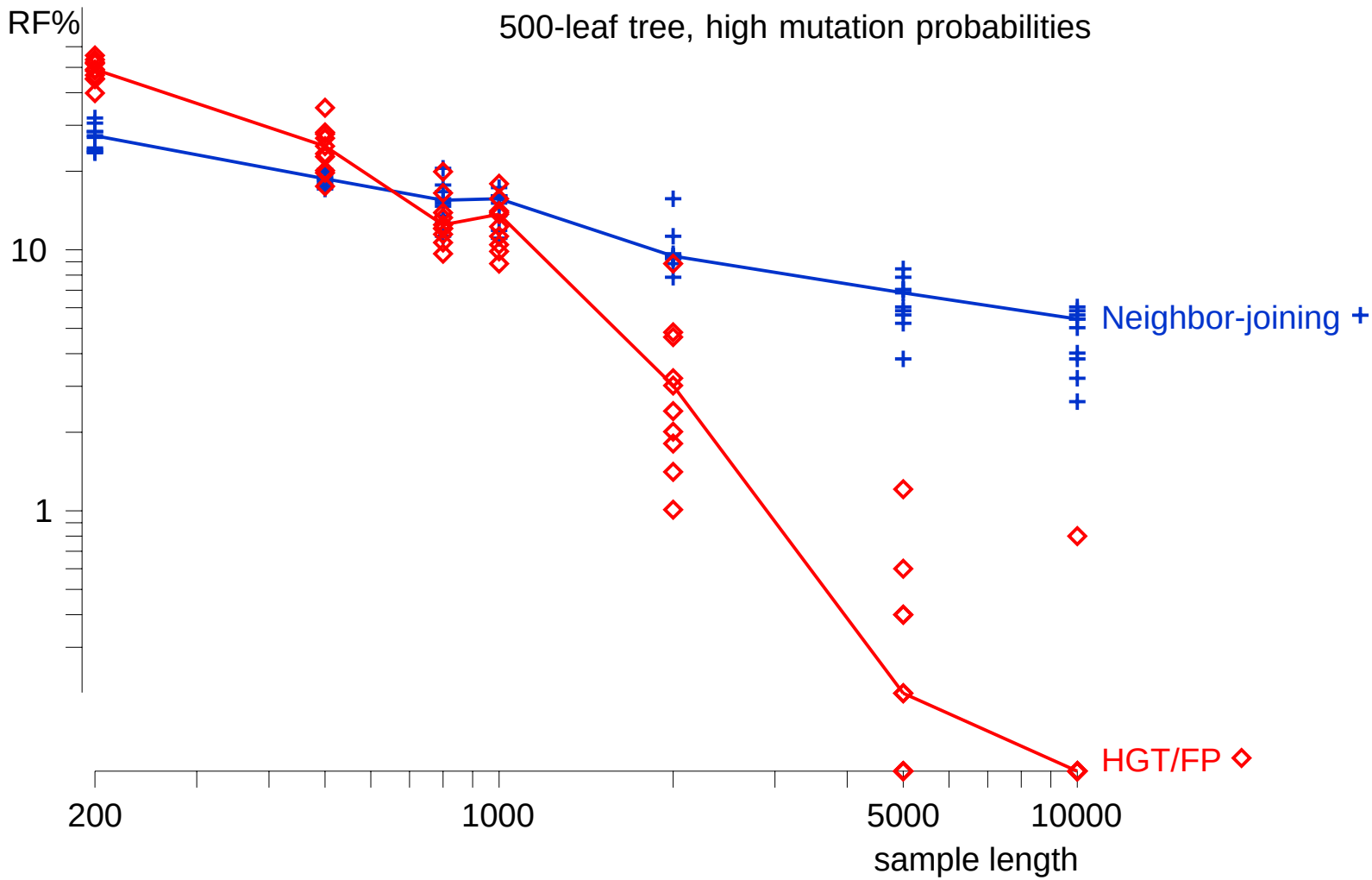
evaluate by Robinson-Foulds distance (1981):
percentage of misplaced internal edges

500-leaf tree

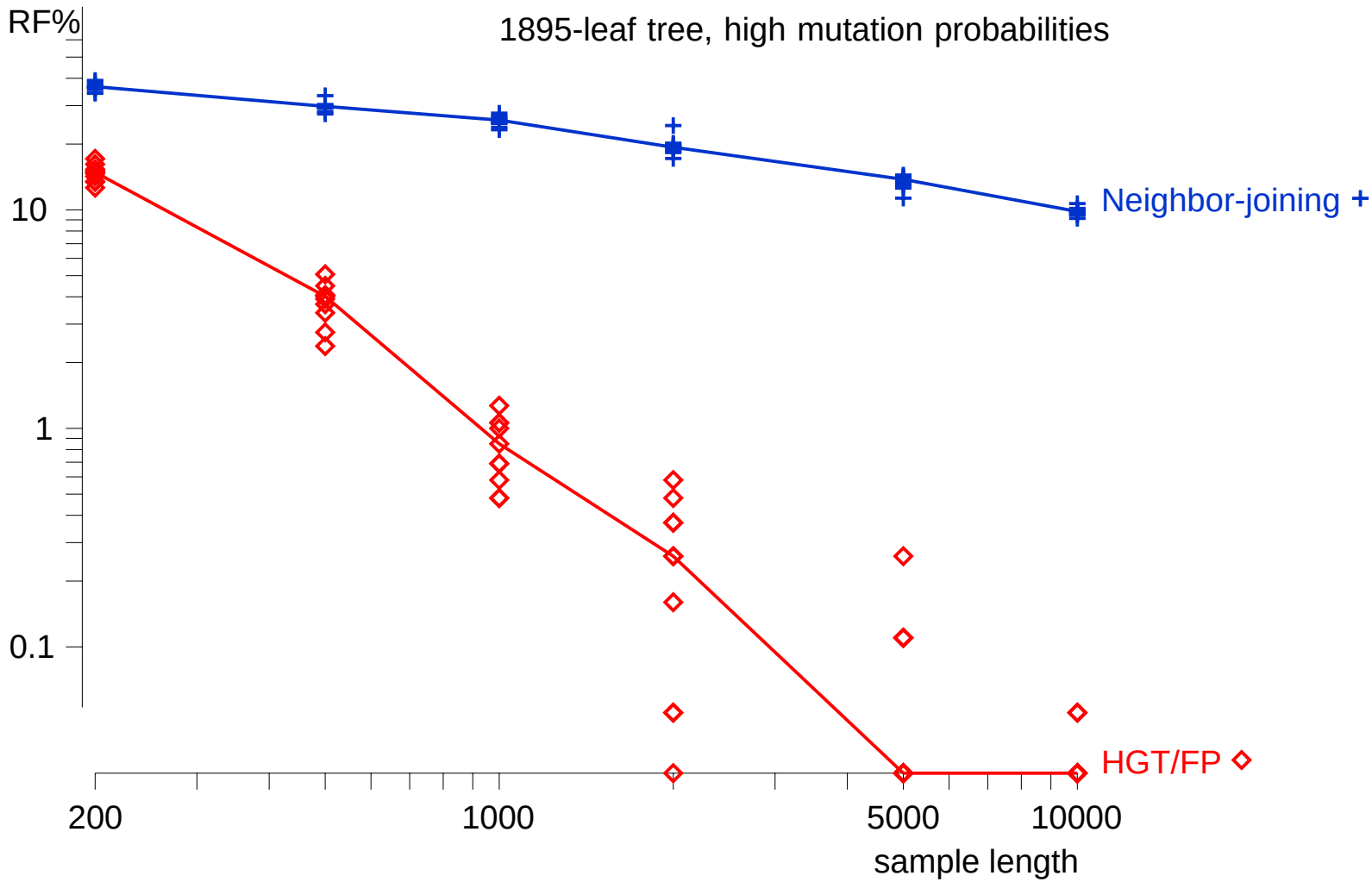


Experimental sample length — 500 leaf tree

varying sample length

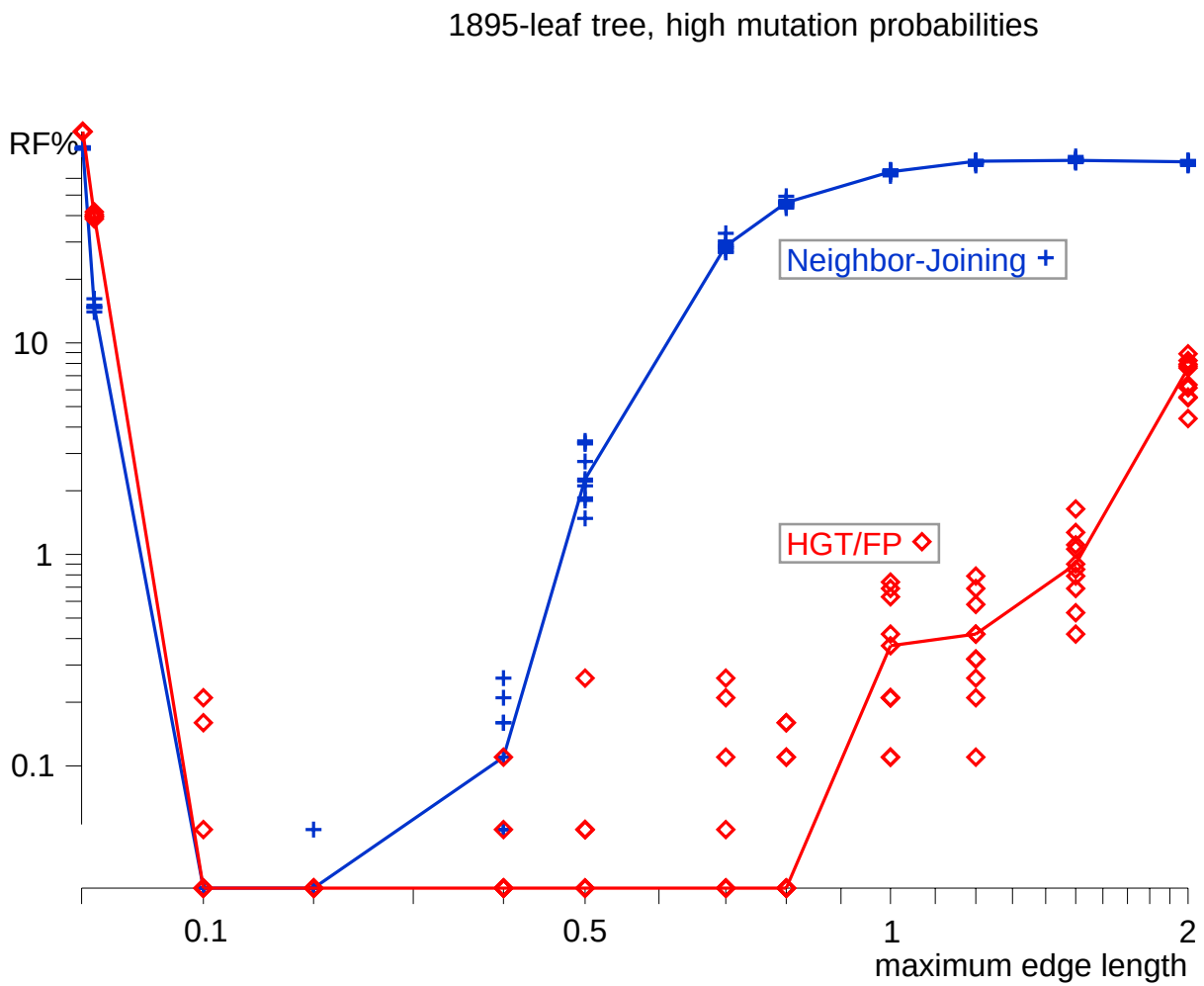


Experimental sample length — 1895 leaf tree

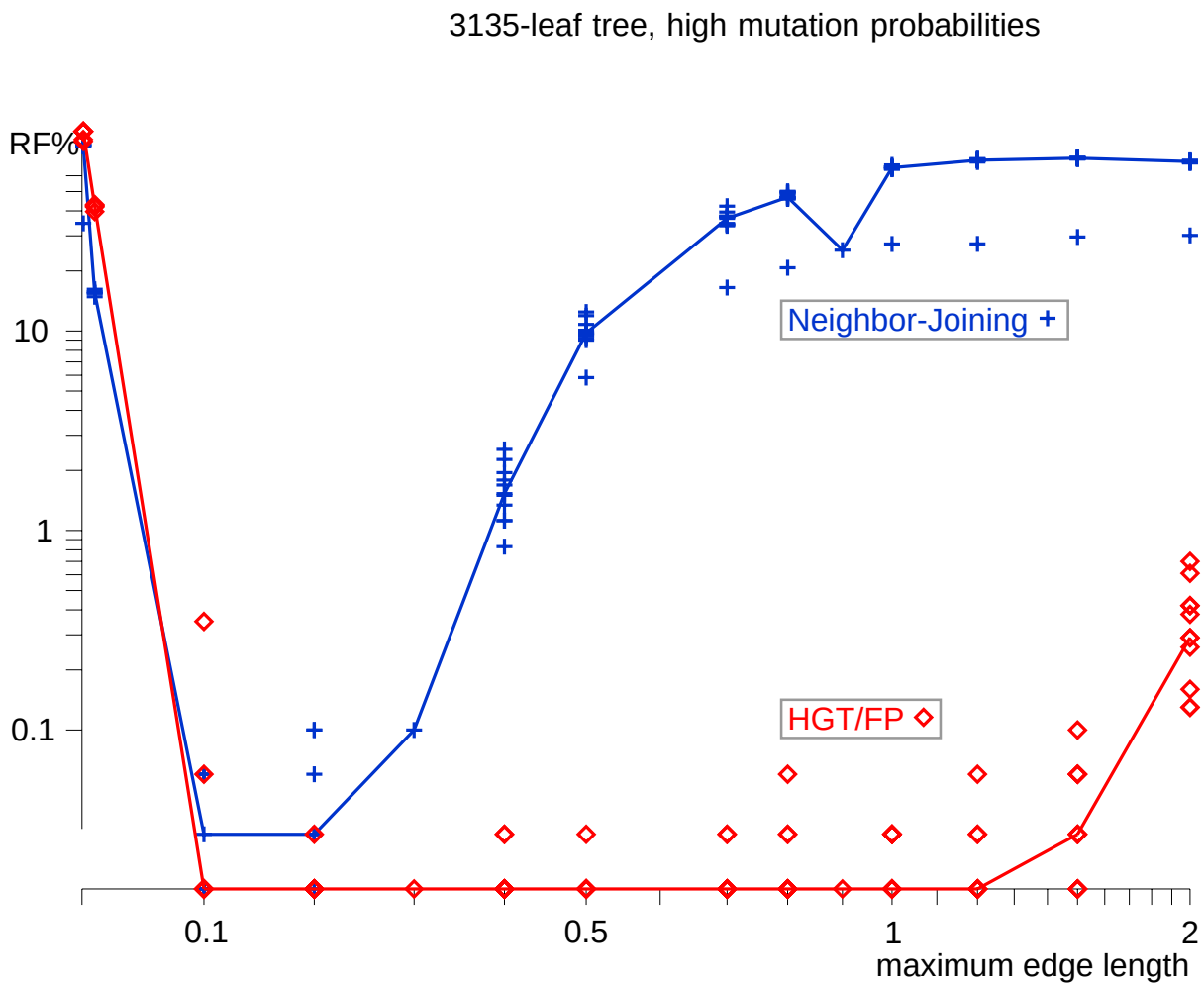


Experimental success — 1895 leaf tree

varying mutation probabilities



Experimental success — 3135 leaf tree



Summary

distance-based algorithm with

- polynomial sample size
  (Jukes-Cantor, Kimura's, paralinear, LogDet)
- $\mathcal{O}\left(n^2\right)$ running time
  $\mathcal{O}\left(n\right)$ work space
- good experimental performance on large divergent trees

$\rightarrow$ fastest algorithm with polynomial sample size

`http://www.cs.yale.edu/~csuros-miklos/papers/`