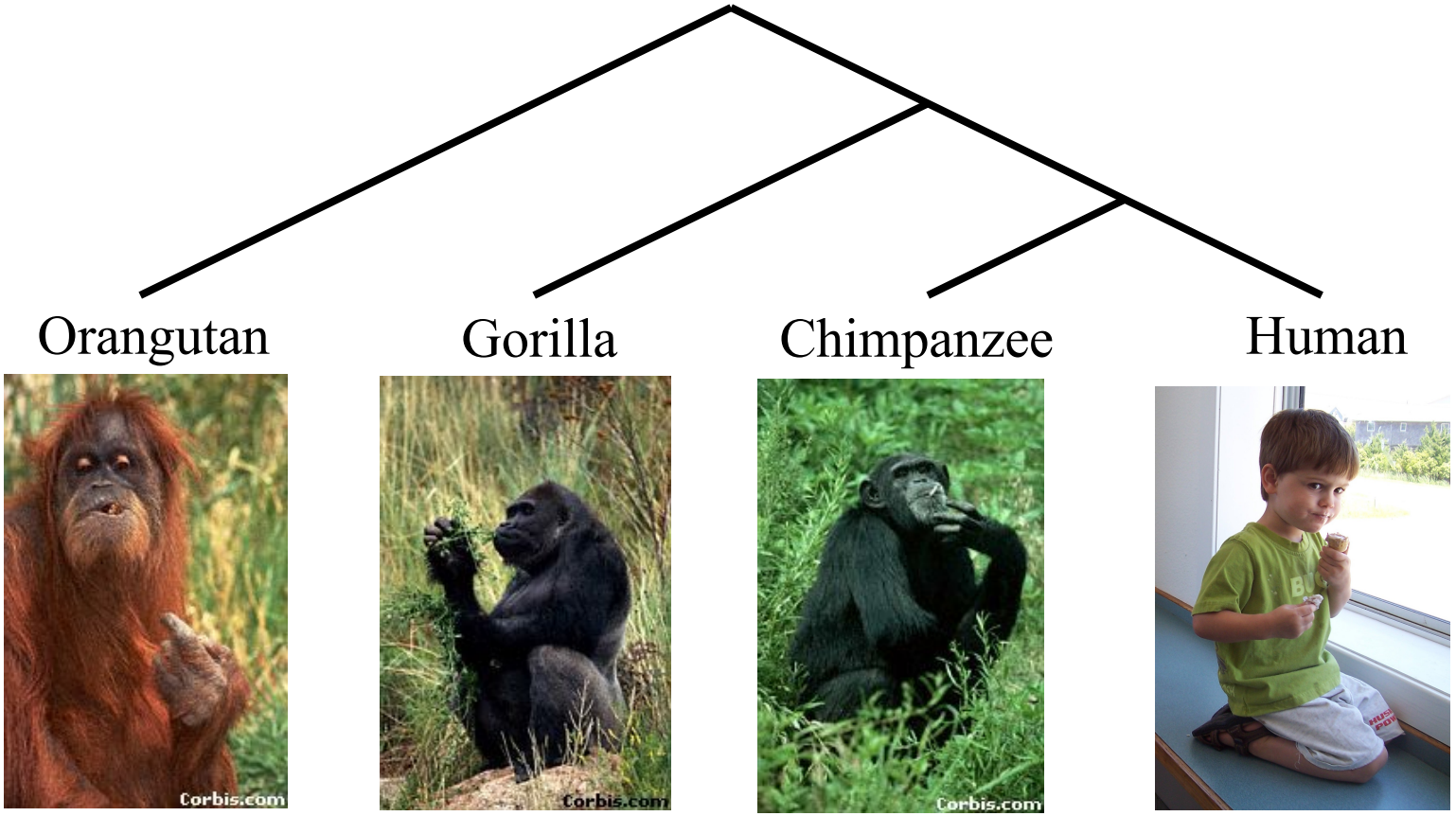
Multiple Sequence Alignment Methods

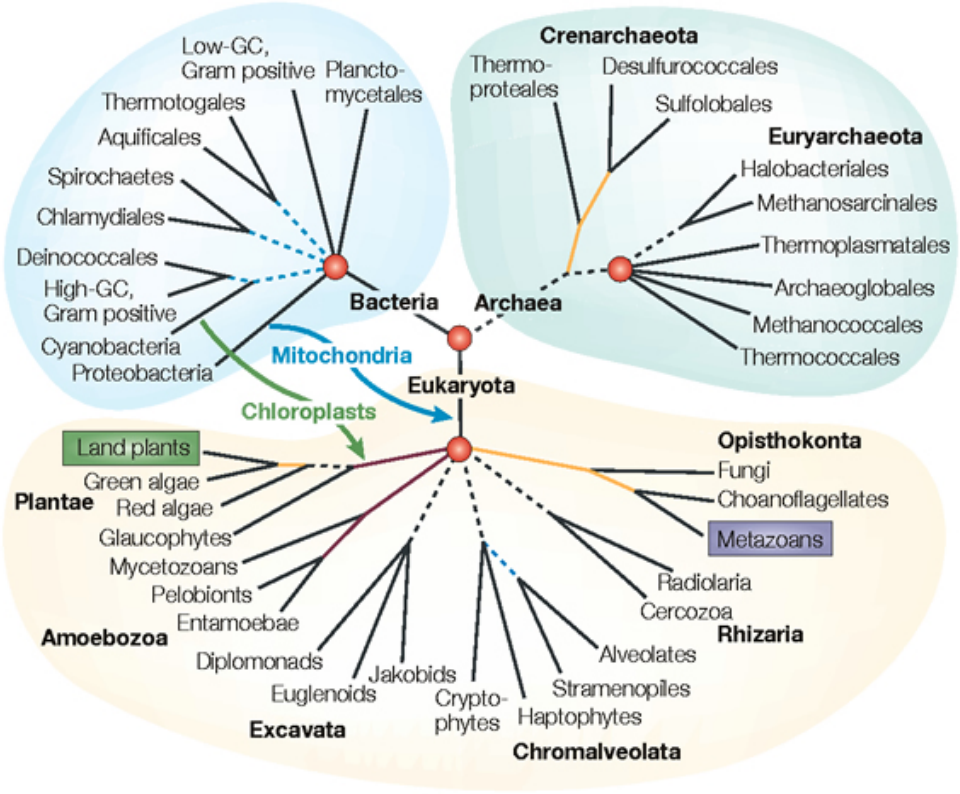
Tandy Warnow
Departments of Bioengineering and
Computer Science
http://tandy.cs.illinois.edu

Species Tree



From the Tree of the Life Website,
University of Arizona

Constructing the Tree of Life: Hard Computational Problems



Nature Reviews | Genetics

NP-hard problems

Large datasets

- 100,000+ sequences
- thousands of genes

"Big data" complexity:

- model misspecification
- fragmentary sequences
- errors in input data
- streaming data

Phylogenomic pipeline

Select taxon set and markers

Gather and screen sequence data, possibly identify orthologs

Compute multiple sequence alignments for each locus

Compute species tree or network:

- Compute gene trees on the alignments and combine the estimated gene trees, OR
- Estimate a tree from a concatenation of the multiple sequence alignments

Get statistical support on each branch (e.g., bootstrapping)

Estimate dates on the nodes of the phylogeny

Use species tree with branch support and dates to understand biology

Phylogenomic pipeline

Select taxon set and markers

Gather and screen sequence data, possibly identify orthologs

Compute multiple sequence alignments for each locus

Compute species tree or network:

- Compute gene trees on the alignments and combine the estimated gene trees, OR
- Estimate a tree from a concatenation of the multiple sequence alignments

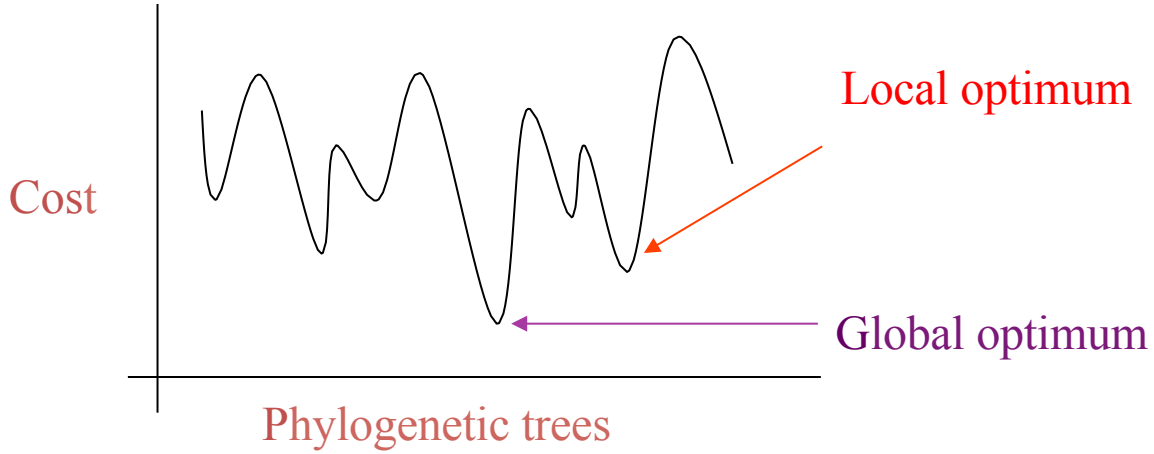
Get statistical support on each branch (e.g., bootstrapping)

Estimate dates on the nodes of the phylogeny

Use species tree with branch support and dates to understand biology

Phylogenetic reconstruction methods

1 Hill-climbing heuristics for hard optimization criteria (Maximum Parsimony and Maximum Likelihood)



2 Polynomial time distance-based methods: Neighbor Joining, FastME, etc.

3. Bayesian methods

Solving maximum likelihood (and other hard optimization problems) is… unlikely

# of Taxa	# of Unrooted Trees
4	3
5	15
6	105
7	945
8	10395
9	135135
10	2027025
20	2.2×10^{20}
100	4.5×10^{190}
1000	2.7×10^{2900}

The Real Problem!

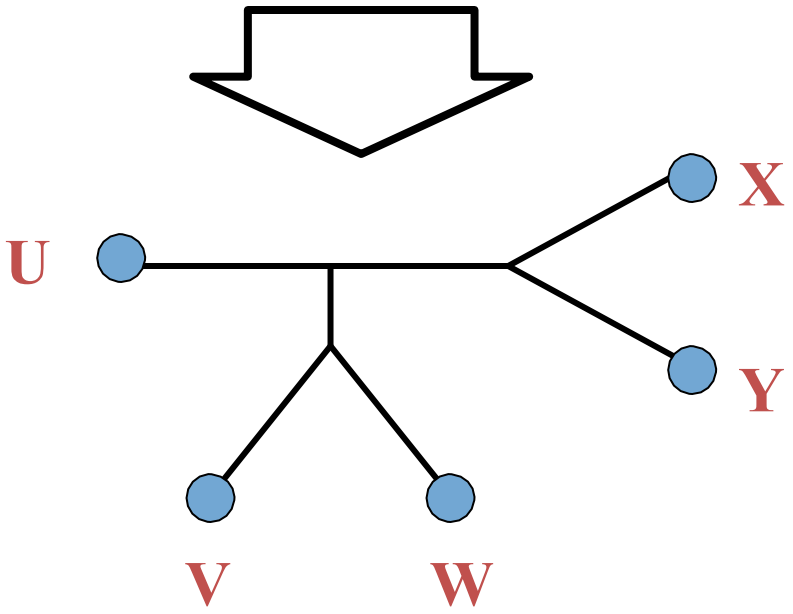
U AGGGCATGA

V AGAT

W TAGACTT

X TGCACAA

Y TGCGCTT



Input: unaligned sequences

```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

Phase 1: Alignment

```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



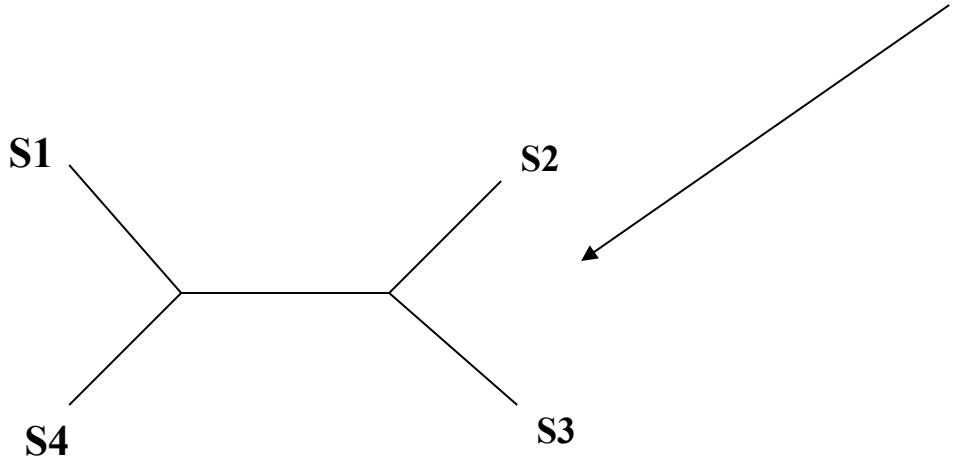
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

Phase 2: Construct tree

```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

→

```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```



Two-phase estimation

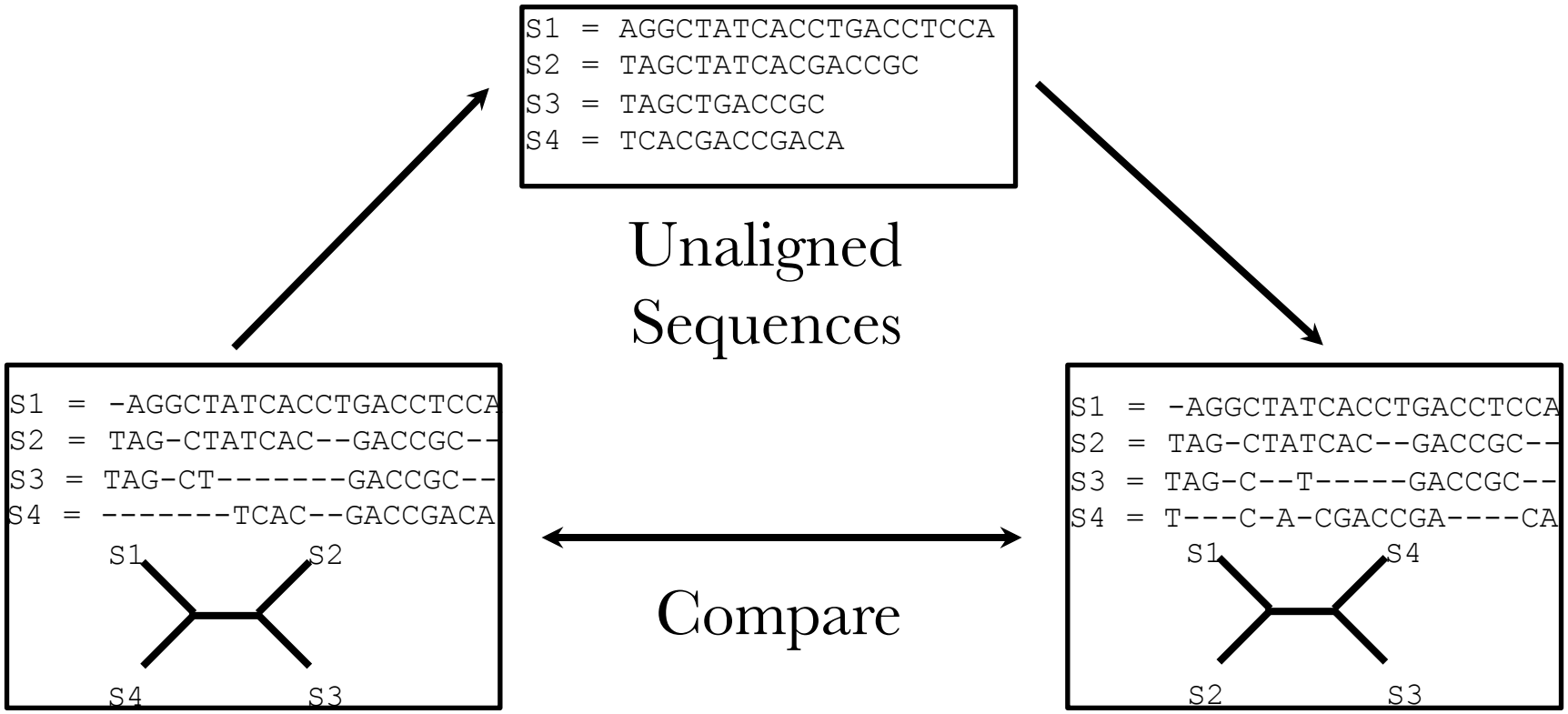
Alignment methods

- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- Probalign
- MAFFT
- Muscle
- Di-align
- T-Coffee
- Prank (PNAS 2005, Science 2008)
- Opal (ISMB and Bioinf. 2007)
- *FSA (PLoS Comp. Bio. 2009)*
- *Infernal (Bioinf. 2009)*
- Etc.

Phylogeny methods

- Bayesian MCMC
- Maximum parsimony
- Maximum likelihood
- Neighbor joining
- FastME
- UPGMA
- Quartet puzzling
- Etc.

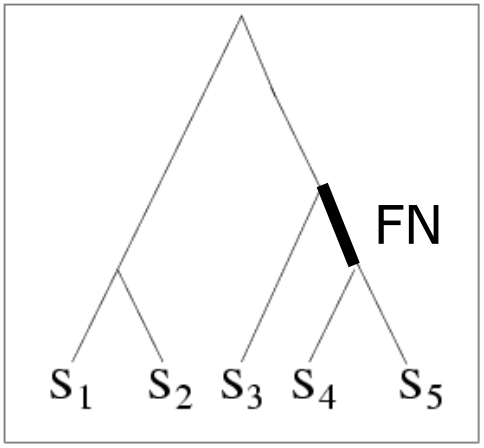
Simulation Studies



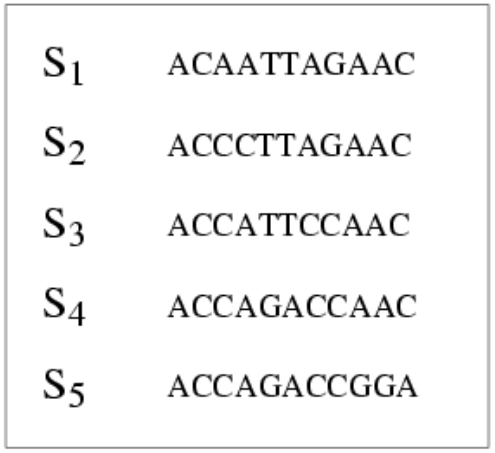
True tree and alignment

Estimated tree and alignment

Quantifying Error



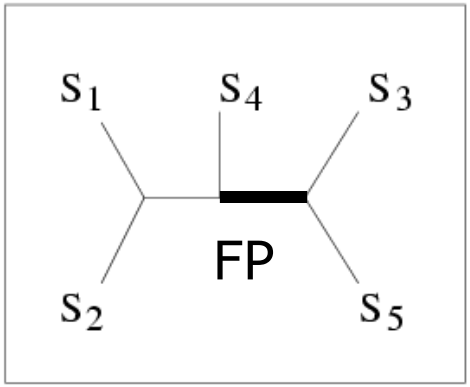
TRUE TREE



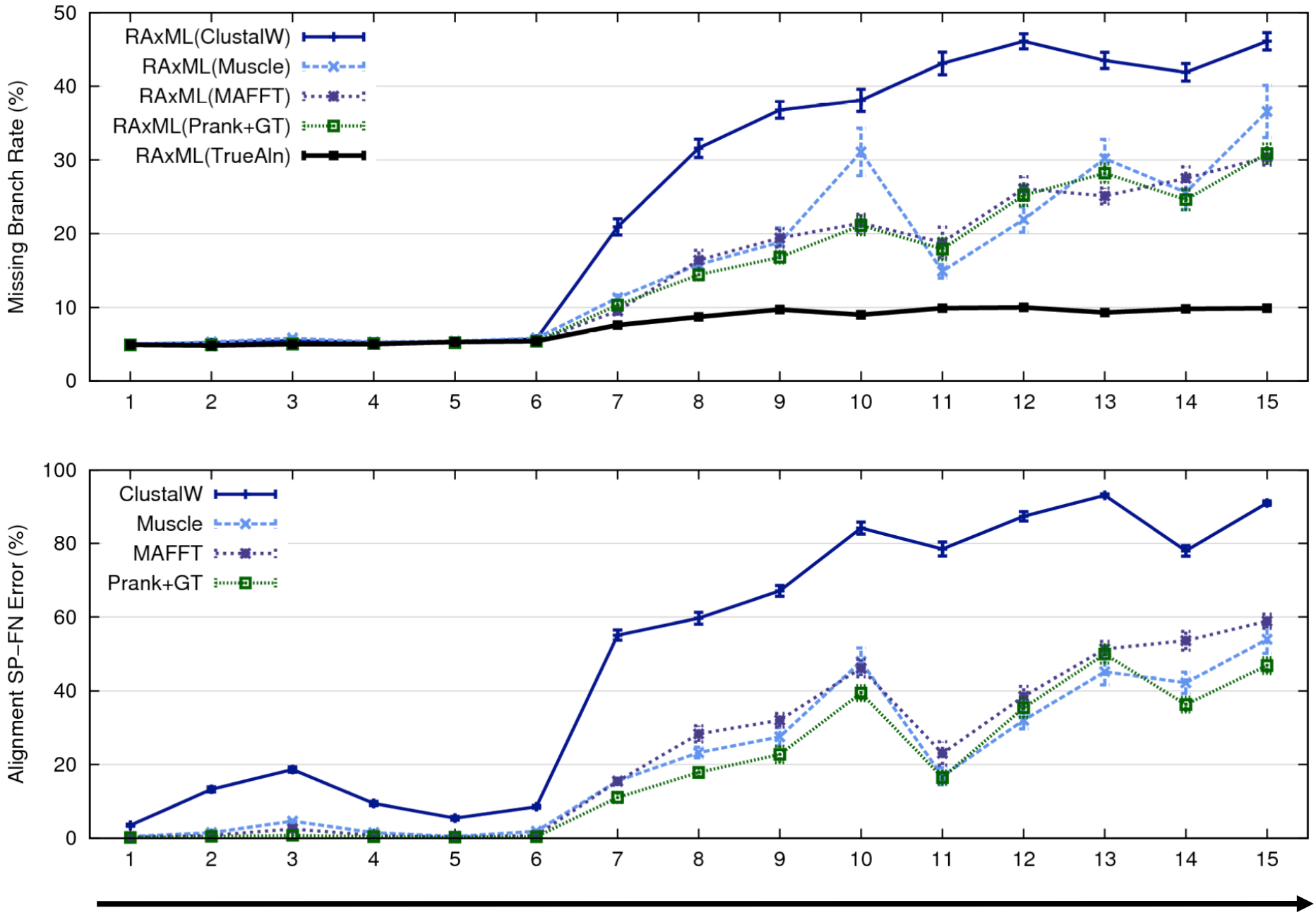
DNA SEQUENCES

FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate



INFERRED TREE


Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Alignment SP–FN Error (%)
ClustalW
Muscle
MAFFT
Prank+GT
1000 taxon models, ordered by difficulty (Liu et al., 2009)

Major Challenges

- **Phylogenetic analyses:** standard methods have *poor accuracy* on even moderately large datasets, and the most accurate methods are enormously *computationally intensive* (weeks or months, high memory requirements)
- **Multiple sequence alignment**: key step for many biological questions (protein structure and function, phylogenetic estimation), but few methods can run on large datasets. Alignment accuracy is generally poor for large datasets with high rates of evolution.

Multiple Sequence Alignment (MSA): *another grand challenge*[1]

```
S1 = AGGCTATCACCTGACCTCCA          S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC              S2 = TAG-CTATCAC--GACCGC--
S3 = TAGCTGACCGC                   S3 = TAG-CT-------GACCGC--
   …                                  …
Sn = TCACGACCGACA          ⟶      Sn = -------TCAC--GACCGACA
```

Novel techniques needed for scalability and accuracy

NP-hard problems and large datasets

Current methods do not provide good accuracy

Few methods can analyze even moderately large datasets

Many important applications besides phylogenetic estimation

[1] Frontiers in Massive Data Analysis, National Academies Press, 2013

Generalized Tree Alignment

- Input: set S of sequences and function for gap costs.
- Output: Tree T and sequences at the internal nodes to minimize the total cost on the tree (sum of edit distances on the edges). Note that the output also defines a multiple sequence alignment!

NP-hard to find the best solution!

Software for GTA (treelength optimization)

- POY is the most well-known method for co-estimating alignments and trees using treelength criteria (however – note that the developers of POY say to ignore the alignment and only use the tree).
- BeeTLe (Better Tree Length) is a heuristic that is guaranteed to always be as least as accurate as POY for the treelength criterion.
- The accuracy of the final tree depends on the edit distance formulation – as noted by several studies. Affine gap penalties are more biologically realistic than simple gap penalties.

Gap penalties

- Simple gap penalties: cost of a gap of length L is cL for some constant c>0
- Affine gap penalties: cost of a gap of length L is cL + c', for some pair of constants c and c'
- Other gap penalties are also possible (e.g., cost could be L + c log L)

Treelength questions

- Is BeeTLe actually better than POY at the treelength problem (as promised)?
- Is it better to use affine than simple gap penalties?
- How accurate are the alignments?
- How accurate are the trees, compared to
  - Maximum Parsimony analyses of good alignments
  - Maximum Likelihood analyses of good alignments

Treelength questions

- Is BeeTLe actually better than POY at the treelength problem (as promised)? – YES!
- Is it better to use affine than simple gap penalties?
- How accurate are the alignments?
- How accurate are the trees, compared to
  - Maximum Parsimony analyses of good alignments
  - Maximum Likelihood analyses of good alignments

Alignment Error/Accuracy

- SPFN: percentage of homologies in the true alignment that are *not* recovered (false negative homologies)
- SPFP: percentage of homologies in the estimated alignment that are false (false positive homologies)

- TC: total number of columns correctly recovered
- SP-score: percentage of homologies in the true alignment that are recovered
- Pairs score: 1-(avg of SP-FN and SP-FP)

How well do POY and BeeTLe do, compared to other MSA methods?

- We simulated sequences down evolutionary trees with substitutions, insertions, and indels.
- We computed alignments on each dataset using multiple techniques (e.g., POY, BeeTLe, Muscle, Mafft, etc.)
- We computed alignment errors using SPFN

See Liu, K. and T. Warnow, PLOS One 7(3). 2012, "Treelength optimization for phylogeny estimation"

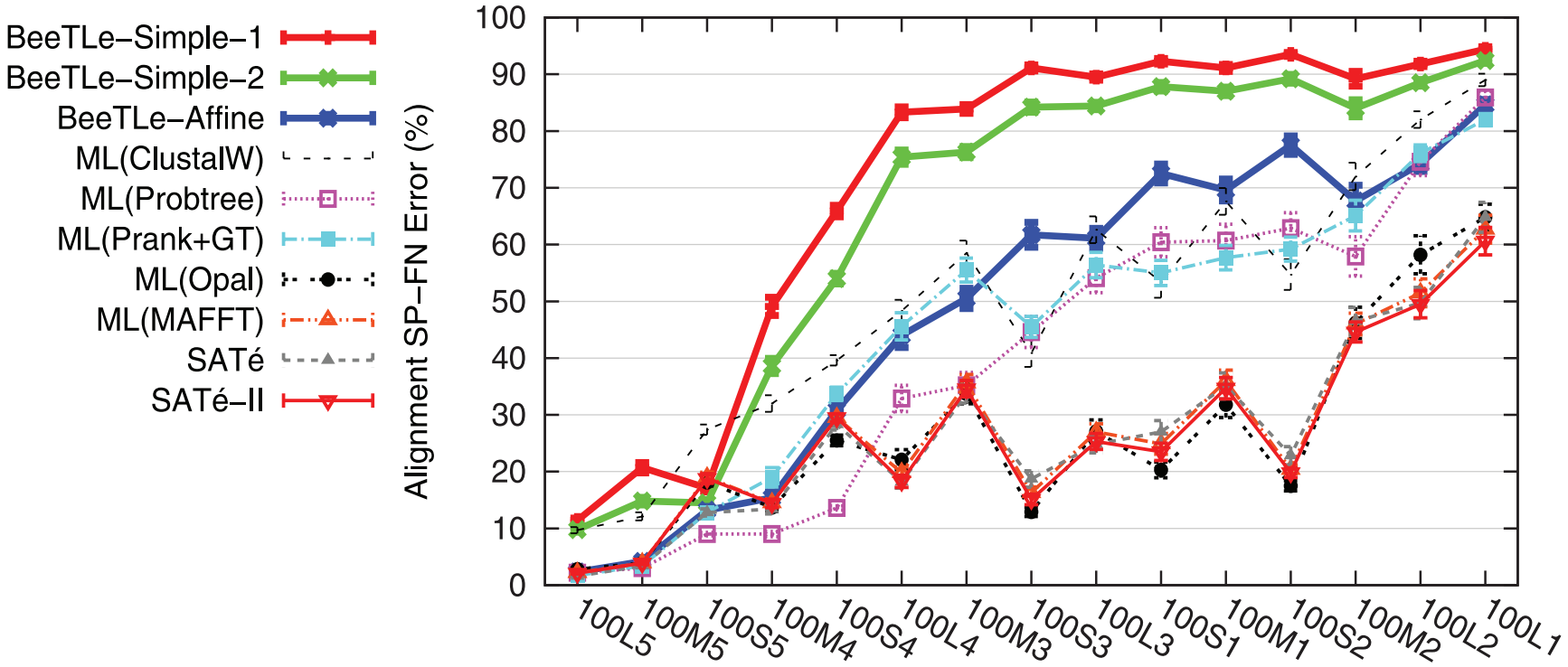


Figure 5. Alignment SP-FN error of different methods on 100-taxon model conditions. Averages and standard error bars are shown; $n = 20$ for each reported value.
doi:10.1371/journal.pone.0033104.g005

Simulated 100-sequence DNA datasets with varying rates of evolution
Results from Liu and Warnow, PLoS ONE 2012

Observations

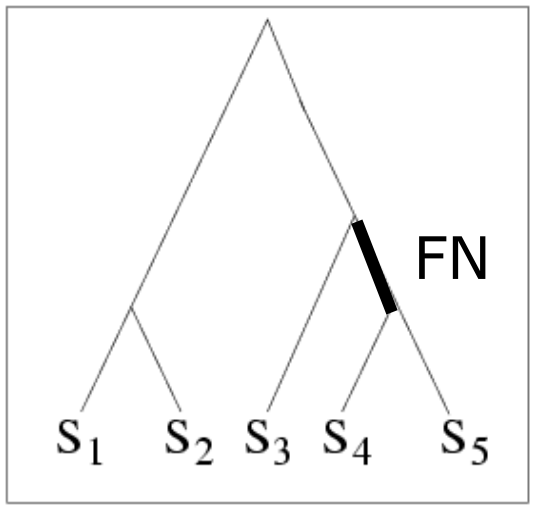
Choice of gap penalty has big impact on BeeTLE – with affine gap penalties much better than simple gap penalties.

Even so, alignments produced by BeeTLe are not nearly as accurate as alignments produced by other methods.

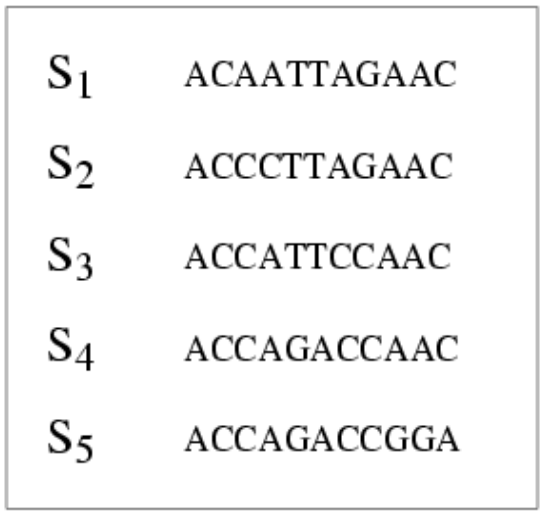
The best accuracy is obtained using Opal, MAFFT, SATe, or SATe-II.

Tree Estimation using Treelength

- Beetle produce phylogenetic (evolutionary) trees as well as alignments.
- Given an alignment, we can compute phylogenetic trees on multiple sequence alignments using many methods.
- Examples of tree estimation methods:
  - Maximum Parsimony
  - Maximum Likelihood

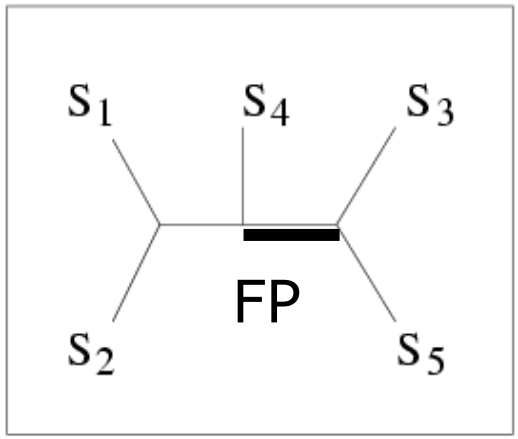


TRUE TREE



DNA SEQUENCES

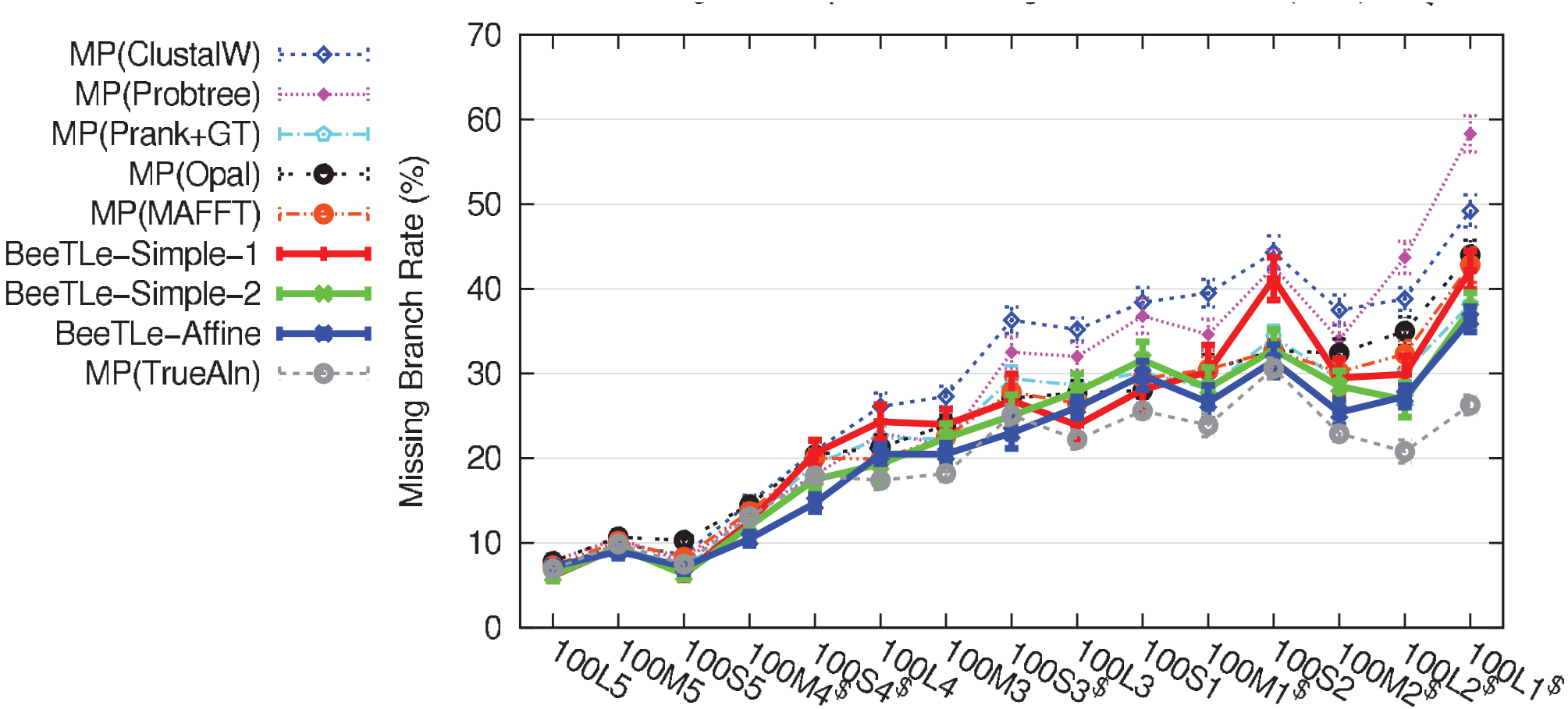
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)



INFERRED TREE

50% error rate

Maximum Parsimony (MP) on different alignments



Simulated 100-sequence DNA datasets with varying rates of evolution
Results from Liu and Warnow, PLoS ONE 2012

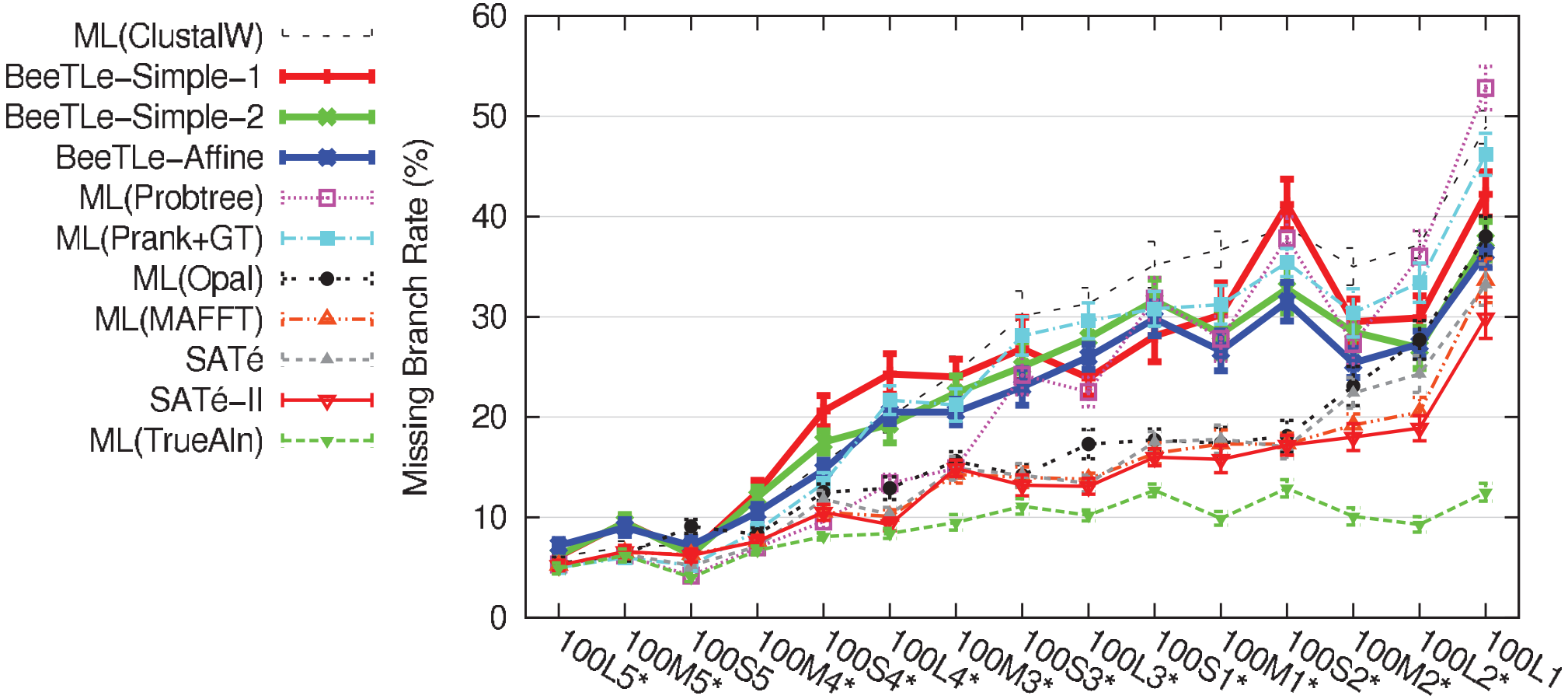
Observations, MP trees

MP trees computed using different alignment methods, compared to BeeTLe trees

- Choice of gap penalty still has an impact, with affine gap penalties better than simple gap penalties.
- BeeTLe-affine better than the best MP trees

Not examined: MP on BeeTLe alignments and MP on SATe or SATe-II alignments.

Maximum Likelihood (ML) on different alignments



Simulated 100-sequence DNA datasets with varying rates of evolution
Results from Liu and Warnow, PLoS ONE 2012

Observations, ML trees

ML trees computed using different alignment methods, compared to BeeTLe trees

- Choice of gap penalty still has an impact, with affine gap penalties better than simple gap penalties.
- BeeTLe-affine worse than nearly all ML trees (exception is ClustalW).
- Best trees obtained using SATe, SATe-II, Opal and MAFFT

Not examined: ML on BeeTLe alignments

Overall Observations

Maximum Likelihood (ML) better at estimating trees than Maximum Parsimony (MP).

The best alignments are obtained using SATe or SATe-2. Opal and MAFFT are also good.

The best trees are obtained using ML on good alignments.

Generalized Tree Alignment – not that good for either alignment or tree estimation.

But…

It is intriguing that BeeTLe alignments are so bad, while trees aren't that bad…

The edit distance function has an impact, so maybe a better edit distance function would result in good alignments and trees.

In other words, we don't know the real answer yet.