Computational and mathematical challenges involved in very large-scale phylogenetics

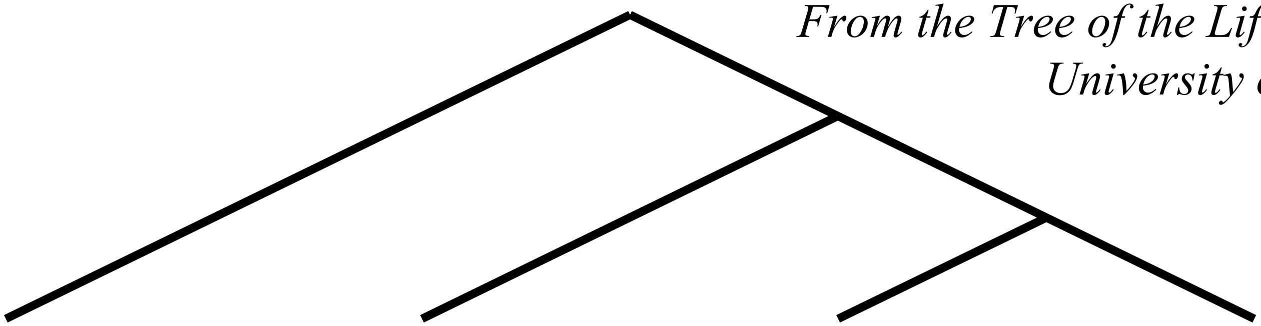
Tandy Warnow

The University of Texas at Austin

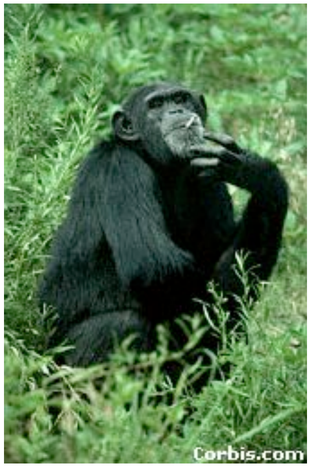
CIPRES

Species phylogeny

From the Tree of the Life Website, University of Arizona



Orangutan Gorilla Chimpanzee Human



How did life evolve on earth?

- Courtesy of the Tree of Life project

An international effort to understand how life evolved on earth

Biomedical applications: drug design, protein structure and function prediction, biodiversity

Phylogenetic estimation is a “Grand Challenge”: millions of taxa, NP-hard optimization problems

The CIPRES Project

(Cyber-Infrastructure for Phylogenetic Research)
www.phylo.org

This project is funded by the NSF under a Large ITR grant

- ***ALGORITHMS and SOFTWARE: scaling to millions of sequences (open source, freely distributed)***
- ***MATHEMATICS/PROBABILITY/STATISTICS: Obtaining better mathematical theory under complex models of evolution***
- ***DATABASES: Producing new database technology for structured data, to enable scientific discoveries***
- ***SIMULATIONS: The first million taxon simulation under realistically complex models***
- ***OUTREACH: Museum partners, K-12, general scientific public***
- ***PORTAL available to all researchers***

Step 1: Gather data

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

Step 2: Multiple Sequence Alignment

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

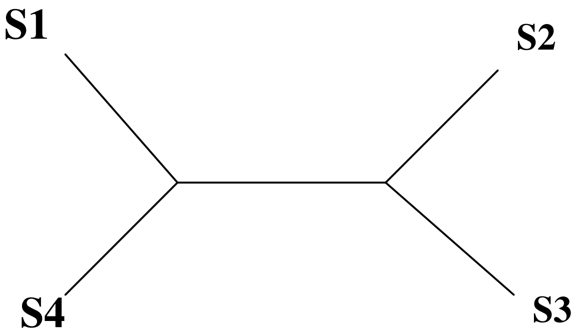
→

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

Step 3: Construct tree

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

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



Performance criteria

- Estimated alignments are evaluated with respect to the ***true alignment***. Studied both in simulation and on real data.
- Estimated trees are evaluated for “topological accuracy” with respect to the ***true tree.*** Typically studied in simulation.
- Methods for these problems can also be evaluated with respect to an optimization criterion (e.g., maximum likelihood score) as a function of running time. Typically studied on real data.

Observations

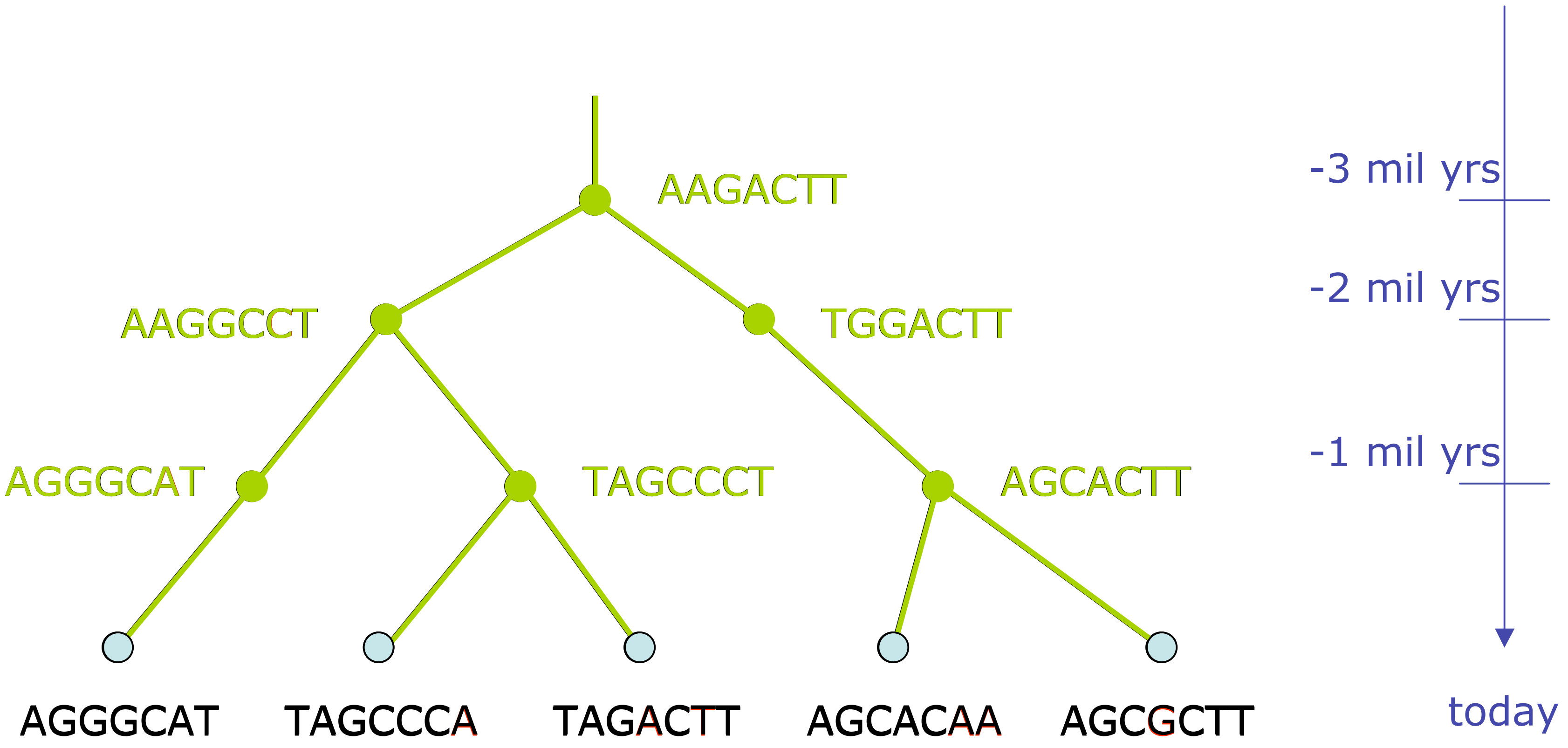
- The best current multiple sequence alignment methods can produce **highly inacccurate alignments on large datasets** (with the result that trees estimated on these alignments are also inaccurate).
- The fast (polynomial time) methods produce **highly inaccurate trees** for many datasets.
- Heuristics for NP-hard optimization problems often produce highly accurate trees, but can take **months** to reach solutions on large datasets.

This talk

- Part 1: Improving the topological accuracy of polynomial time phylogeny reconstruction methods (and *absolute fast converging* methods)
- Part 2: Improving heuristics for NP-hard optimization problems (getting better solutions faster)
- Part 3: Simultaneous Alignment and Tree estimation (SATe)
- Part 4: Conclusions

Part 1: Improving polynomial time methods (and absolute fast converging methods)

DNA Sequence Evolution



AAGACTT

AAGGCCT

TGGACTT

AGGGCAT

TAGCCCT

AGCACTT

AGGGCAT TAGCCCA TAGACTT AGCACAA AGCGCTT

-3 mil yrs

-2 mil yrs

-1 mil yrs

today

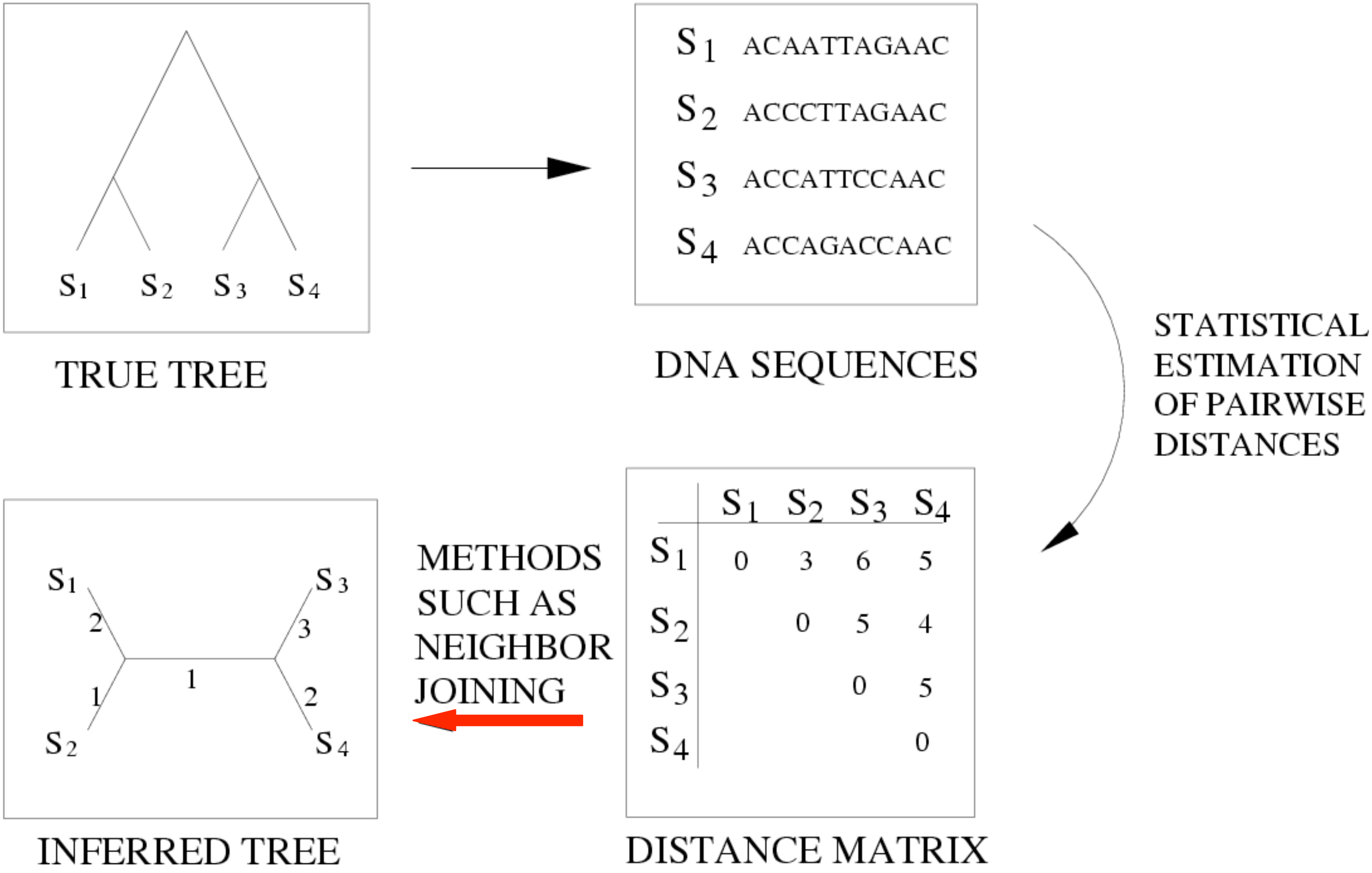
Markov models of single site evolution

Simplest (Jukes-Cantor):

- The model tree is a pair (T,{e,p(e)}), where T is a rooted binary tree, and p(e) is the probability of a substitution on the edge e.
- The state at the root is random.
- If a site changes on an edge, it changes with equal probability to each of the remaining states.
- The evolutionary process is Markovian.

More complex models (such as the General Markov model) are also considered, often with little change to the theory.

Distance-based Phylogenetic Methods



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

DNA SEQUENCES

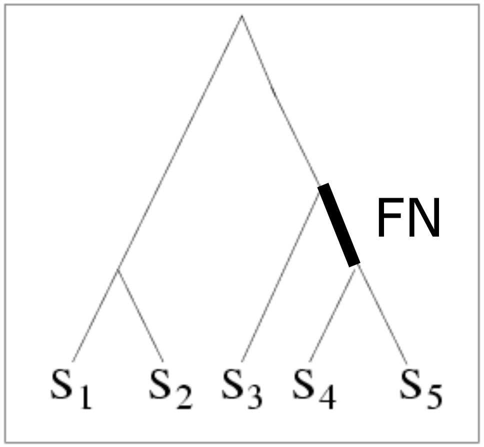
STATISTICAL
ESTIMATION
OF PAIRWISE
DISTANCES

	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

DISTANCE MATRIX

METHODS
SUCH AS
NEIGHBOR
JOINING

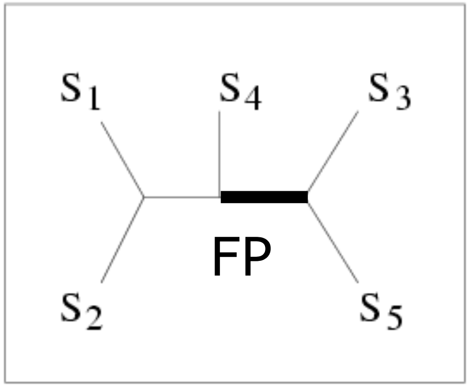
INFERRED TREE



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

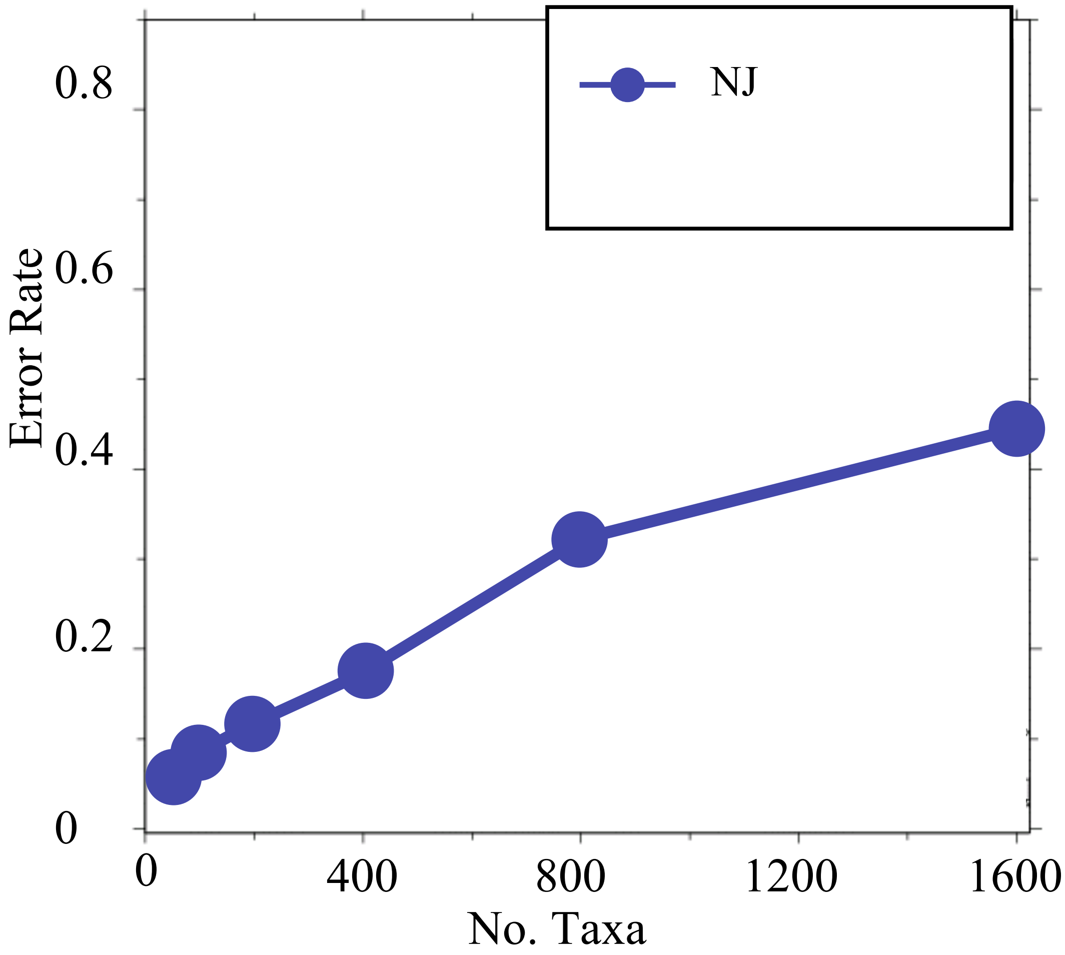


INFERRED TREE

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

50% error rate

Neighbor joining has poor performance on large diameter trees *[Nakhleh et al. ISMB 2001]*

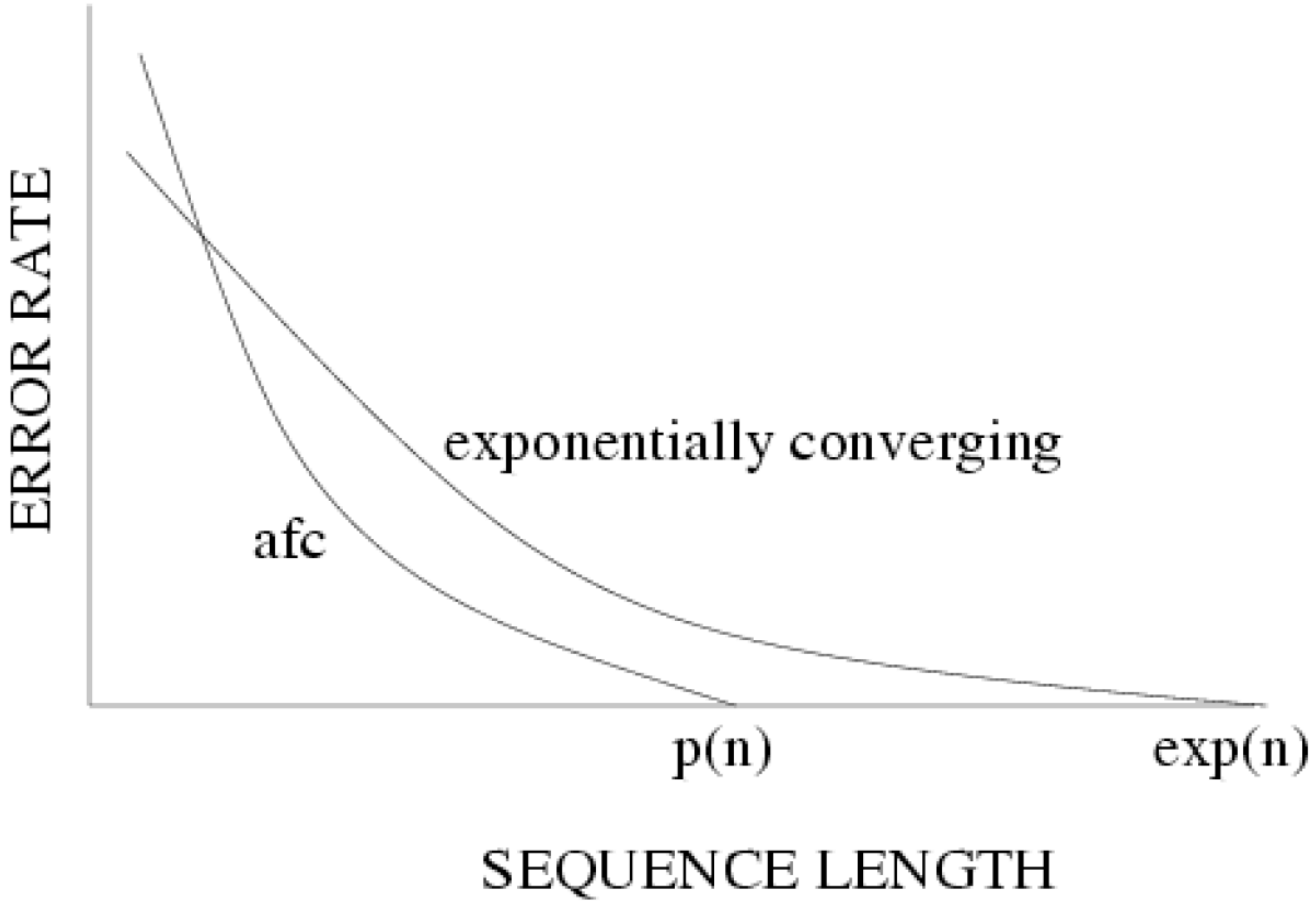


Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

Error rates reflect proportion of incorrect edges in inferred trees.

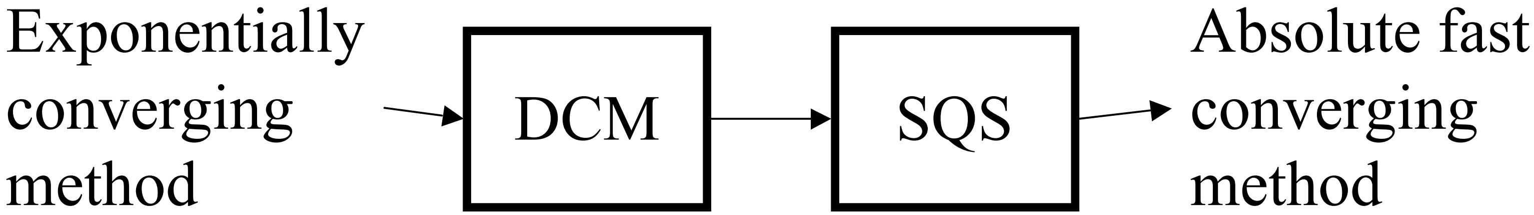
- Theorem: Neighbor joining (and some other distance-based methods) will return the true tree with high probability provided sequence lengths are **exponential** in the diameter of the tree (Erdos et al., Atteson).

Statistical consistency, exponential convergence, and absolute fast convergence (afc)



DCM1

Warnow, St. John, and Moret, SODA 2001



Exponentially converging method

DCM

SQS

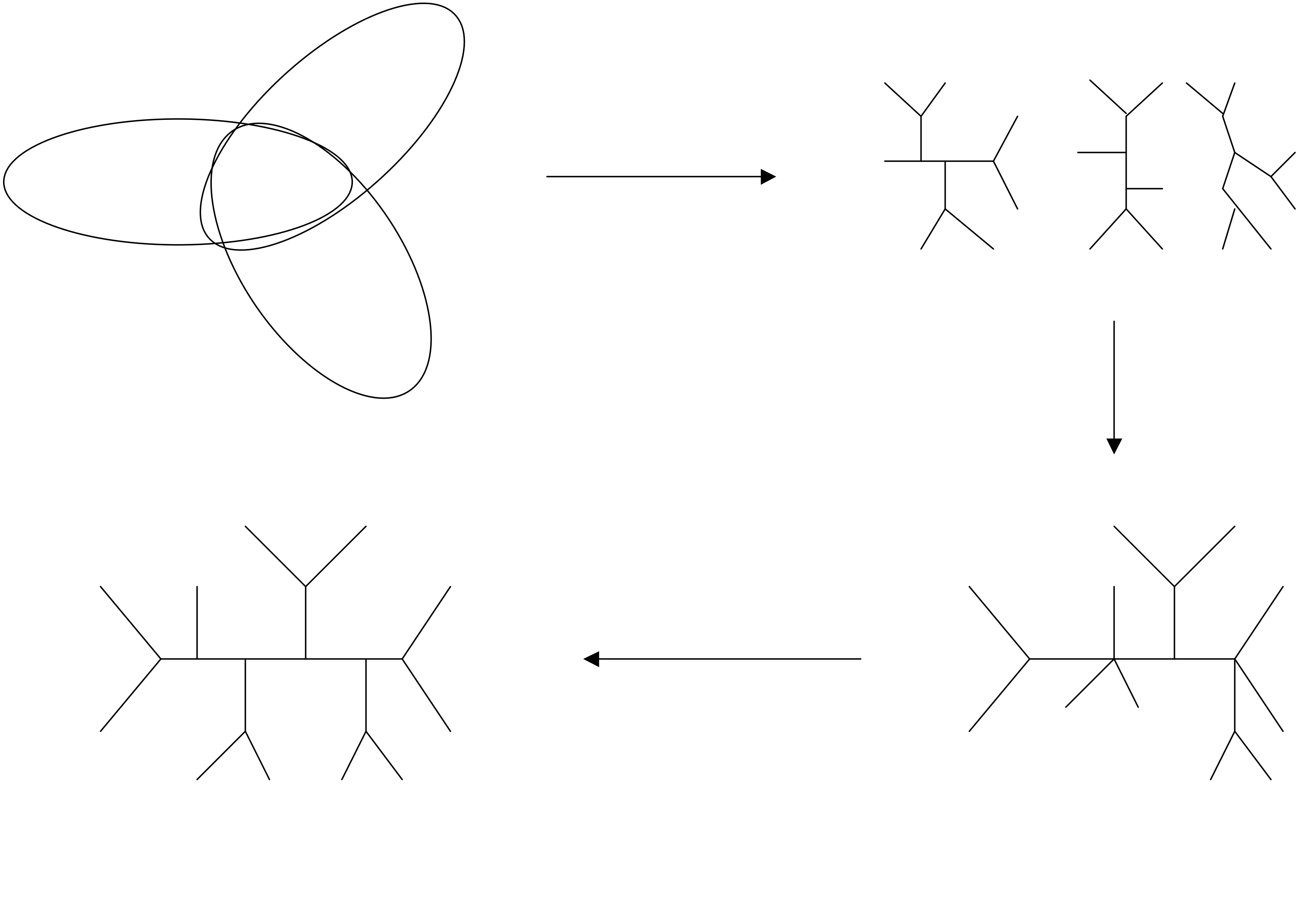
Absolute fast converging method

- A two-phase procedure which reduces the sequence length requirement of methods. The DCM phase produces a collection of trees, and the SQS phase picks the “best” tree.
- The “base method” is applied to subsets of the original dataset. When the base method is NJ, you get DCM1-NJ.

Graph-theoretic divide-and-conquer (DCM's)

- Define a triangulated (i.e. **chordal**) graph so that its vertices correspond to the input taxa
- Compute a **decomposition** of the graph into overlapping subgraphs, thus defining a decomposition of the taxa into overlapping subsets.
- Apply the "**base method**" to each subset of taxa, to construct a subtree
- **Merge** the subtrees into a single tree on the full set of taxa.

DCM (cartoon)



Some properties of chordal graphs

- Every chordal graph has at most n maximal cliques, and these can be found in polynomial time: ***Maxclique*** decomposition.
- Every chordal graph has a vertex separator which is a maximal clique: ***Separator-component*** decomposition.
- Every chordal graph has a perfect elimination scheme: ***enables us to merge correct subtrees and get a correct supertree back, if subtrees are big enough.***

DCM1 Decompositions

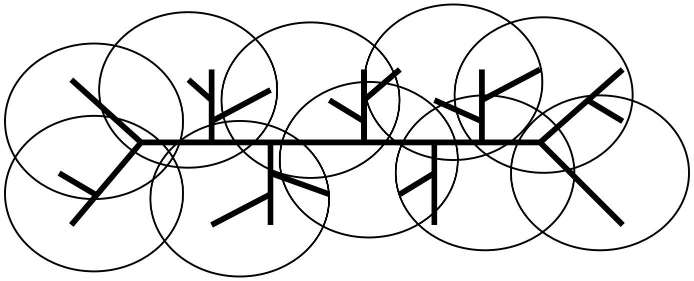
Input: Set S of sequences, distance matrix d, threshold value $q \in \{d_{ij}\}$

1. Compute threshold graph

$$G_q = (V, E), V = S, E = \{(i, j) : d(i, j) \leq q\}$$

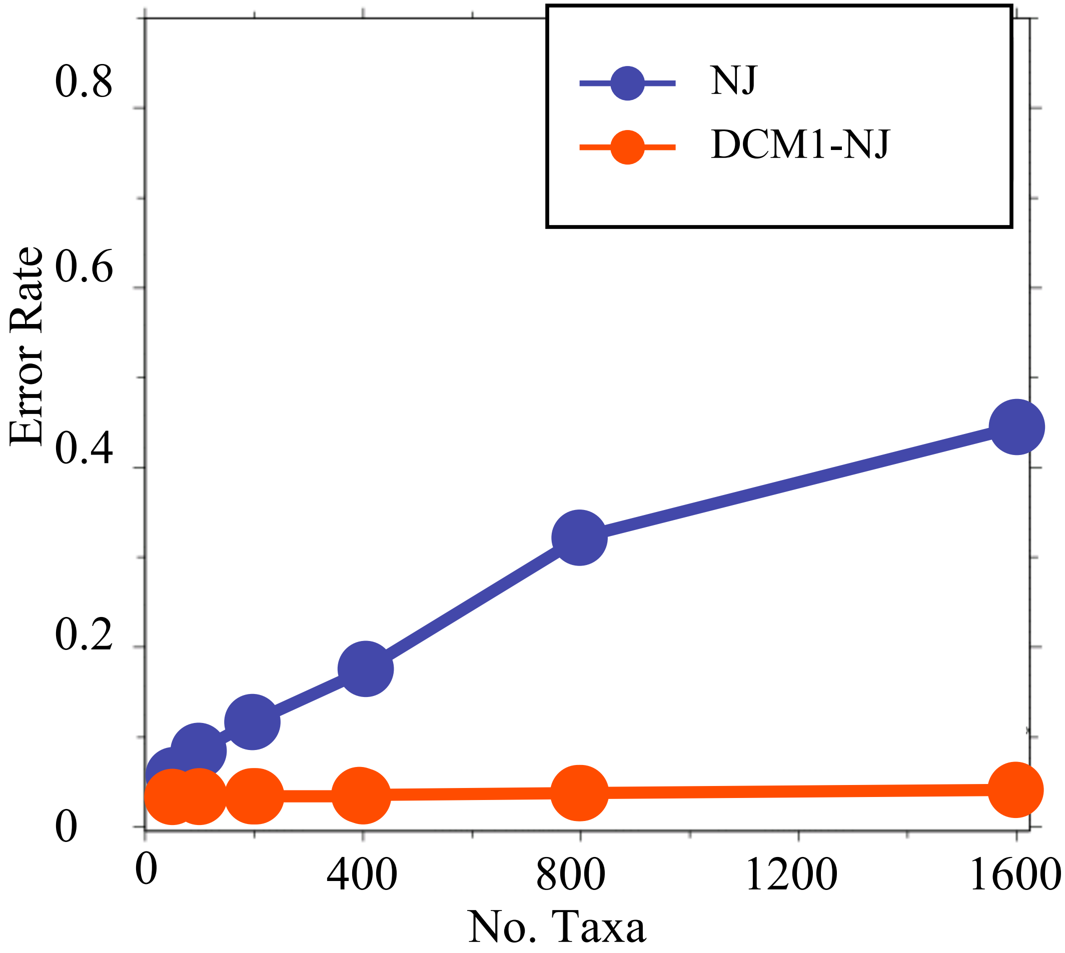
2. Perform minimum weight triangulation (note: if d is an additive matrix, then the threshold graph is provably chordal).

DCM1 decomposition : Compute maximal cliques



DCM1-boosting distance-based methods

[Nakhleh et al. ISMB 2001]



Theorem: DCM1-NJ converges to the true tree from polynomial length sequences

However,

- The best phylogenetic accuracy tends to be from computationally intensive methods (and most molecular phylogeneticists prefer these methods).
- Unfortunately, these approaches can take weeks or more, just to reach decent local optima.

- Conclusion: ***We need better heuristics for NP-hard optimization methods!***

Part 2: Improved heuristics for NP-hard optimization problems

- Rec-I-DCM3: Roshan, Williams, Moret, and Warnow
- Part of the CIPRES software distribution and portal

Standard problem: Maximum Parsimony (Hamming distance Steiner Tree)

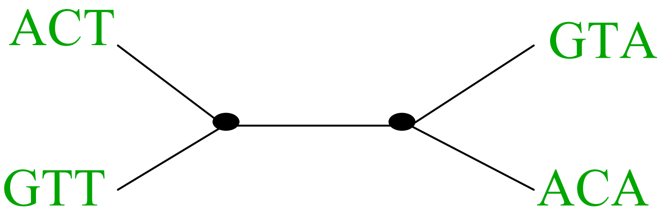
- **Input**: Set *S* of *n* aligned sequences of length k
- **Output**: A phylogenetic tree *T*
  - leaf-labeled by sequences in *S*
  - additional sequences of length *k* labeling the internal nodes of *T*

such that $\sum_{(i,j)\in E(T)} H(i,j)$ is minimized.

Maximum parsimony (example)

- **Input**: Four sequences
  - ACT
  - ACA
  - GTT
  - GTA
- **Question**: which of the three trees has the best MP scores?

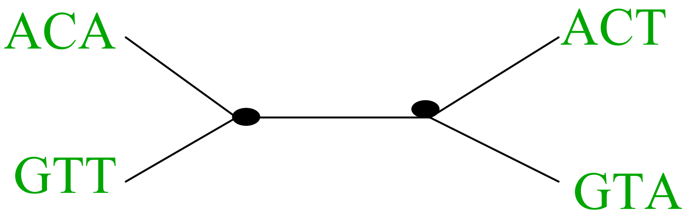
Maximum Parsimony


ACT

GTT

GTA

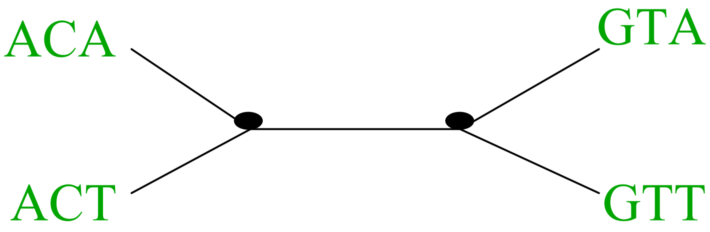
ACA


ACA

GTT

ACT

GTA

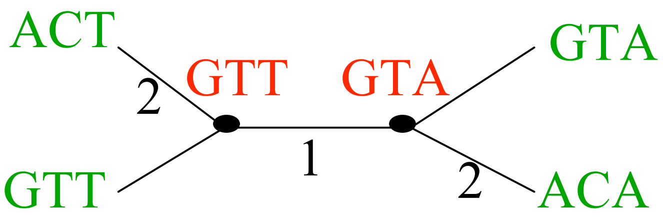

ACA

ACT

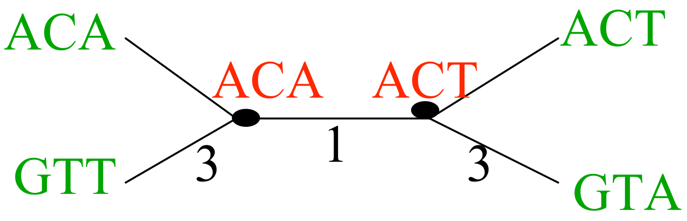
GTA

GTT

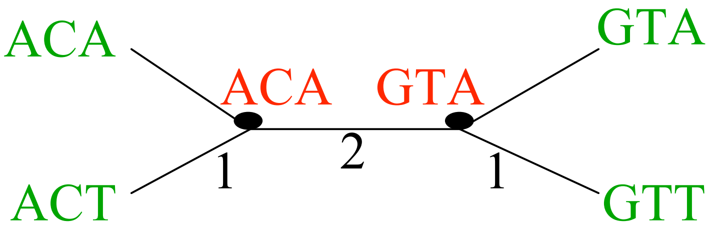
Maximum Parsimony



MP score = 5



MP score = 7

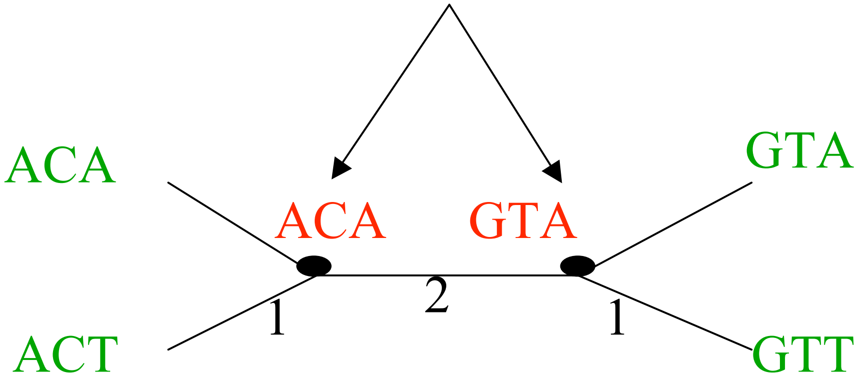


MP score = 4

Optimal MP tree

Maximum Parsimony: computational complexity

Optimal labeling can be computed in linear time O(nk)



MP score = 4

Finding the optimal MP tree is **NP-hard**

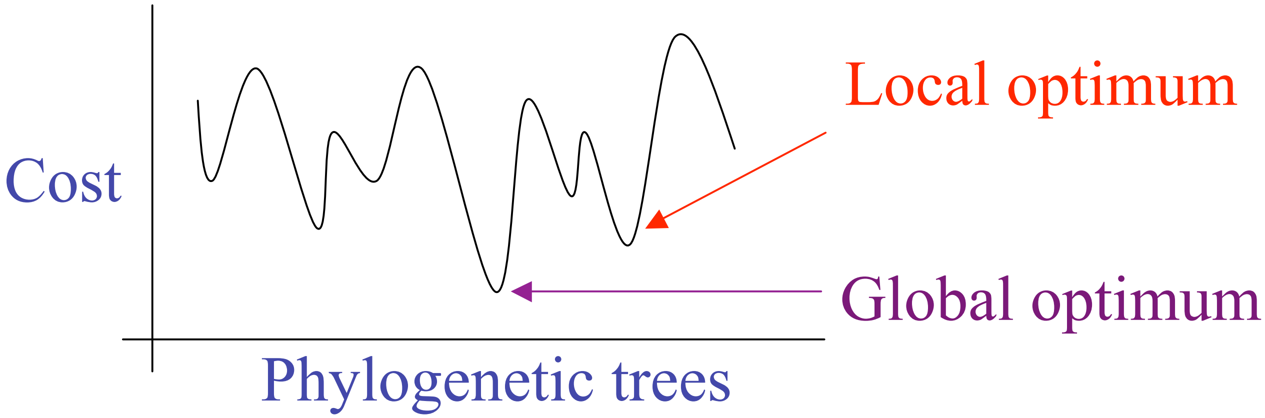
Maximum Likelihood (ML)

- Given: stochastic model of sequence evolution (e.g. Jukes-Cantor) and a set S of sequences
- Objective: Find tree T and parameter values so as to maximize the probability of the data.

Preferred by some systematists, but even harder than MP in practice.

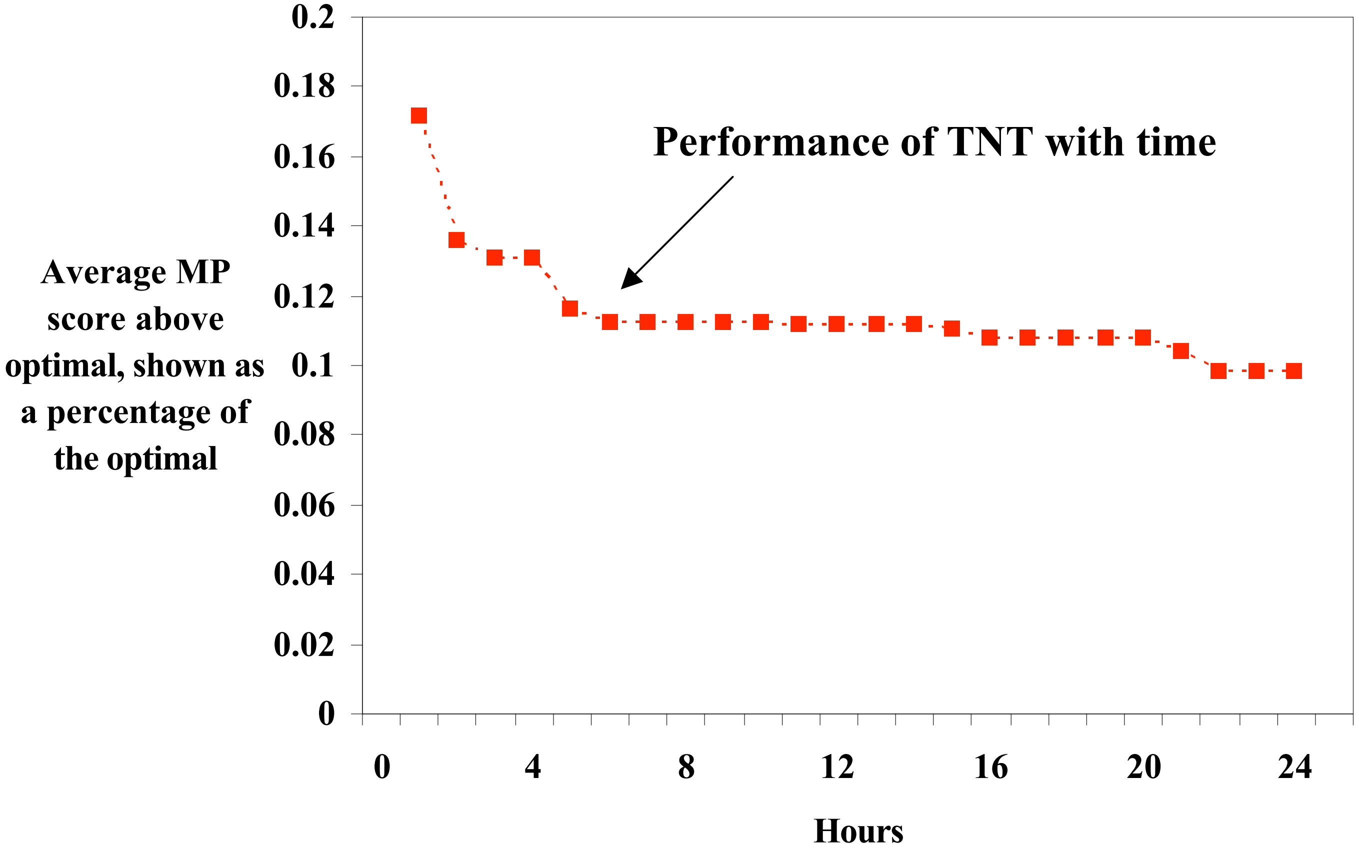
Approaches for “solving” MP (and other NP-hard problems in phylogeny)

1. Hill-climbing heuristics (which can get stuck in local optima)
2. Randomized algorithms for getting out of local optima
3. Approximation algorithms for MP (based upon Steiner Tree approximation algorithms).



Problems with current techniques for MP

Shown here is the performance of a TNT heuristic maximum parsimony analysis on a real dataset of almost 14,000 sequences. ("Optimal" here means ***best score to date***, using any method for any amount of time.) Acceptable error is below 0.01%.



Rec-I-DCM3: a new technique (Roshan et al.)

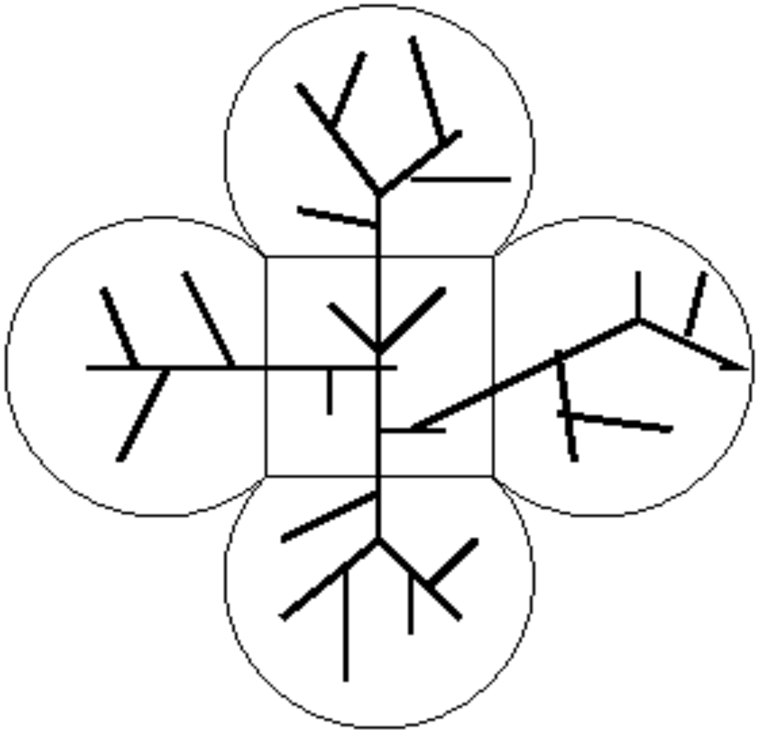
- Combines a new decomposition technique (DCM3) with recursion and iteration, to produce a novel approach for escaping local optima
- Demonstrated here on MP (maximum parsimony), but also implemented for ML and other optimization problems

The DCM3 decomposition

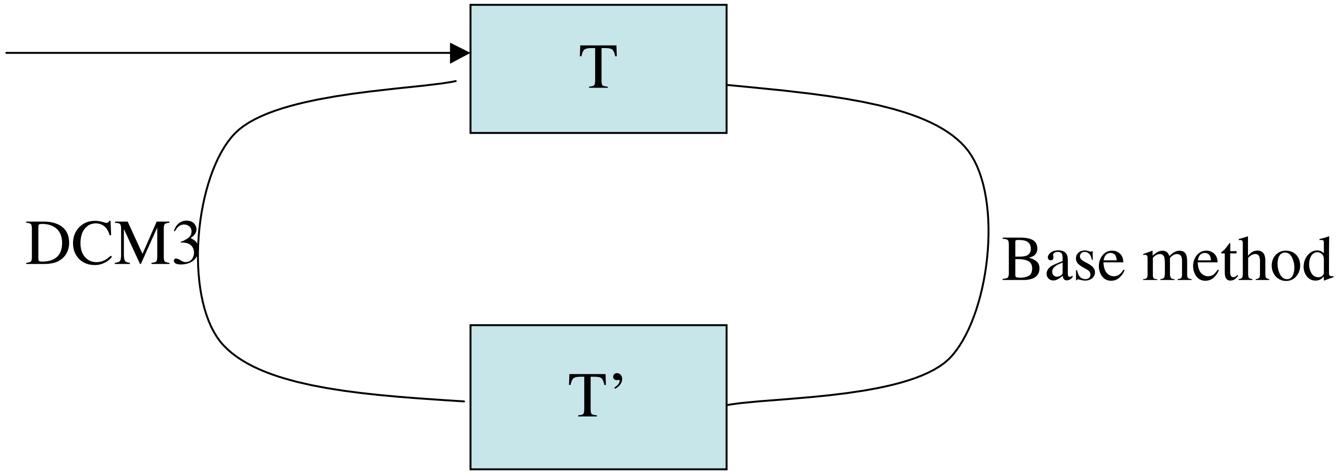
Input: Set *S* of sequences, and guide-tree *T*

1. Compute *short subtree* graph *G(S,T)*, based upon *T*
2. Find clique separator in the graph *G(S,T)* and form subproblems

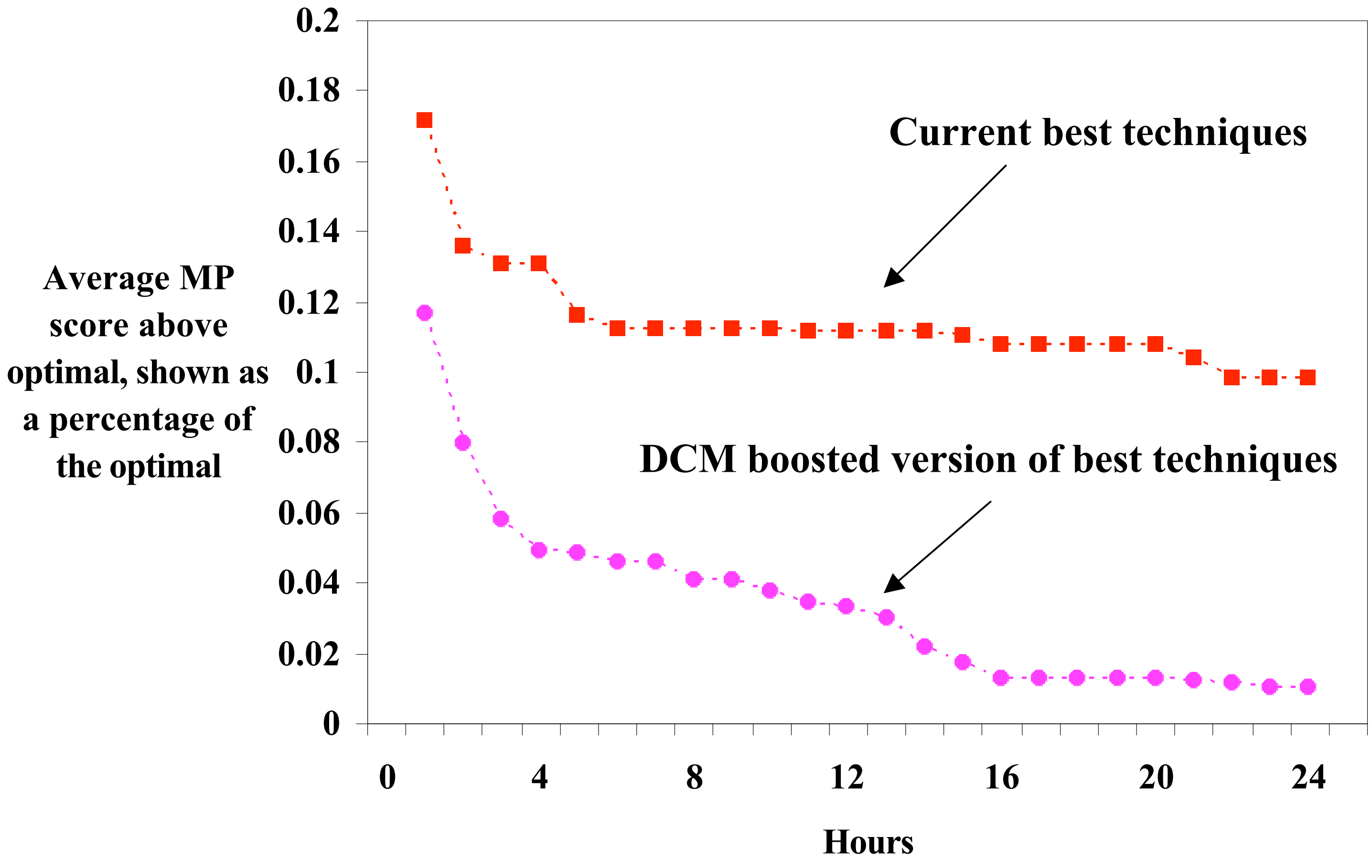
DCM3 decompositions
(1) can be obtained in **O(n)** time (the short subtree graph is **triangulated**)
(2) yield small subproblems
(3) can be used iteratively



Iterative-DCM3



Rec-I-DCM3 significantly improves performance (Roshan et al.)



Comparison of TNT to Rec-I-DCM3(TNT) on one large dataset

Part 3: Multiple sequence alignment

- SATe (Simultaneous Alignment and Tree estimation)
- Developers: Liu, Nelesen, Linder, and Warnow
- unpublished

Multiple Sequence Alignment

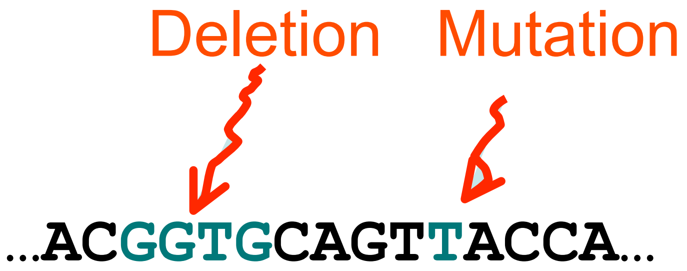
```
AGGCTATCACCTGACCTCCA  -AGGCTATCACCTGACCTCCA
TAGCTATCACGACCGC      TAG-CTATCAC--GACCGC--
TAGCTGACCGC           TAG-CT-------GACCGC--
```

Notes:
1. We insert gaps (dashes) to each sequence to make them “line up”.
2. Nucleotides in the same column are presumed to have a common ancestor (i.e., they are “homologous”).

Indels and substitutions at the DNA level

…**ACGGTGCAGTTACCA**…

Indels and substitutions at the DNA level

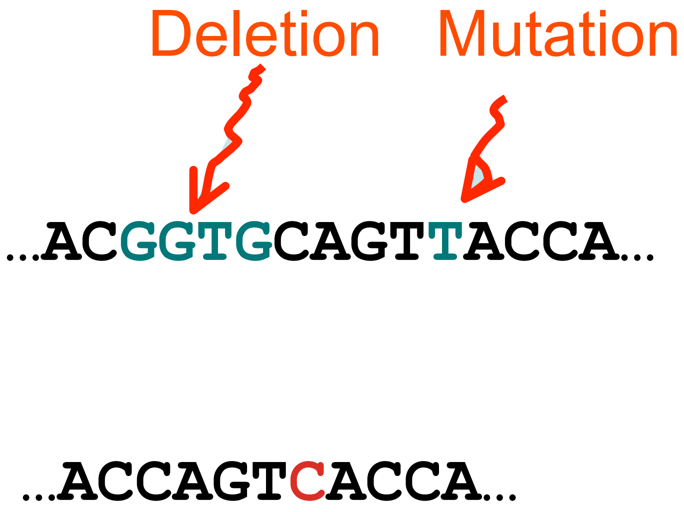


Deletion

Mutation

...ACGGTGCAGTTACCA...

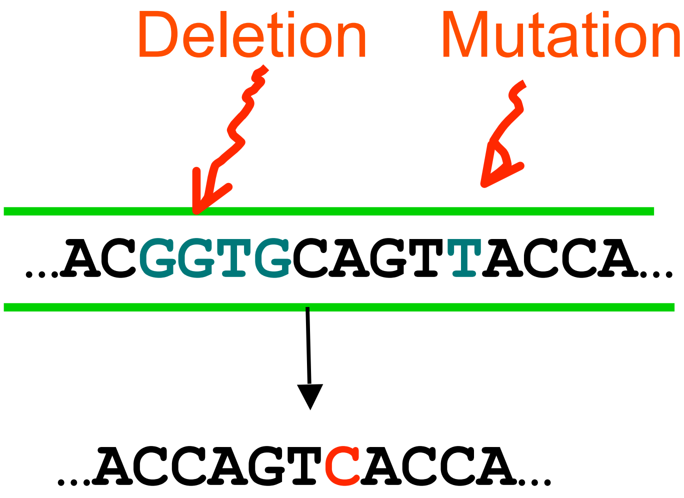
Indels and substitutions at the DNA level



Deletion Mutation

...ACGGTGCAGTTACCA...

...ACCAGTCACCA...



Deletion

Mutation

...ACGGTGCAGTTACCA...

...ACCAGTCACCA...

The true multiple alignment is:

...ACGGTGCAGTTACCA...

...AC----CAGTCACCA...

Basic observations about standard two-phase methods

- Clustal is the standard multiple alignment method used by systematists.
- However, many new MSA methods improve on ClustalW, with ProbCons and MAFFT the two best MSA methods.
- The best current two-phase techniques are obtained by computing maximum likelihood trees on ProbCons or MAFFT alignments (joint work with Wang, Leebens-Mack, and dePamphilis - unpublished).

New method: SATe

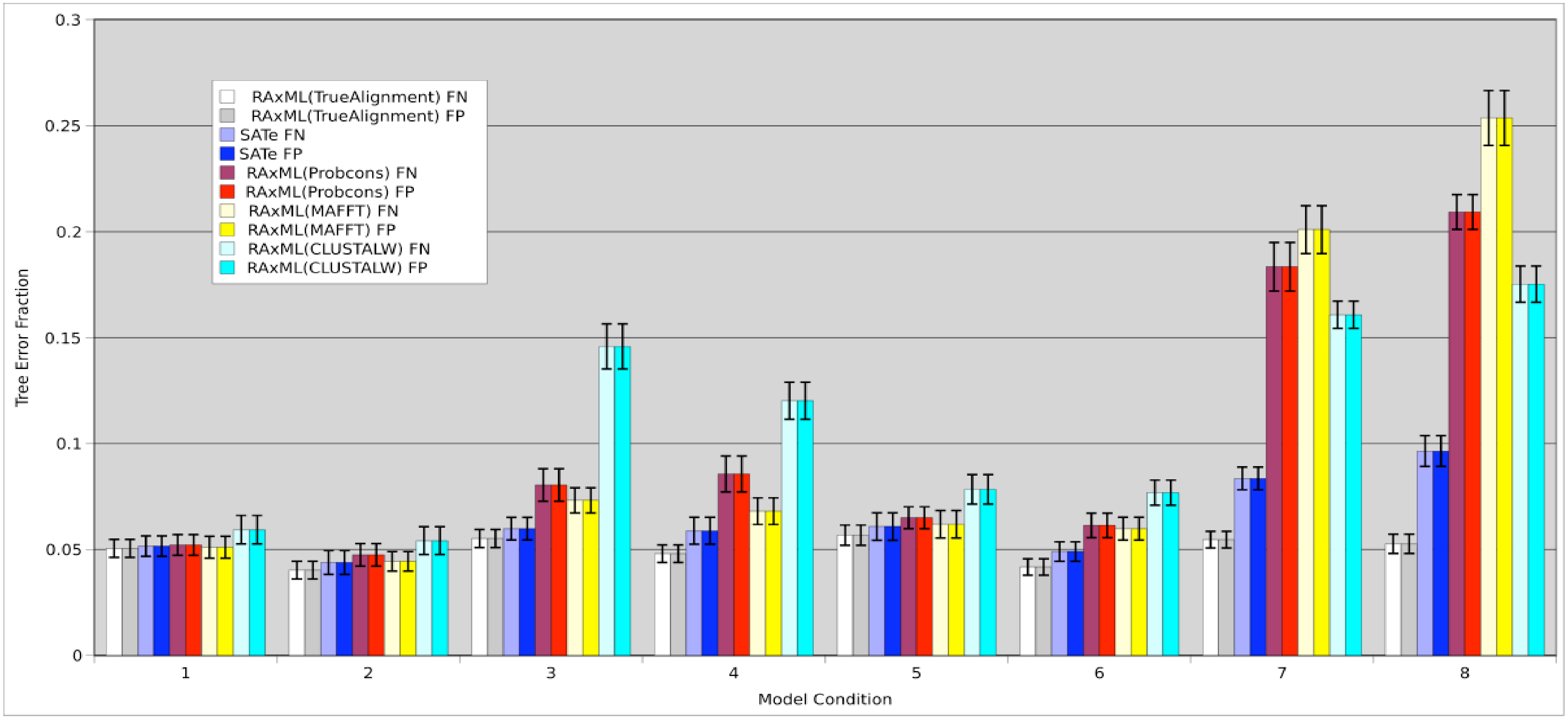
(Simultaneous Alignment and Tree estimation)

- Developers: Warnow, Linder, Liu, Nelesen, and Zhao.
- Basic technique: iteratively ***propose alignments*** (using various techniques), and ***compute maximum likelihood trees*** for these alignments.
- Unpublished.

Simulation study

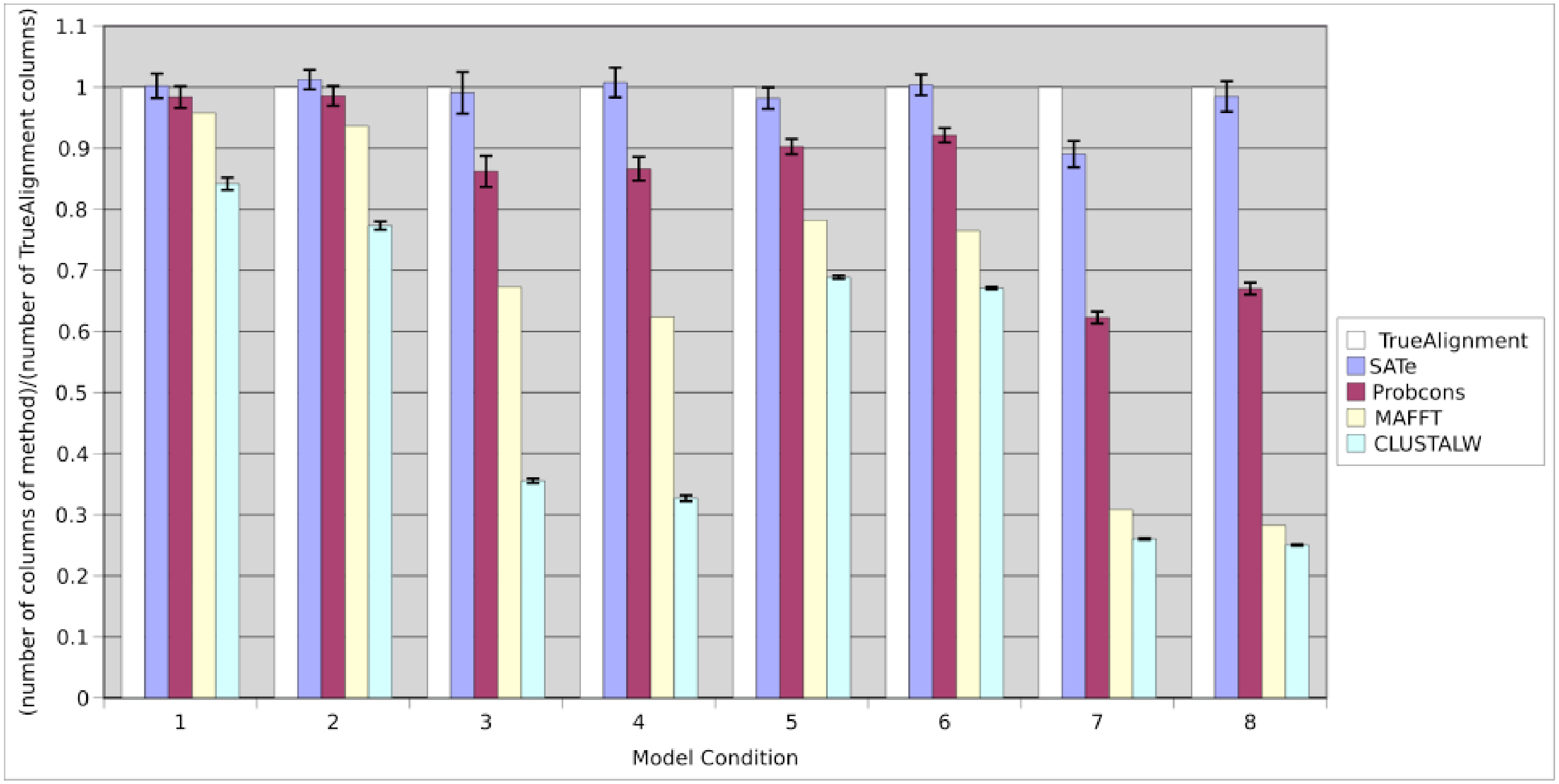
- 100 taxon model trees, 1000 sites at the root
- DNA sequences evolved with indels and substitutions (using ROSE).
- We vary the gap length distribution, probability of gaps, and probability of substitutions, to produce 8 model conditions: models 1-4 have "**long gaps**" and 5-8 have "**short gaps**".
- We compare SATe to maximum likelihood trees (using RAxML) on various alignments (including the true alignment), each method limited to 24 hours.

Error rates refer to the proportion of incorrect edges.



Errors in estimating alignments

- Normalized number of columns in the estimated alignment relative to the true alignment.



Summary of SATe

- SATe produces more accurate trees than the best current two-phase method, especially when the evolutionary process has many gap events.
- SATe alignments do not compress the data ("over-align") as much as standard MSA methods, most of which are based upon progressive alignment.

Future work

- Our current research is focused on extending SATe to estimate maximum likelihood under models that include gap events.
- Evolution is more complex than just indels and substitutions: we need methods that can handle ***genome rearrangements*** and ***duplications.***

Acknowledgements

- Funding: NSF, The David and Lucile Packard Foundation, The Program in Evolutionary Dynamics at Harvard, and The Institute for Cellular and Molecular Biology at UT-Austin.
- Collaborators: Peter Erdos, Daniel Huson, Randy Linder, Kevin Liu, Bernard Moret, Serita Nelesen, Usman Roshan, Mike Steel, Katherine St. John, Laszlo Szekely, Tiffani Williams, and David Zhao.
- Thanks also to Li-San Wang and Serafim Batzoglou (slides)