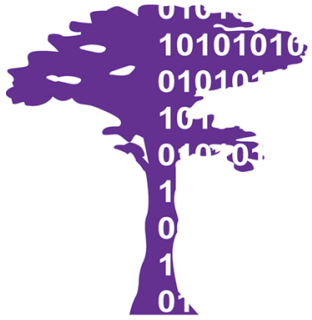
The Tree of Life:
Algorithmic and Software Challenges

Tandy Warnow

The University of Texas at Austin



CIPRES

How did life evolve on earth?

Courtesy of the Tree of Life project

Evolution informs about everything in biology

- Big genome sequencing projects just produce data – so what?
- Evolutionary history relates all organisms and genes, and helps us understand and predict
  - interactions between genes (genetic networks)
  - drug design
  - predicting functions of genes
  - influenza vaccine development
  - origins and spread of disease
  - origins and migrations of humans

The CIPRES Project

(Cyber-Infrastructure for Phylogenetic Research)

The US National Science Foundation funds this project, which has the following major components:

- ***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***
- ***See www.phylo.org for more about CIPRES.***

CIPRES algorithms research

- Heuristics for NP-hard problems in phylogeny reconstruction
- Compact representation of sets of trees
- Reticulate evolution reconstruction
- Gene order phylogeny
- Genomic and multiple sequence alignment
- New phylogeny estimation methods with improved sequence length requirements
- Ancestral sequence reconstruction
- Gene family evolution
- Simultaneous estimation of trees and alignments

CIPRES software

- Improvements and extensions to existing software (MrBayes and Phycas, Mesquite, POY)
- Fast maximum likelihood and maximum parsimony software (using Rec-I-DCM3 boosting)
- Software libraries (for phylogeny estimation method development)
- Portal for phylogenetic analysis (fast ML and MP currently enabled, POY and MrBayes coming shortly).

All ***open-source***

Estimating large phylogenies

Necessary, desirable, but difficult:

- Computationally hard: Many datasets (including the “Tree of Life”) are big, and optimization problems are NP-hard
- Desirable and/or necessary: Taxonomic sampling enables more accurate study of adaptive evolution

Over the last decade or so, there has been tremendous progress in developing fast methods for statistical estimation of phylogenies with greatly improved accuracy (both with respect to topologies, and with respect to optimization problems).

Is the problem solved? Not at all.

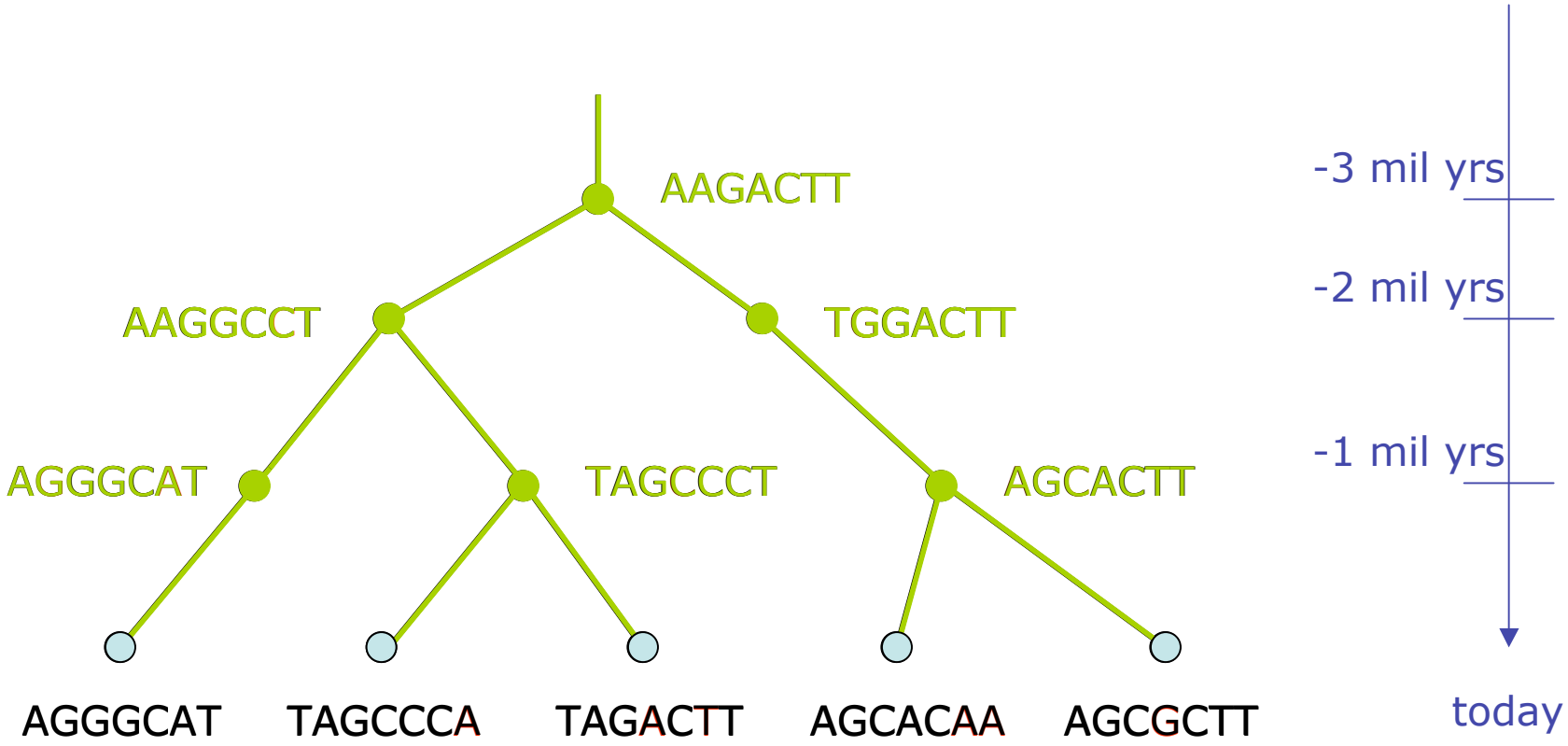
This talk

- Progress on large-scale phylogeny estimation
  - absolute fast-converging methods
  - improved heuristics for NP-hard optimization problems
  - simultaneous estimation of alignments and trees
- Problems that still need to be addressed

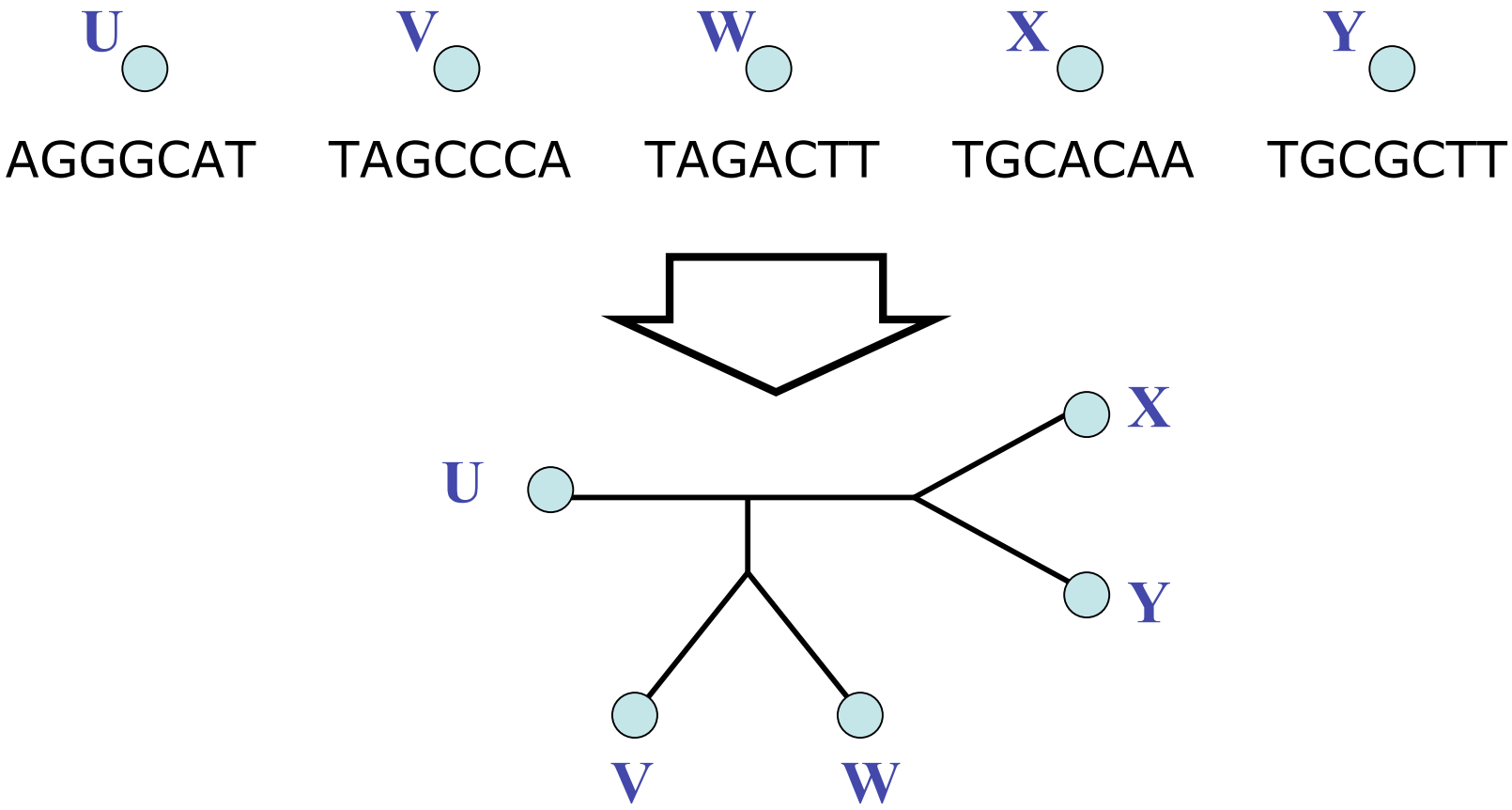
Steps in a phylogenetic analysis

- Gather data
- Align sequences
- Reconstruct phylogeny on the multiple alignment - often obtaining a large number of trees
- Compute consensus (or otherwise estimate the reliable components of the evolutionary history)
- Perform post-tree analyses.

DNA Sequence Evolution

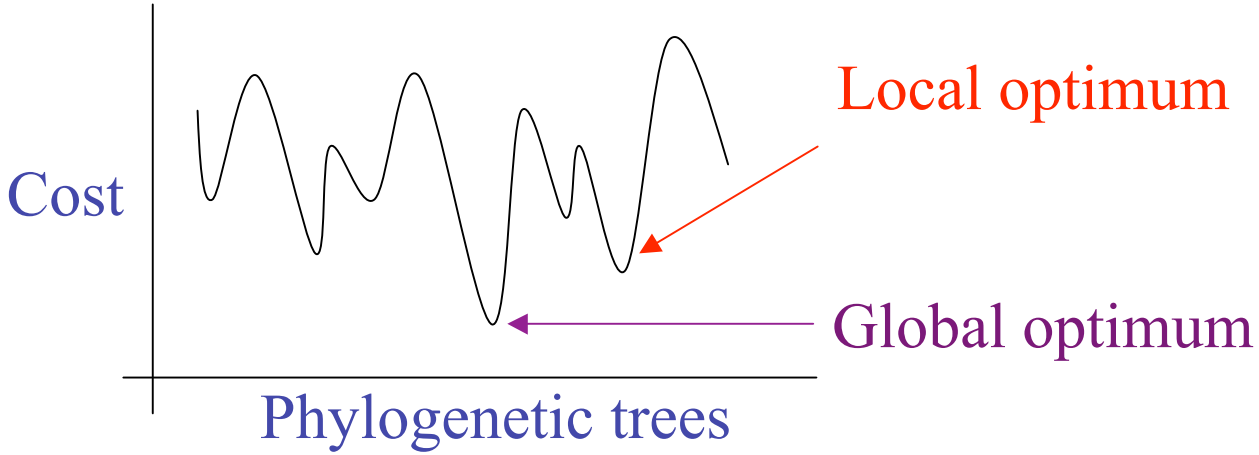


What about phylogeny reconstruction methods?



Phylogenetic reconstruction methods

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



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

Performance criteria

- Running time.
- Space.
- Statistical performance issues (e.g., statistical consistency and sequence length requirements)
- “Topological accuracy” with respect to the underlying ***true tree.*** Typically studied in simulation.
- Accuracy with respect to a particular criterion (e.g. tree length or likelihood score), on real data.

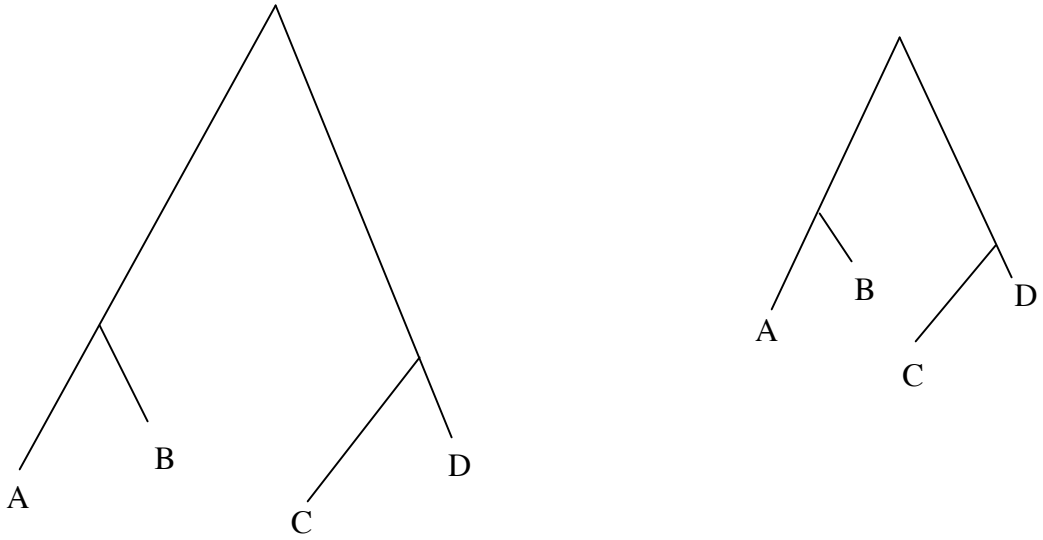
Markov models of single site evolution

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.

Modelling variation between characters: Rates-across-sites



- If a site (i.e., character) is twice as fast as another on one edge, it is twice as fast everywhere.
- The distribution of the rates is typically assumed to be gamma.

Identifiability and statistical consistency

- A model is *identifiable* if it is uniquely characterized by the probability distribution it defines.
- A phylogenetic reconstruction method is *statistically consistent* under a model if the probability that the method reconstructs the true tree goes to 1 as the sequence length increases.

Identifiability results

- The “standard” Markov models (from Jukes-Cantor to the General Markov model) are identifiable.
- These models are also identifiable when sites draw rates from a gamma distribution (easy to prove if the distribution is known, and harder to prove if the distribution must be estimated - *cf. Allman and Rhodes*).
- However, mixed models are often not identifiable (*cf*. *Matsen and Steel*), nor are some models in which sites draw rates from more complex distributions.

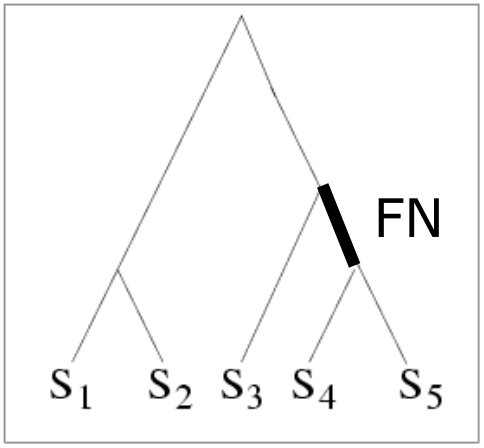
Phylogeny estimation typically is done under identifiable models.

Theoretical results I

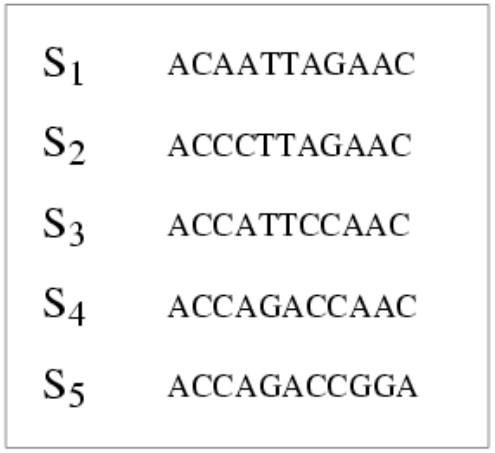
- Neighbor Joining is polynomial time, and statistically consistent.
- Maximum Parsimony is NP-hard, and even exact solutions are not statistically consistent.
- Maximum Likelihood is NP-hard, but exact solutions are statistically consistent

What about performance on finite data?

Quantifying Error



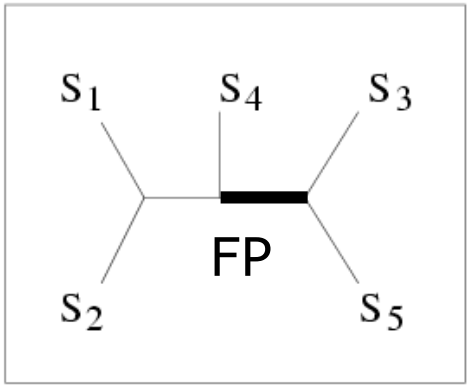
TRUE TREE



DNA SEQUENCES

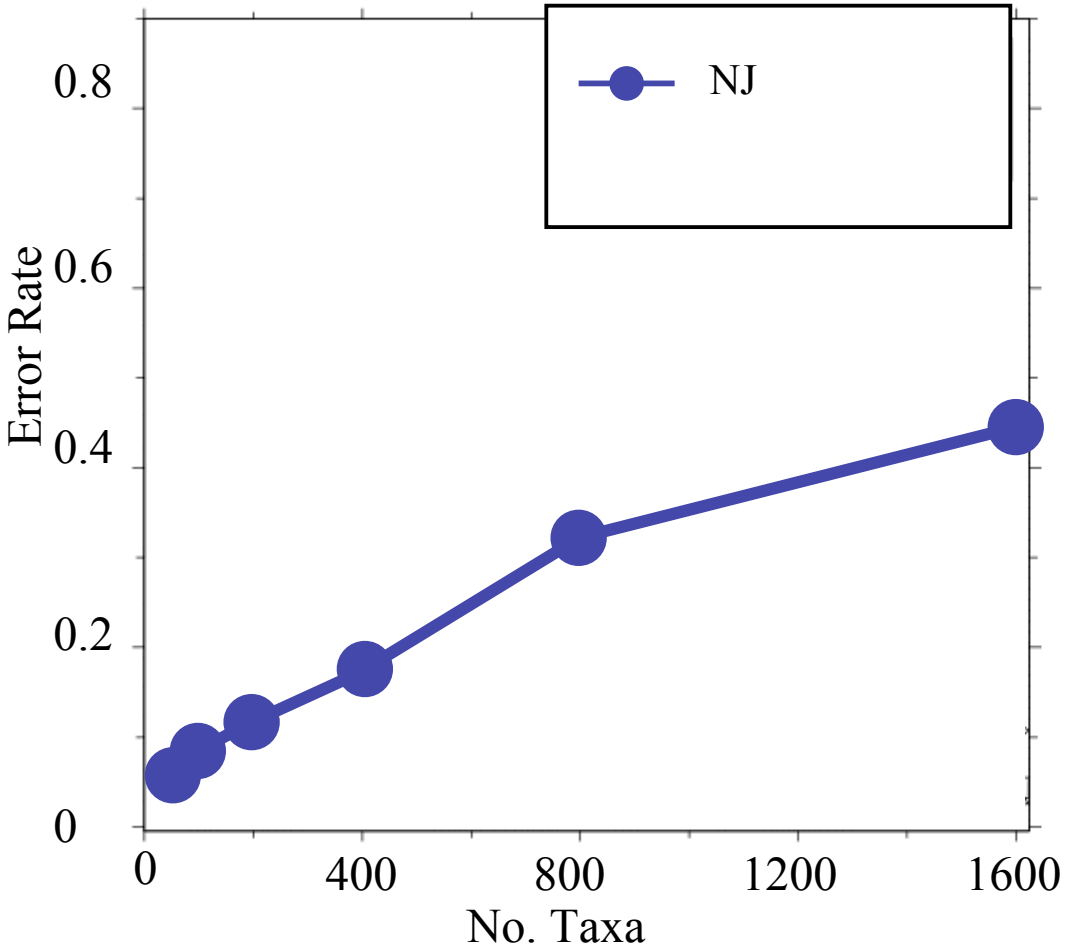
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate



INFERRED TREE

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.

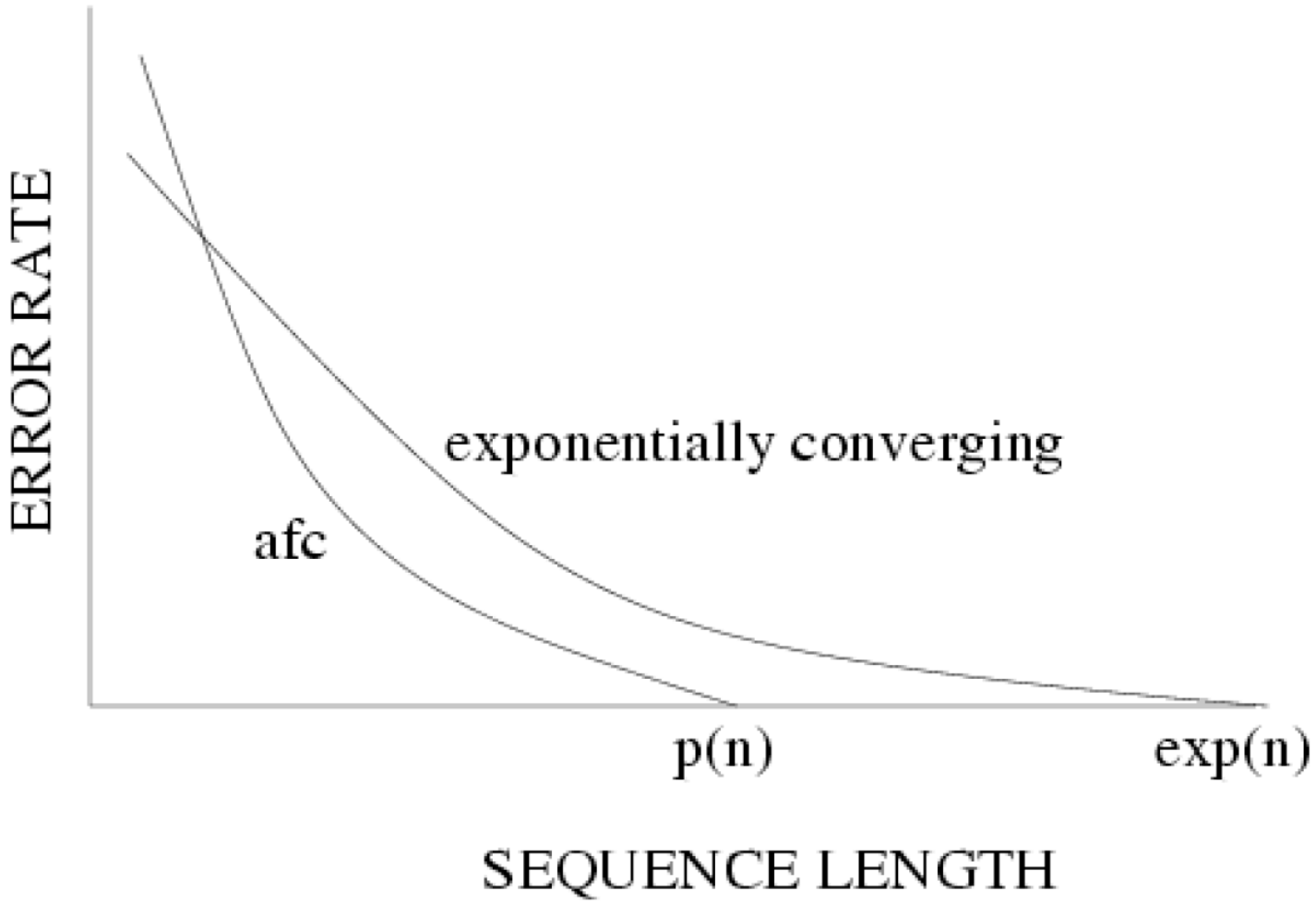
Error rates reflect proportion of incorrect edges in inferred trees.

Theoretical results II

- 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). Exponential lower bound for caterpillar trees: Lacey and Chang.

- Maximum likelihood will return the true tree with high probability provided sequence lengths are **exponential** in the number of taxa (Steel and Szekely).

Exponential convergence and absolute fast convergence (afc)

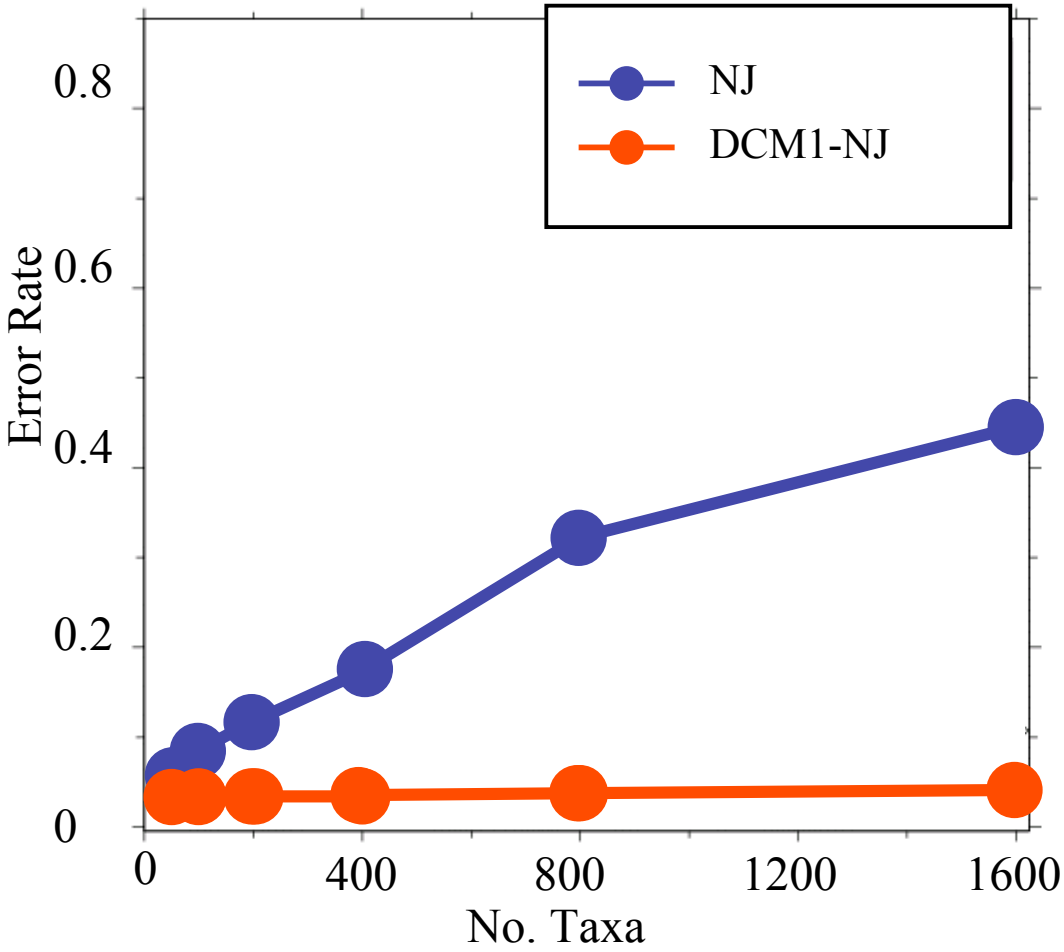


Afc methods

- The “short quartet” methods (Erdos et al.) were the first (1995)
- DCM-boosting for distance-based methods (Huson, Warnow, St. John, Moret, and others)
- Mossel, Rao and others have recently developed new techniques based upon estimating ancestral sequences
- Others (e.g. Gronau and Moran)

DCM1-boosting distance-based methods

[Nakhleh et al. ISMB 2001]

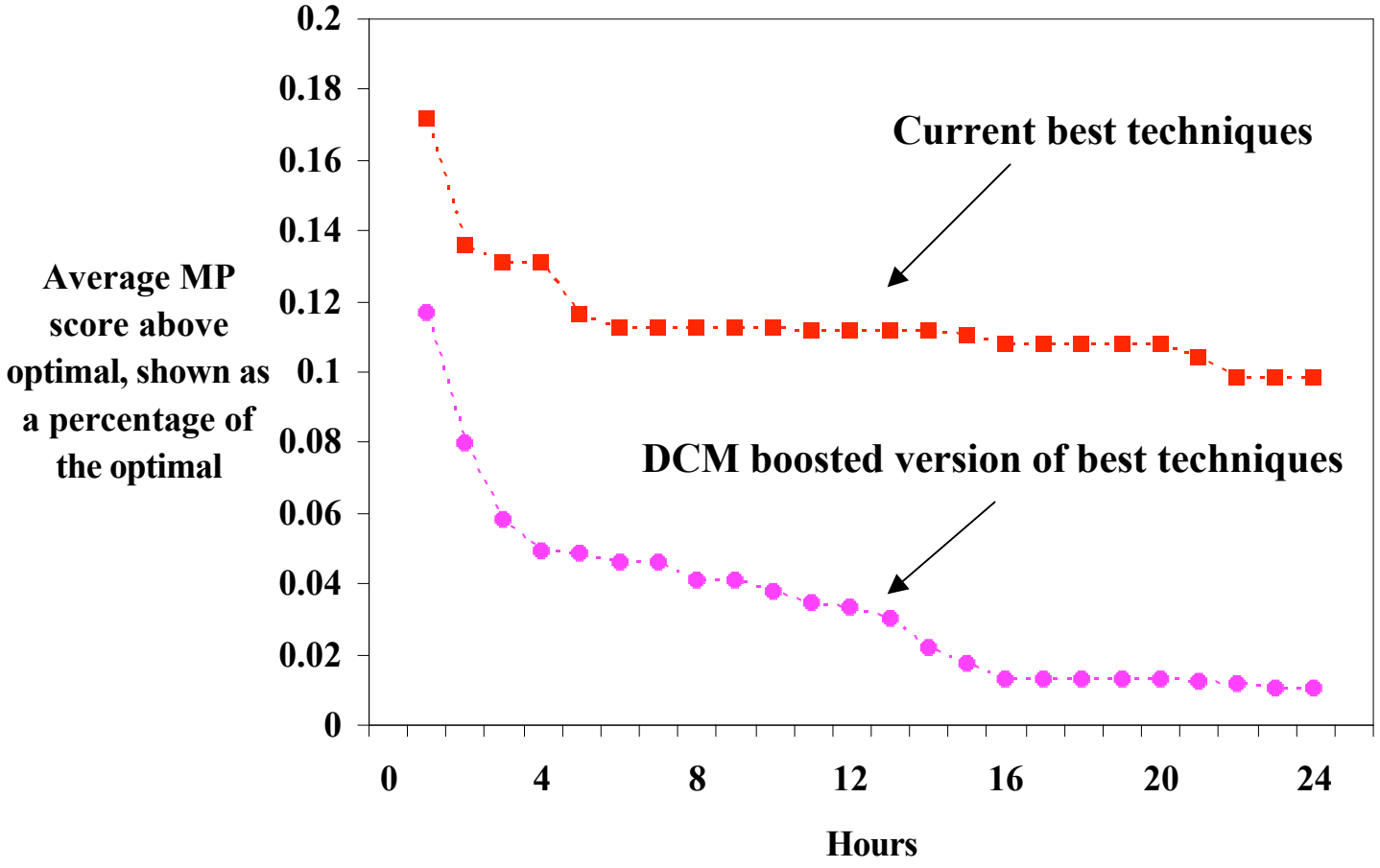


Theorem:
DCM1-NJ converges to the true tree from polynomial length sequences

Large datasets

- Better accuracy is obtained through good heuristics for NP-hard optimization (esp. maximum likelihood)
- CIPRES has developed new “boosters” for large-scale optimization routines

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



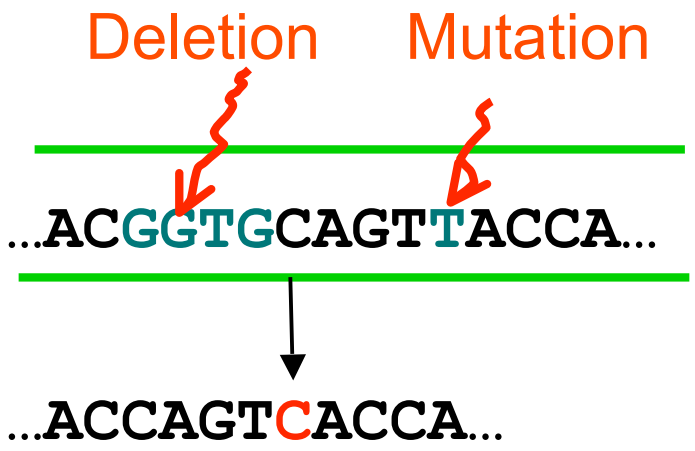
Comparison of TNT to Rec-I-DCM3(TNT) on one large dataset

All well and good…

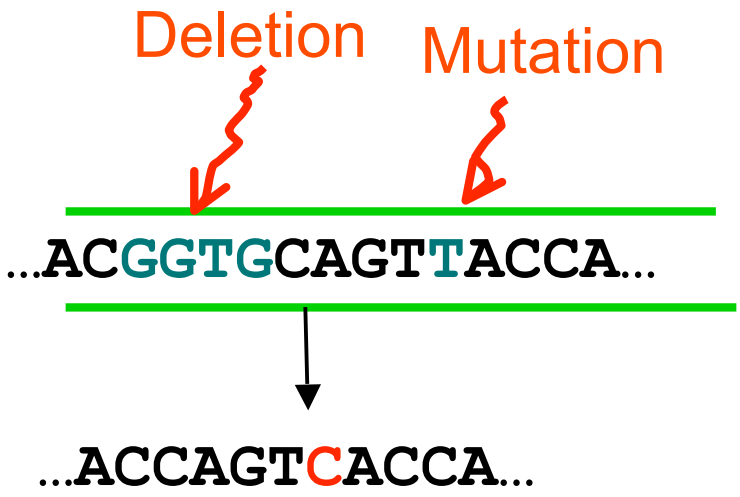
- But evolution is more complicated than that!

Steps in a phylogenetic analysis

- Gather data
- ***Align sequences***
- Reconstruct phylogeny on the multiple alignment - often obtaining a large number of trees
- Compute consensus (or otherwise estimate the reliable components of the evolutionary history)
- Perform post-tree analyses.



indels also occur!



The true pairwise alignment is:

...ACGGTGCAGTTACCA...

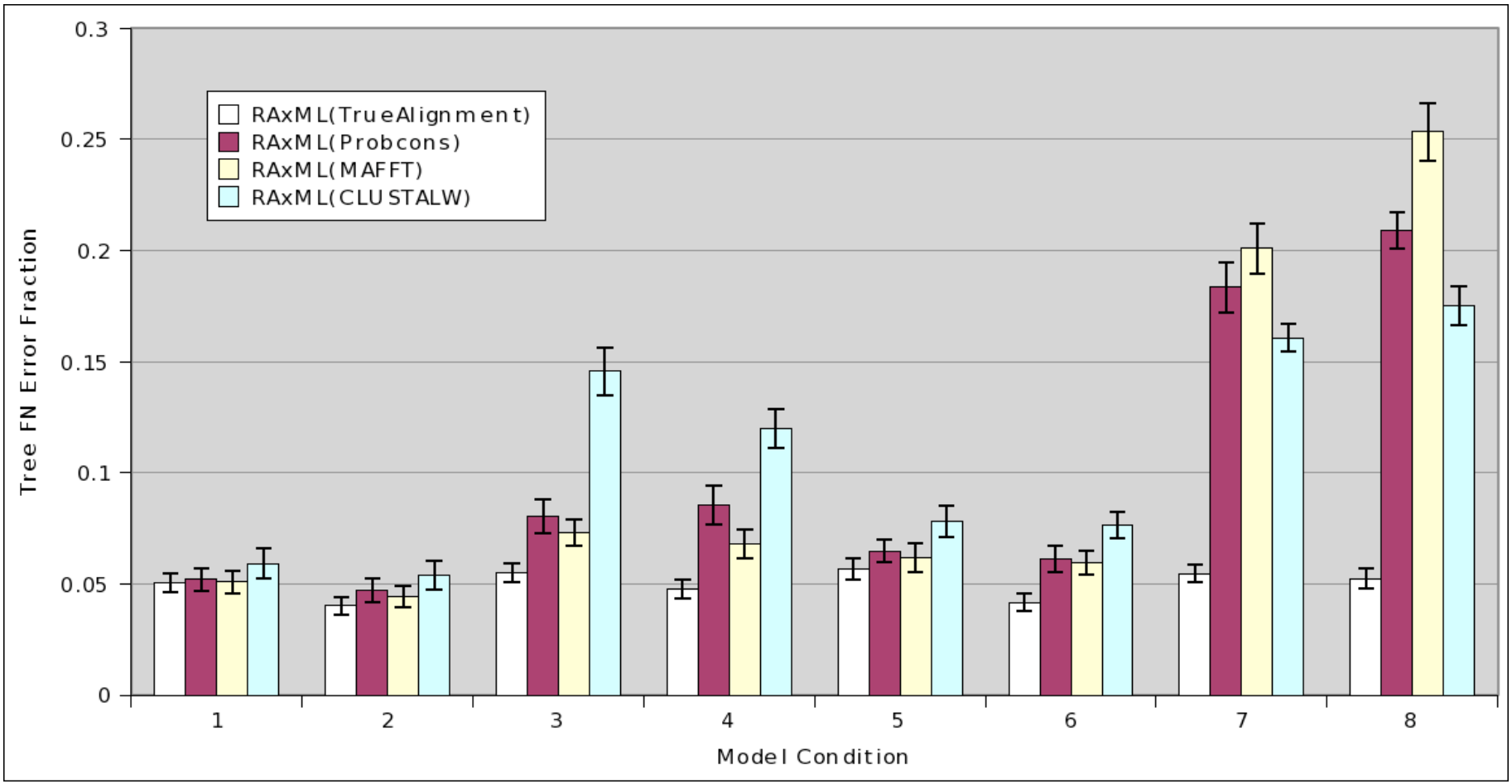
...AC----CAGTCACCA...

The **true multiple alignment** on a set of homologous sequences is obtained by tracing their evolutionary history, and extending the pairwise alignments on the edges to a multiple alignment on the leaf sequences.

Simulation study

- 100 taxon model trees (generated by r8s and then modified, so as to deviate from the molecular clock).
- DNA sequences evolved under ROSE (indel events of blocks of nucleotides, plus HKY site evolution). The root sequence has 1000 sites.
- 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 compared RAxML on various alignments (including the true alignment).

Non-coding DNA evolution



Models 1-4 have “long gaps”, and models 5-8 have “short gaps”

Two problems with two-phase methods

- Current MSA methods have high error rates when sequences evolve with many indels and substitutions.
- Current phylogeny estimation methods treat indel events inadequately (either treating as missing data, or giving too much weight to each gap).

Simultaneous estimation?

- Several Bayesian methods for simultaneous estimation of trees and alignments have been developed, but none can be applied to datasets with more than (approx.) 20 sequences.
- POY attempts to solve the NP-hard “minimum length tree” problem, where gaps contribute to the length of the tree and can be applied to large datasets. However, its performance on simulated data isn’t competitive with the best two-phase methods (unpublished data).

New method: SATe

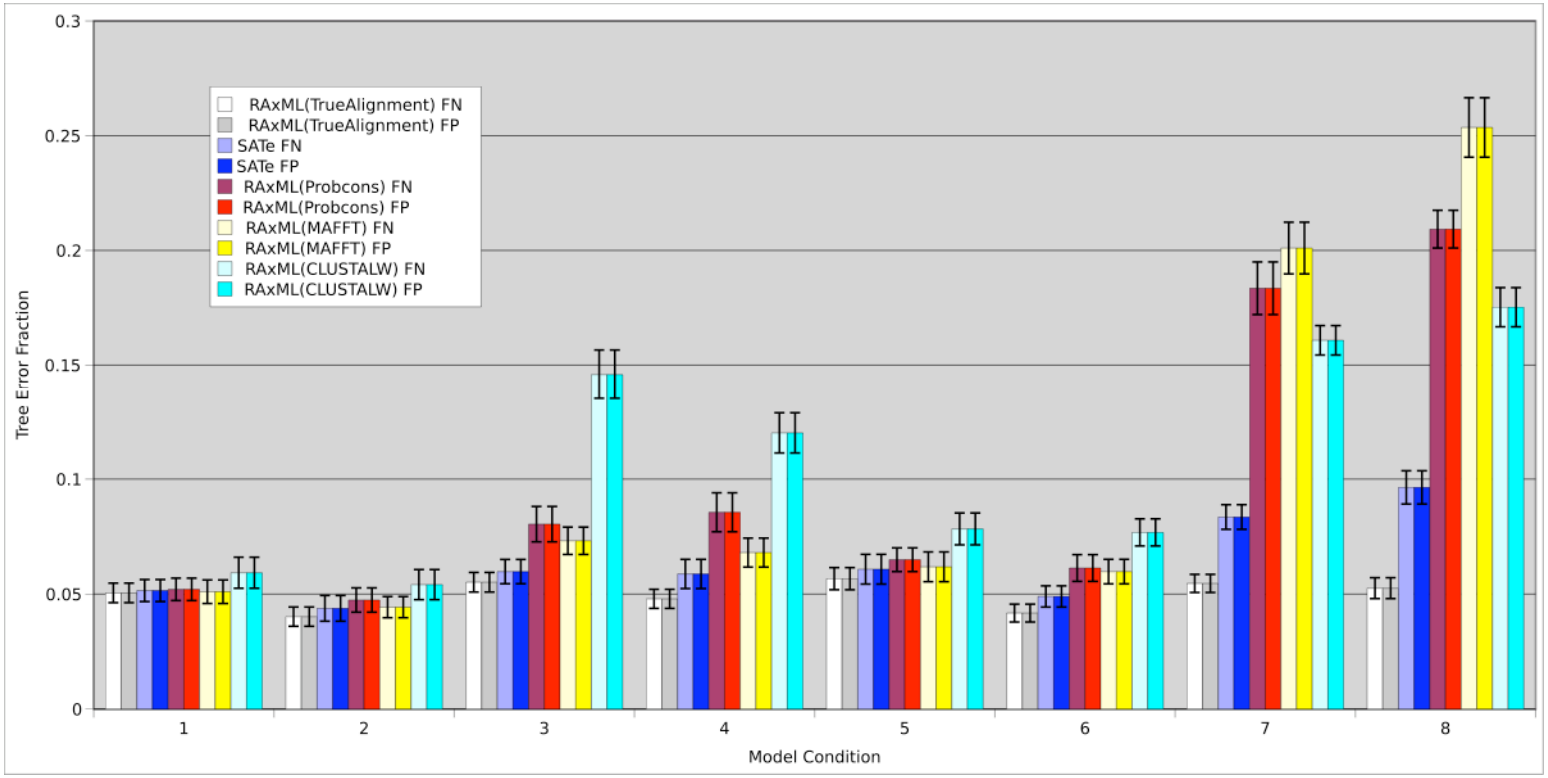
(Simultaneous Alignment and Tree estimation)

- Developers: Warnow, Linder, Liu, Nelesen, and Zhao.
- Basic technique: ***heuristically propose different alignments*** and ***compute maximum likelihood trees*** for these alignments under GTR+Gamma+I.
- Unpublished.

Topological accuracy

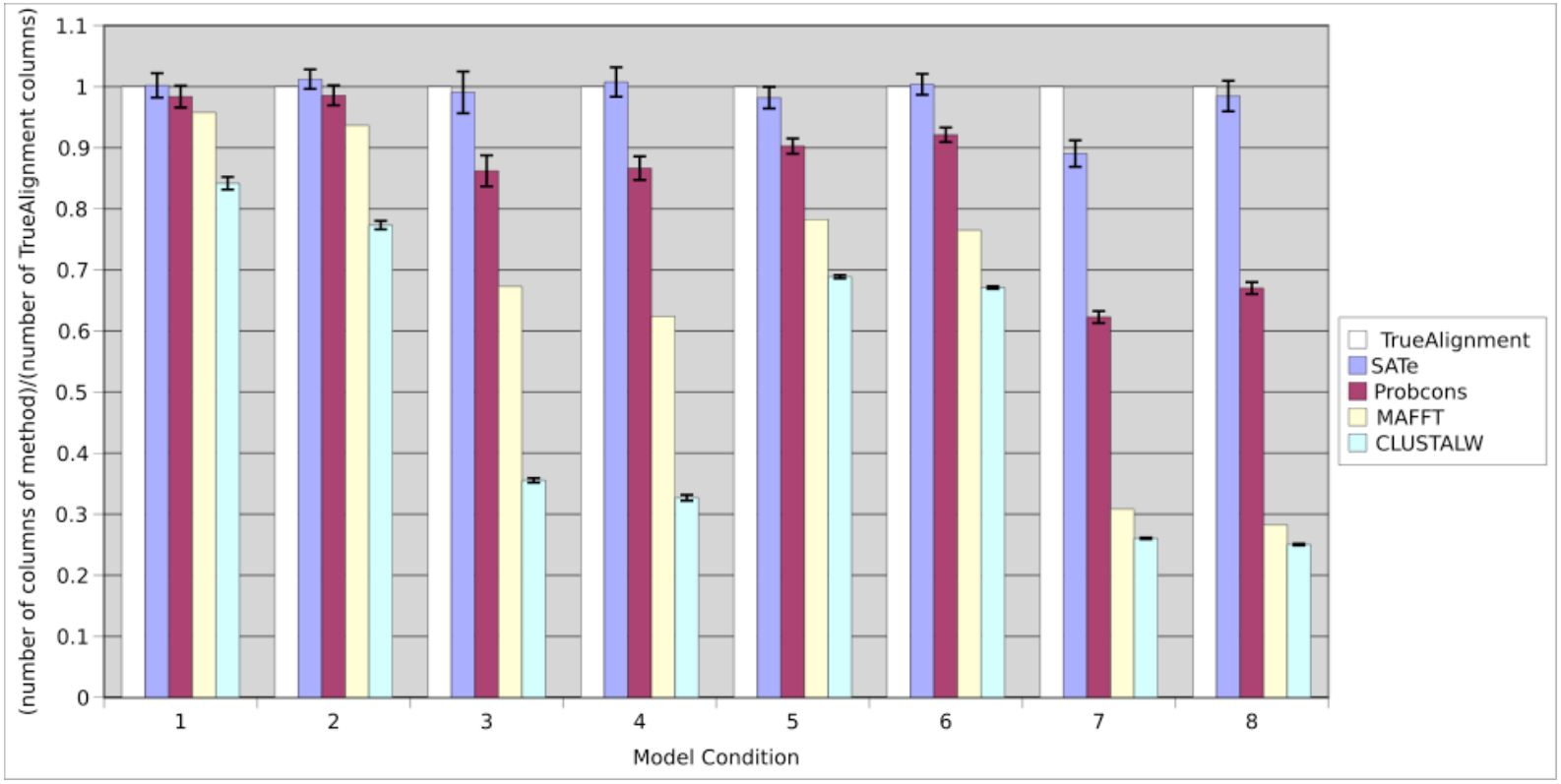
FN (false negative): proportion of correct edges missing from the estimated tree

FP (false positive): proportion of incorrect edges in the estimated tree



Alignment accuracy

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



Multiple sequence alignment

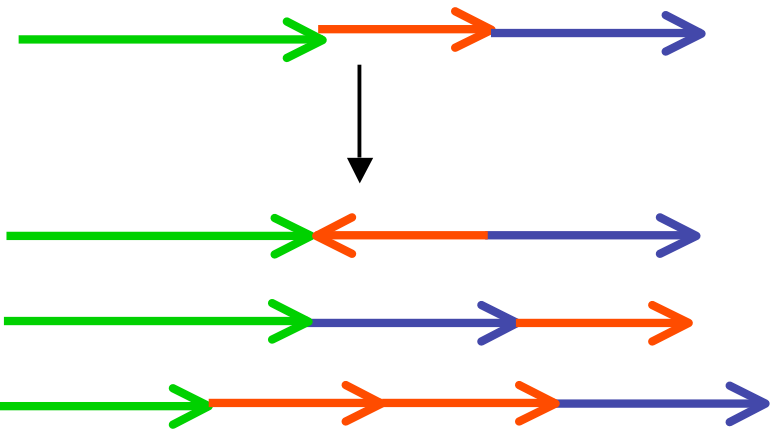
- SATe gives an improvement over standard two-phase methods, but better performance is needed.

- We conjecture that ML estimation under models that include gaps should yield good results.

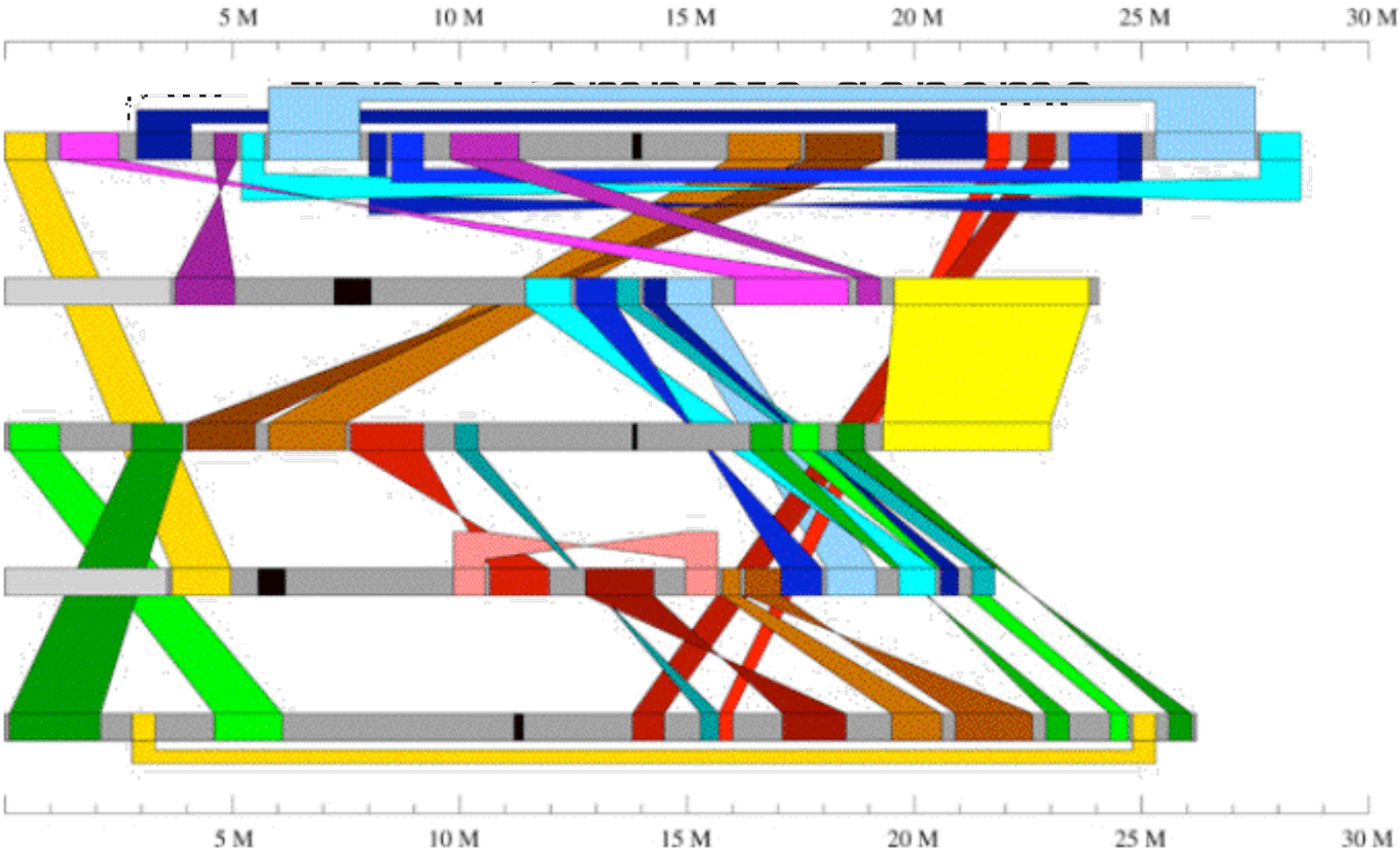
- But evolution is more complicated than that!

Genome-scale evolution

(REARRANGEMENTS)



Whole genome processes



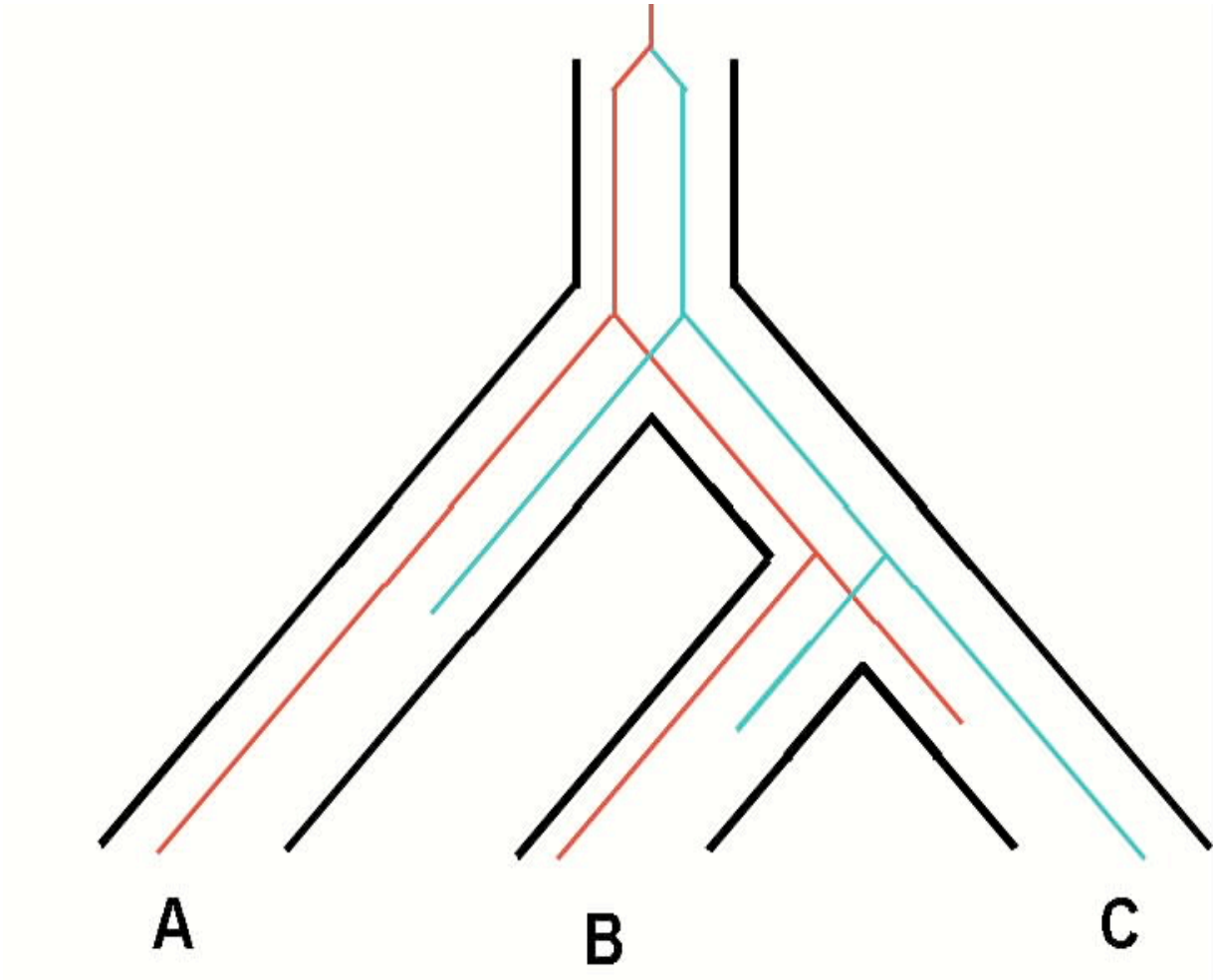
Whole genome phylogenetics

- Given collection of whole genomes, find best alignment and phylogeny.

- Previous work: even when the alignment is given, optimization problems are NP-hard (e.g., minimizing the total number of inversions on a fixed tree).
- Effective heuristics exist for some special cases (once the alignment is given).

- But evolution is more complicated than that!

Gene Tree/Species Tree

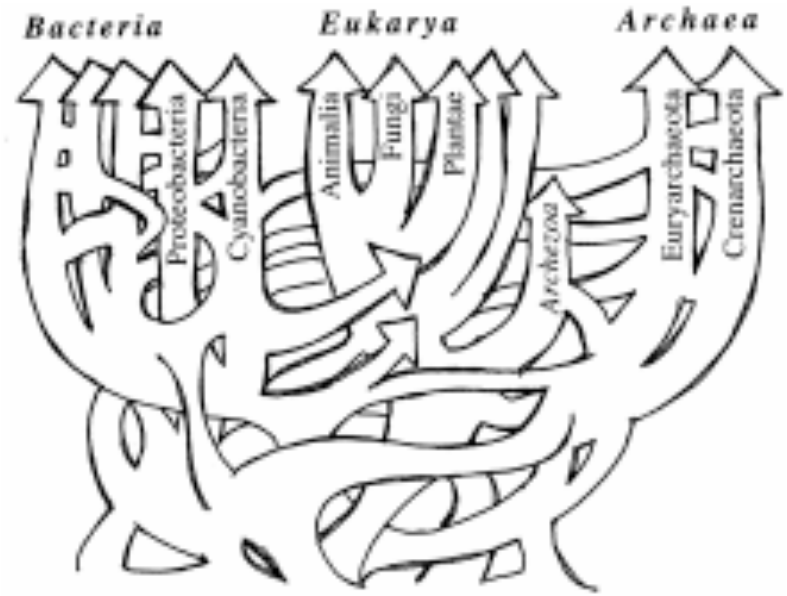


Reconciliation problem

- Given a collection of estimated gene trees, find best species tree
- Previous work: if the true gene tree and species tree are given, the minimum cost duplication and loss history can be estimated.
- Issues: how to handle incomplete resolution, support estimations, etc?

- But evolution is more complicated than that!

The “tree of life” is not a tree



Reticulate evolution (horizontal gene transfer and hybridization) is also a problem

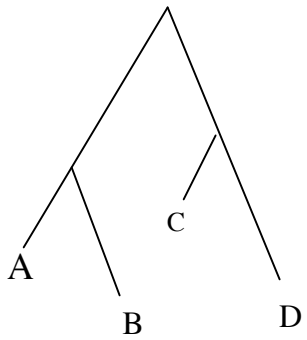
Reticulate evolution detection and reconstruction

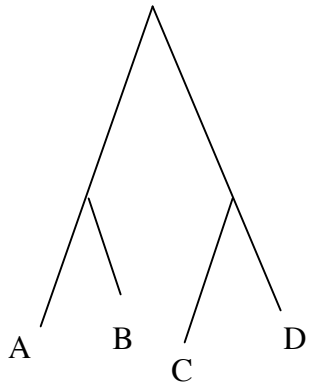
- Previous work: NeighborNet, SplitsTree, Network, etc.
- Main challenge: distinguishing between various processes (finite data, alignment estimation error, homoplasy, model mis-specification, gene tree/species tree distinctions, inadequate analysis) that suggest reticulation

- But evolution is more complicated than that!

Modelling variation between characters: Heterotachy

- A separate random variable for every combination of site and edge - the underlying tree is fixed, but otherwise there are no constraints on variation between sites.





Heterotachy and other mixture models

- Mixture models are not identifiable (Matsen and Steel, and others)
- It is computationally challenging to estimate under these models

Estimating large phylogenies

Necessary, desirable, but difficult:

- Statistically hard: Model-based approaches will need to deal with model misspecification, marker-specific and lineage-specific variation
- Computationally hard: The “Tree of Life” is big, and optimization problems are NP-hard
- Data challenges: missing data, or markers that cannot be aligned, or which evolve too slowly (or too quickly) for the region of interest
- Desirable and/or necessary: Taxonomic sampling enables more accurate study of adaptive evolution

Also:

- Gene tree/species tree differences (for various reasons)
- Reticulation (horizontal gene transfer and hybridization)

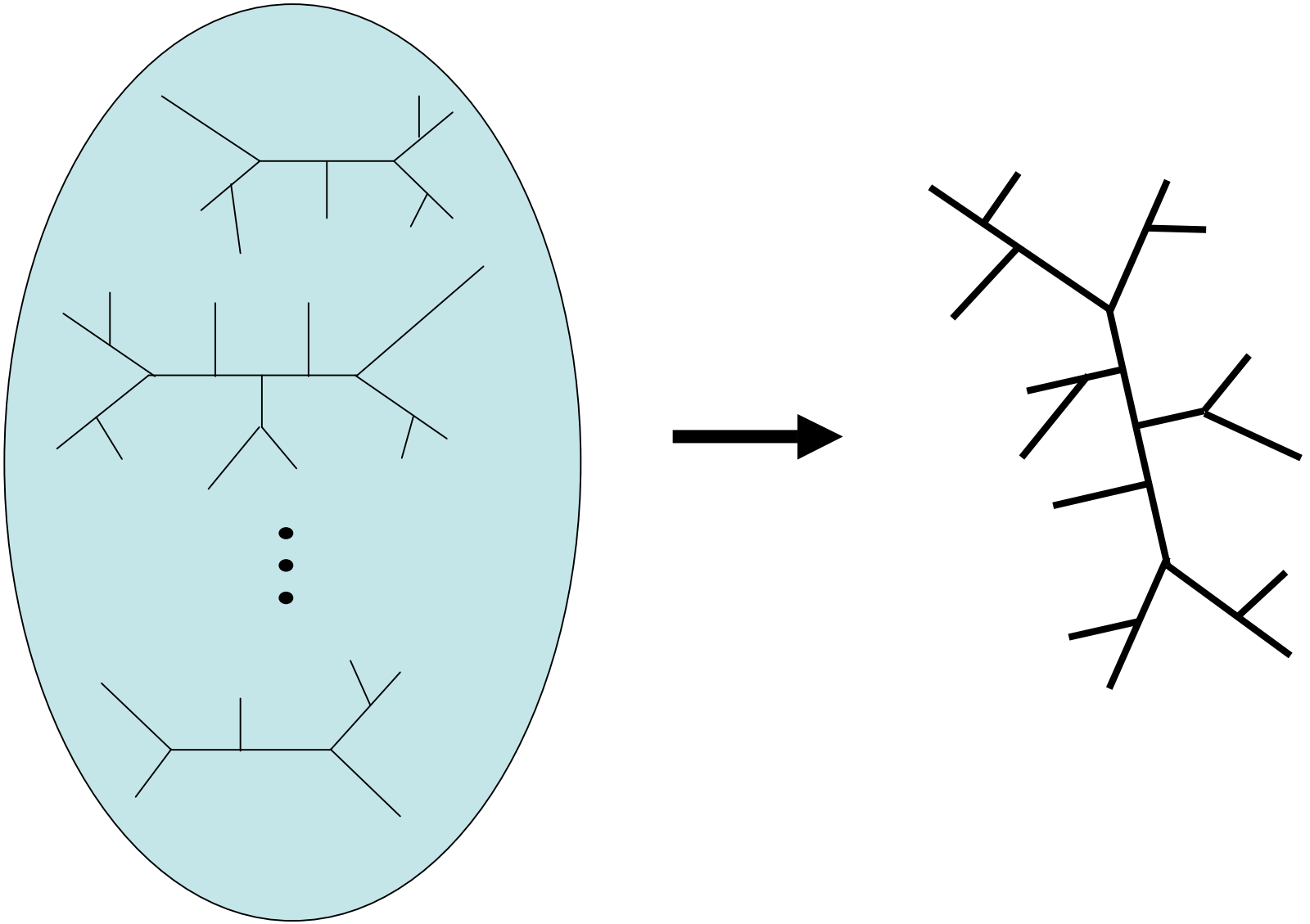
Problems we need to solve

- Simultaneous alignment and tree reconstruction using maximum likelihood
- Whole genome alignment and phylogeny reconstruction
- Reconciling estimates of gene trees into a species phylogeny
- Reticulate evolution detection and reconstruction
- Better supertree methods
- Better visualization tools for multiple alignment and phylogenies
- ***Better models of evolution (for simulation and estimation)***

Acknowledgements

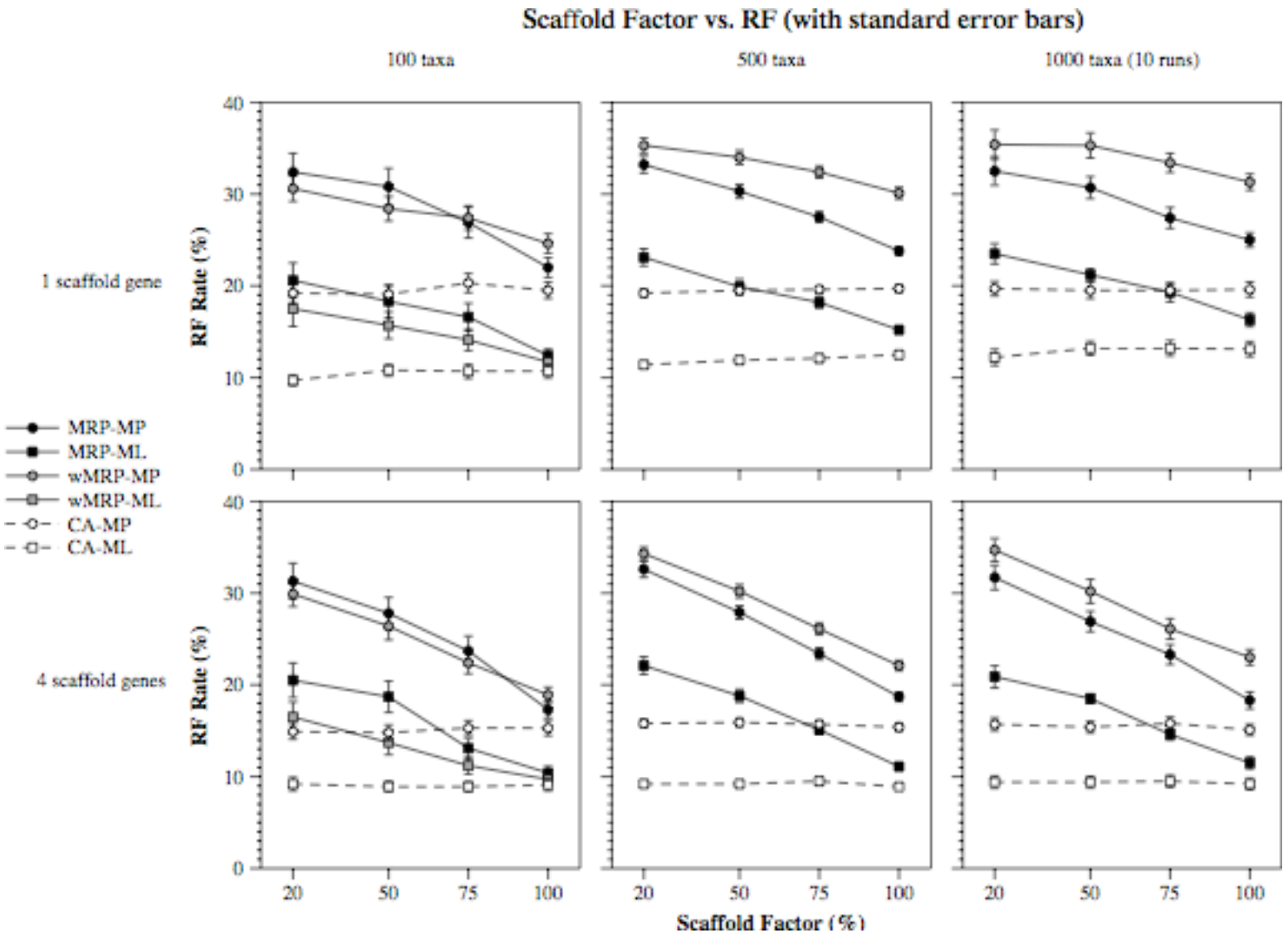
- Funding: NSF, The David and Lucile Packard Foundation, The Program in Evolutionary Dynamics at Harvard, and The Institute for Cellular and Molecular Biology at UT-Austin.
- Collaborators: Peter Erdos, Daniel Huson, Randy Linder, Kevin Liu, Bernard Moret, Serita Nelesen, Usman Roshan, Mike Steel, Katherine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

What is a Supertree Method?



Why Use Supertree Methods?

- Data:
  - Incongruent data types
  - Large amounts of missing data
  - Already have overlapping trees
- Improve performance (because smaller datasets?)


Scaffold Factor vs. RF (with standard error bars)
100 taxa
500 taxa
1000 taxa (10 runs)
1 scaffold gene
4 scaffold genes
RF Rate (%)
RF Rate (%)
0
10
20
30
40
20
50
75
100
Scaffold Factor (%)
MRP-MP
MRP-ML
wMRP-MP
wMRP-ML
CA-MP
CA-ML