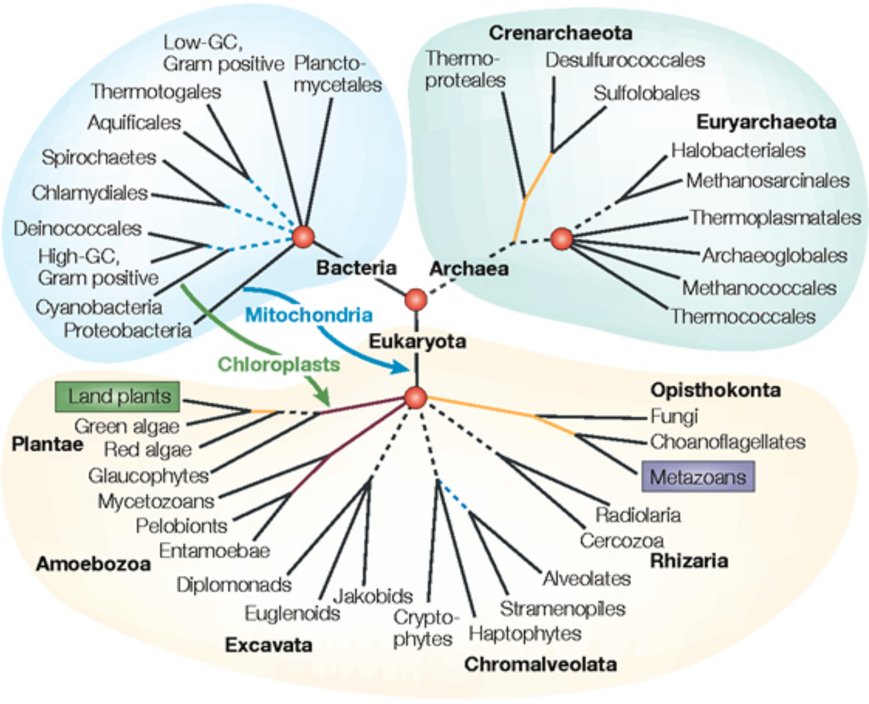
Introduction to Phylogenomics and Metagenomics

Tandy Warnow
The Department of Computer Science
The University of Texas at Austin

The “Tree of Life”



Nature Reviews | Genetics

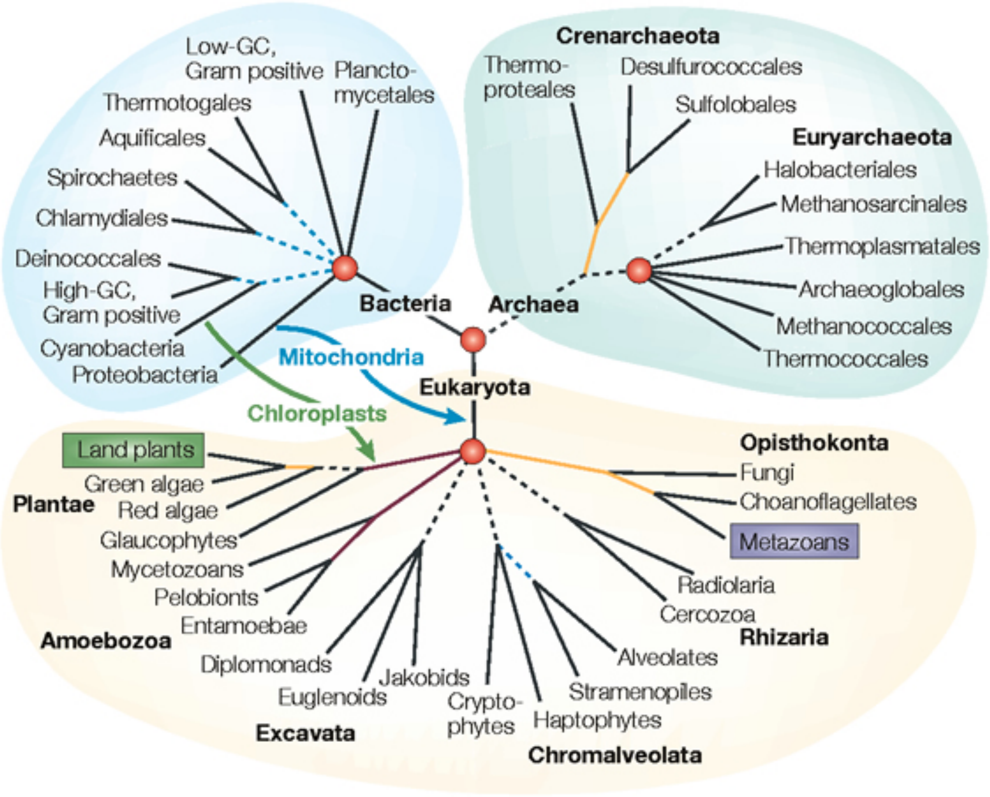
Applications of phylogenies to:

- protein structure and function
- population genetics
- human migrations

Estimating phylogenies is a complex analytical task

Large datasets are very hard to analyze with high accuracy

Phylogenetic Estimation: Big Data Challenges



Nature Reviews | Genetics

NP-hard problems

Large datasets:

100,000+ sequences

10,000+ genes

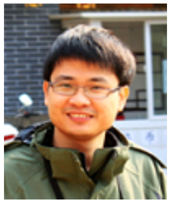
“BigData” complexity

Avian Phylogenomics Project

Erich Jarvis,
HHMI

MTP Gilbert,
Copenhagen

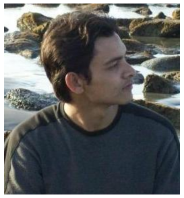
G Zhang,
BGI



T. Warnow
UT-Austin

S. Mirarab
UT-Austin

Md. S. Bayzid
UT-Austin



Plus many many other people…

- Approx. 50 species, whole genomes
- 8000+ genes, UCEs
- Gene sequence alignments and trees computed using SATé

Challenges:

Maximum likelihood tree estimation on multi-million-site sequence alignments

Massive gene tree incongruence

1kp: Thousand Transcriptome Project

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

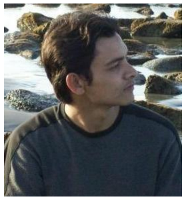
N. Matasci
iPlant

T. Warnow,
UT-Austin

S. Mirarab,
UT-Austin

N. Nguyen,
UT-Austin

Md. S.Bayzid
UT-Austin



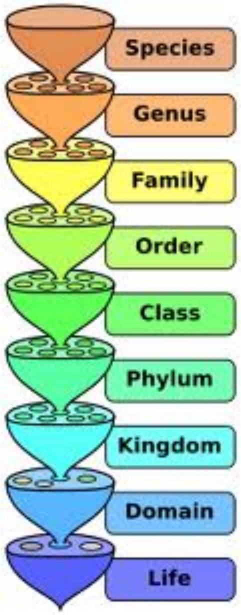
Plus many many other people…

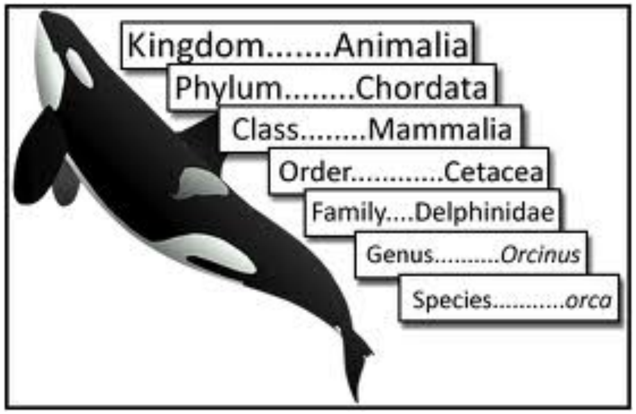
- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)

Challenge:
Alignment of datasets with > 100,000 sequences
Gene tree incongruence

Metagenomic Taxon Identification

Objective: classify short reads in a metagenomic sample





Basic Questions

1. What is this fragment? (Classify each fragment as well as possible.)

2. What is the taxonomic distribution in the dataset? (Note: helpful to use marker genes.)

3. What are the organisms in this metagenomic sample doing together?

Phylogenomic pipeline

- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- Compute multiple sequence alignments for each marker (possibly "mask" alignments)
- Compute species tree or network:
  - Compute gene trees on the alignments and combine the estimated gene trees, OR
  - Perform "concatenation analysis" (aka "combined analysis")
- Get statistical support on each branch (e.g., bootstrapping)
- Use species tree with branch support to understand biology

Phylogenomic pipeline

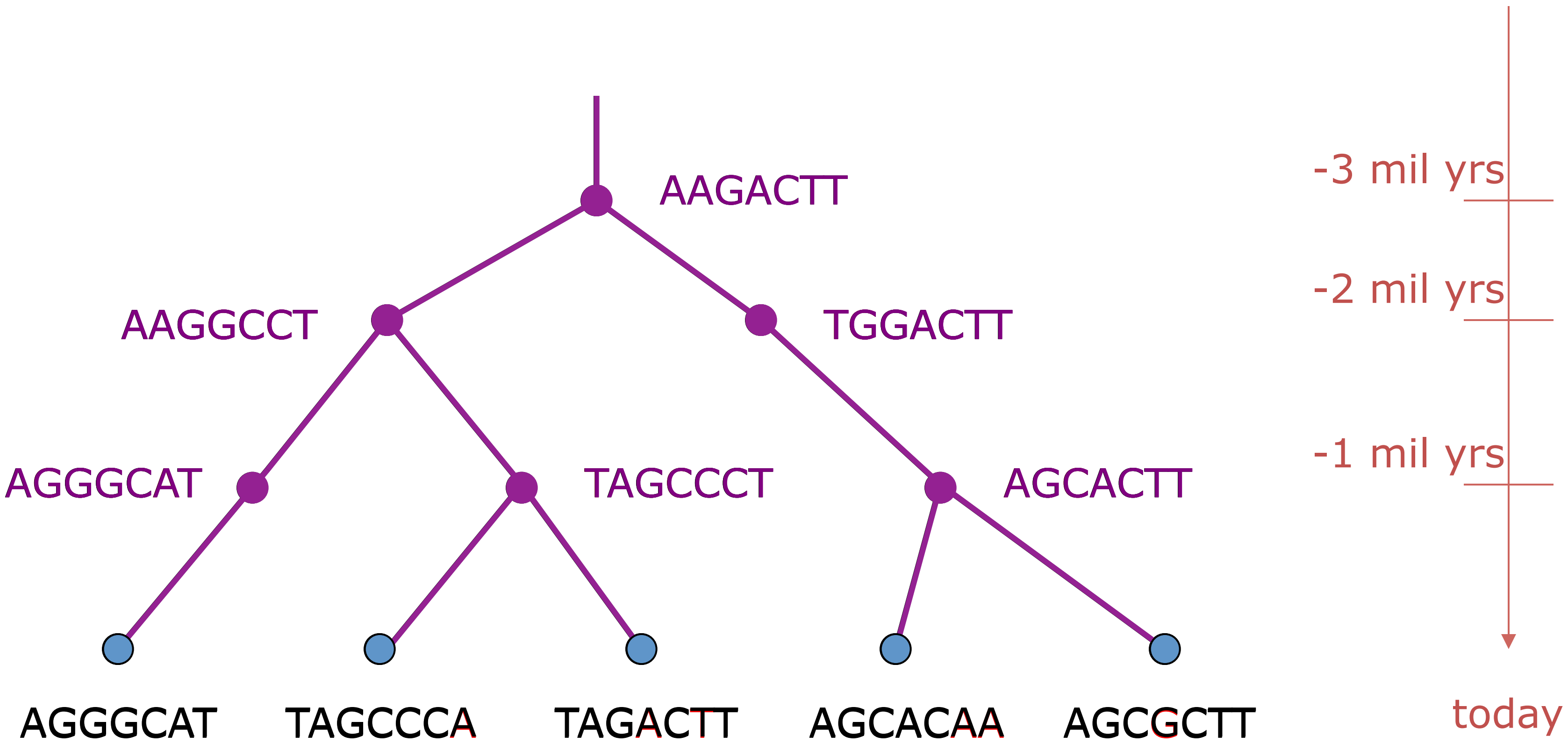
- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- Compute multiple sequence alignments for each marker (possibly “mask” alignments)
- Compute species tree or network:
  - Compute gene trees on the alignments and combine the estimated gene trees, OR
  - Perform “concatenation analysis” (aka “combined analysis”)
- Get statistical support on each branch (e.g., bootstrapping)
- Use species tree with branch support to understand biology

This talk

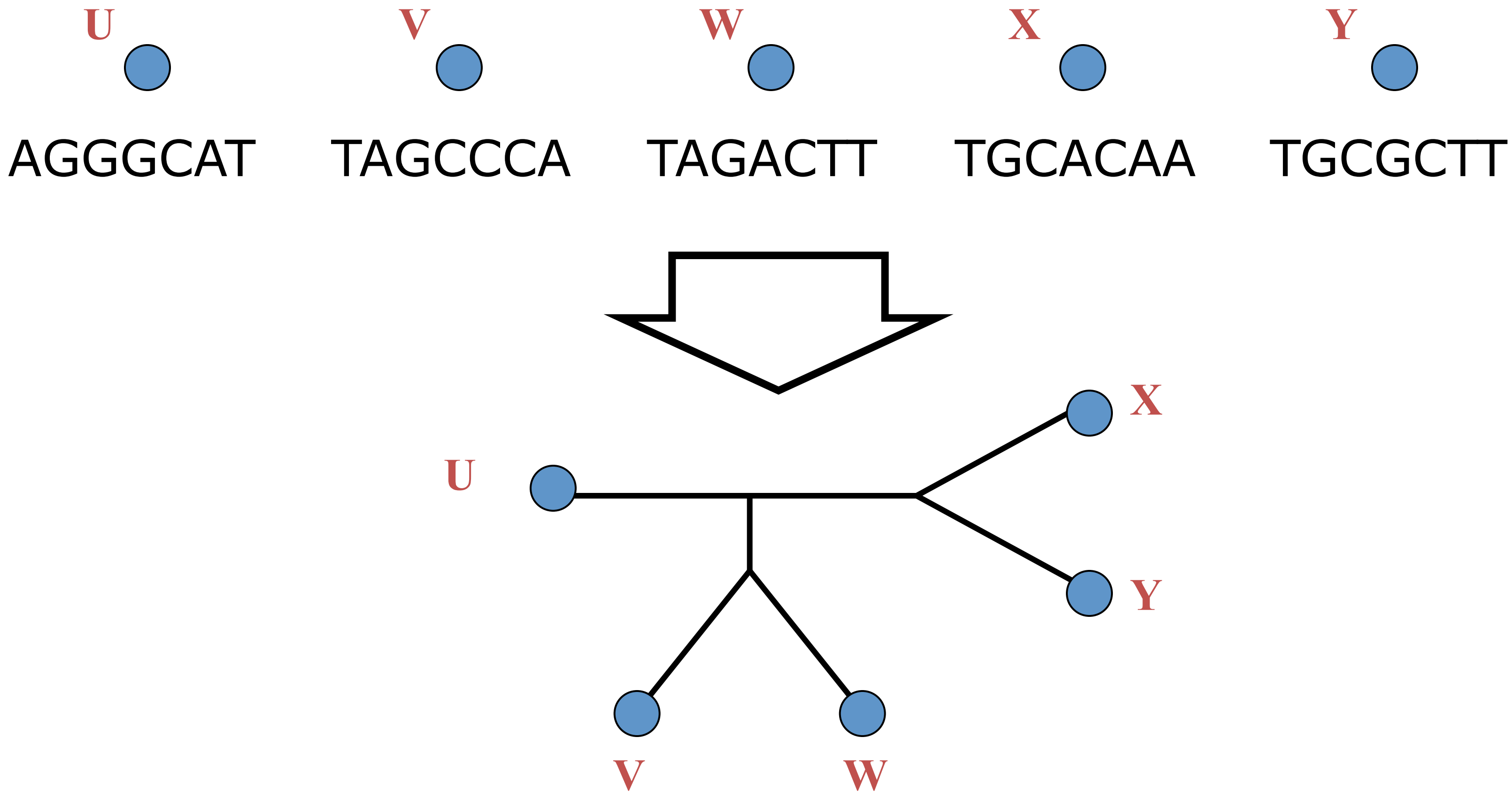
- Phylogeny estimation methods
- Multiple sequence alignment (MSA)
- Species tree estimation methods from multiple gene trees
- Phylogenetic Networks
- Metagenomics
- What we’ll cover this week

Phylogeny Estimation methods

DNA Sequence Evolution

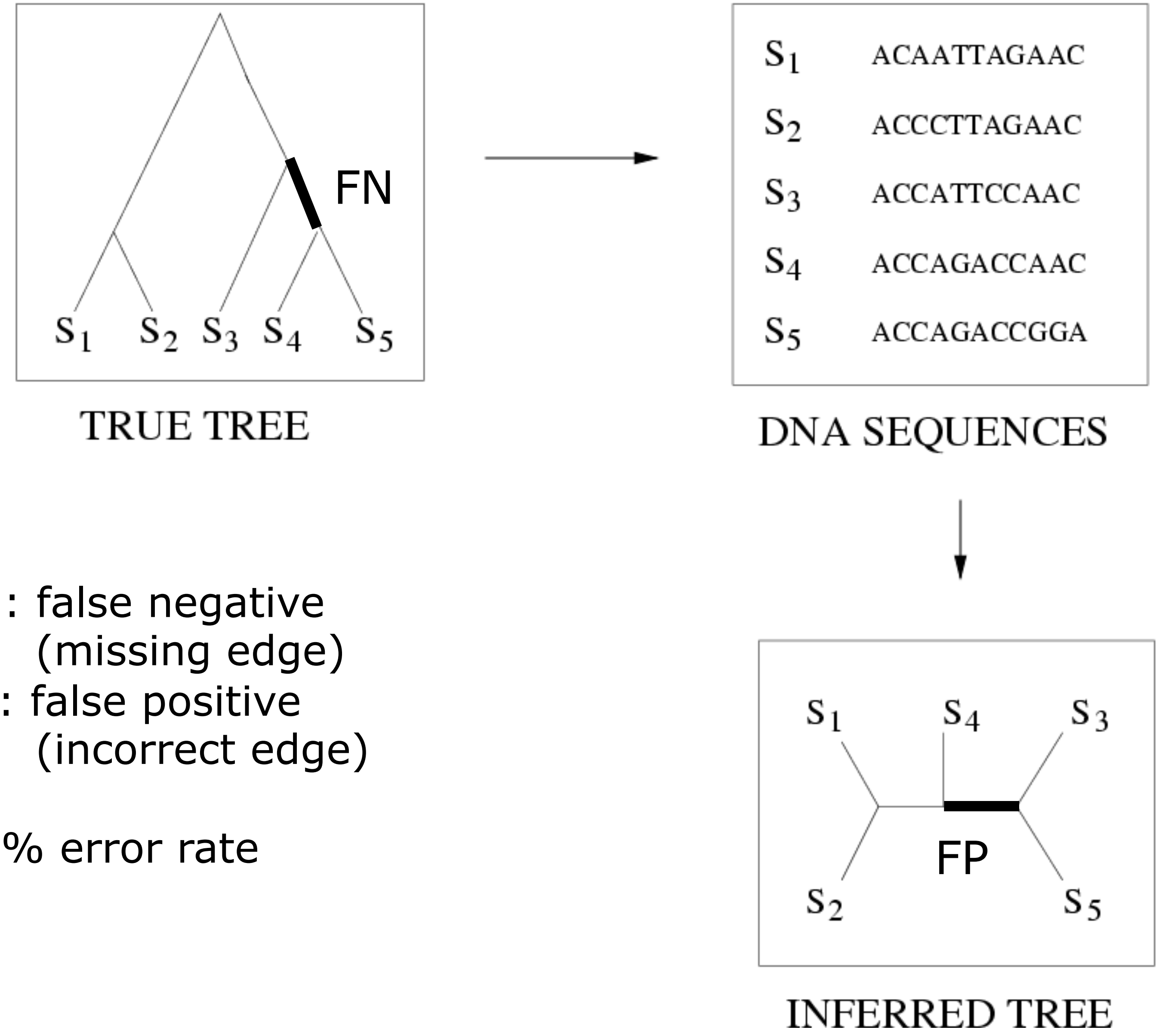


Phylogeny Problem



AGGGCAT TAGCCCA TAGACTT TGCACAA TGCGCTT

Quantifying Error



S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

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

50% error rate

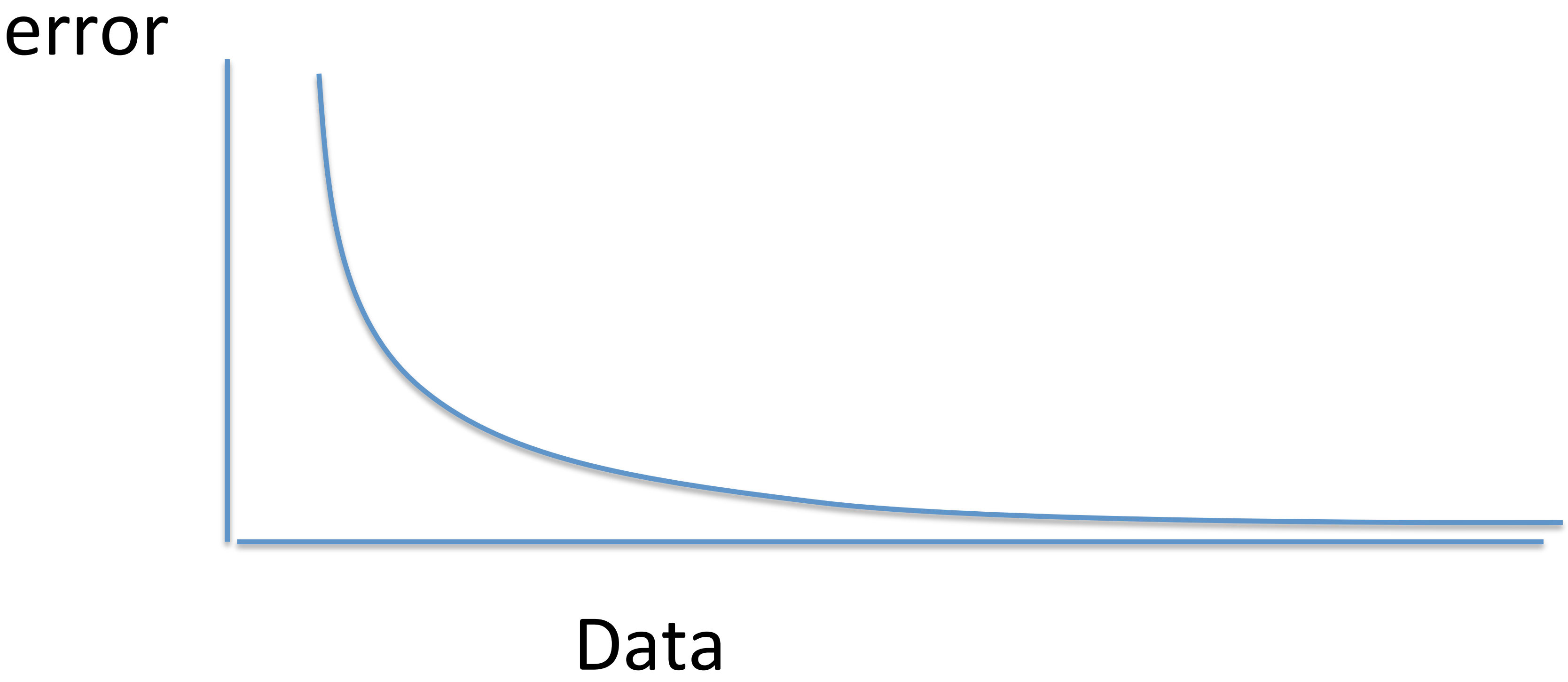
Markov Model of Site Evolution

Simplest (Jukes-Cantor):

- The model tree T is binary and has substitution probabilities p(e) on each edge e.
- The state at the root is randomly drawn from {A,C,T,G} (nucleotides)
- If a site (position) 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.

Statistical Consistency



Data are sites in an alignment

Tree Estimation Methods

- Maximum Likelihood (e.g., RAxML, FastTree, PhyML)
- Bayesian MCMC (e.g., MrBayes)
- Maximum Parsimony (e.g., TNT, PAUP*)
- Distance-based methods (e.g., neighbor joining)
- Quartet-based methods (e.g., Quartet Puzzling)

General Observations

- Maximum Likelihood and Bayesian methods – probably most accurate, have statistical guarantees under many statistical models (e.g., GTR). However, these are often computationally intensive on large datasets.

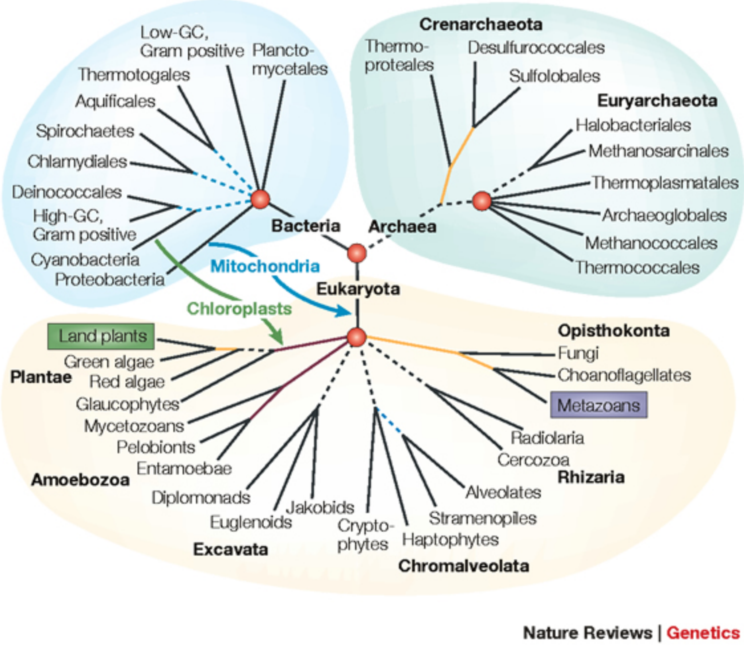
- No statistical guarantees for maximum parsimony (can even produce the incorrect tree with high support) – and MP heuristics are computationally intensive on large datasets.

- Distance-based methods can have statistical guarantees, but may not be so accurate.

General Observations

- Maximum Parsimony and Maximum Likelihood are NP-hard optimization problems, so methods for these are generally heuristic – and may not find globally optimal solutions.
- However, effective heuristics exist that are reasonably good (and considered reliable) for most datasets.
  - MP: TNT (best?) and PAUP* (very good)
  - ML: RAxML (best?), FastTree (even faster but not as thorough), PhyML (not quite as fast but has more models), and others

Estimating The Tree of Life: a *Grand Challenge*



Most well studied problem:

Given DNA sequences, find the Maximum Likelihood Tree

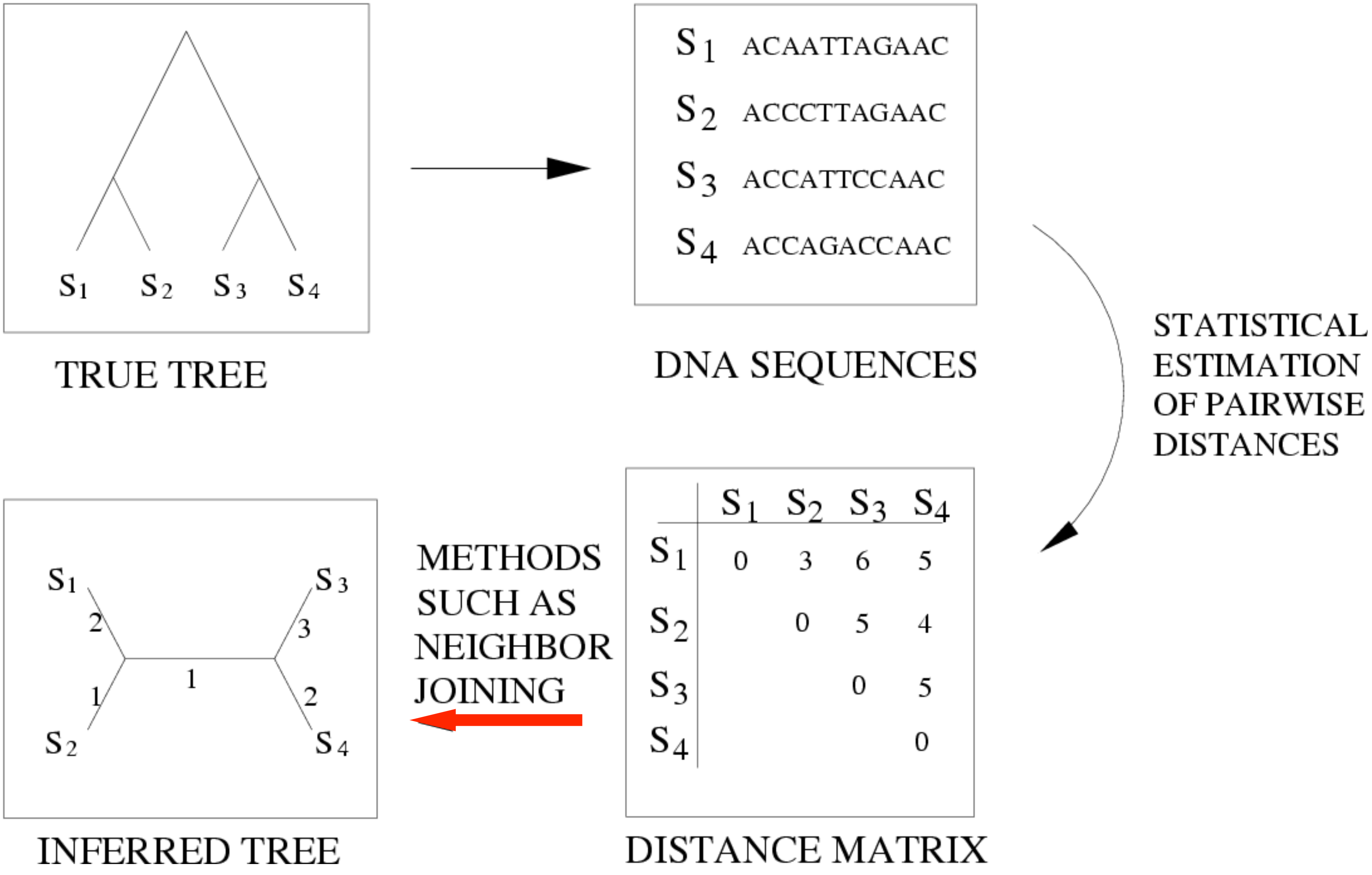
NP-hard, lots of heuristics (RAxML, FastTree-2, PhyML, GARLI, etc.)

More observations

- Bayesian methods: Basic idea – find a distribution of trees with good scores, and so don't return just the single best tree.
- These are even slower than maximum likelihood and maximum parsimony. They require that they are run for a long time so that they "converge". May be best to limit the use of Bayesian methods to small datasets.
- Example: MrBayes.

Distance-based methods

Distance-based estimation



TRUE TREE

S1	ACAATTAGAAC
S2	ACCCTTAGAAC
S3	ACCATTCCAAC
S4	ACCAGACCAAC

DNA SEQUENCES

STATISTICAL
ESTIMATION
OF PAIRWISE
DISTANCES

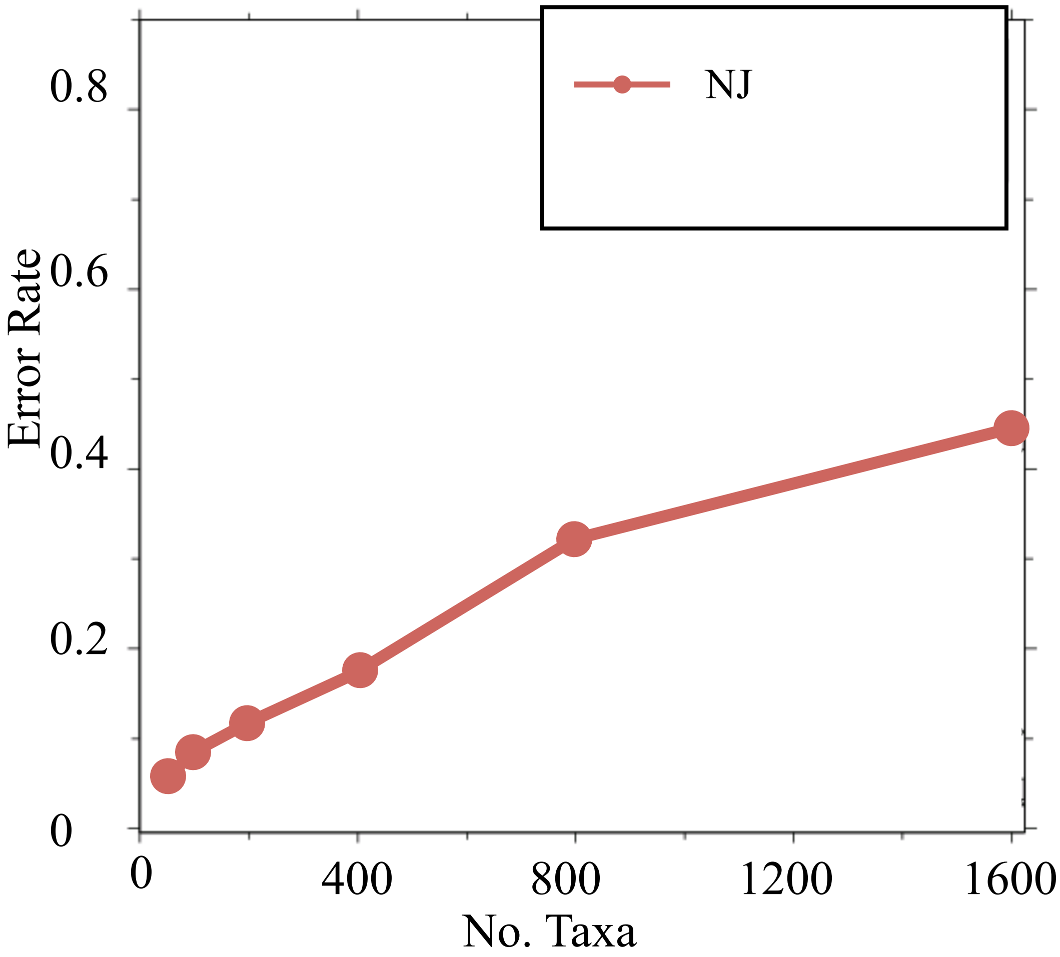
	S1	S2	S3	S4
S1	0	3	6	5
S2		0	5	4
S3			0	5
S4				0

DISTANCE MATRIX

METHODS
SUCH AS
NEIGHBOR
JOINING

INFERRED TREE

Neighbor Joining on large diameter trees

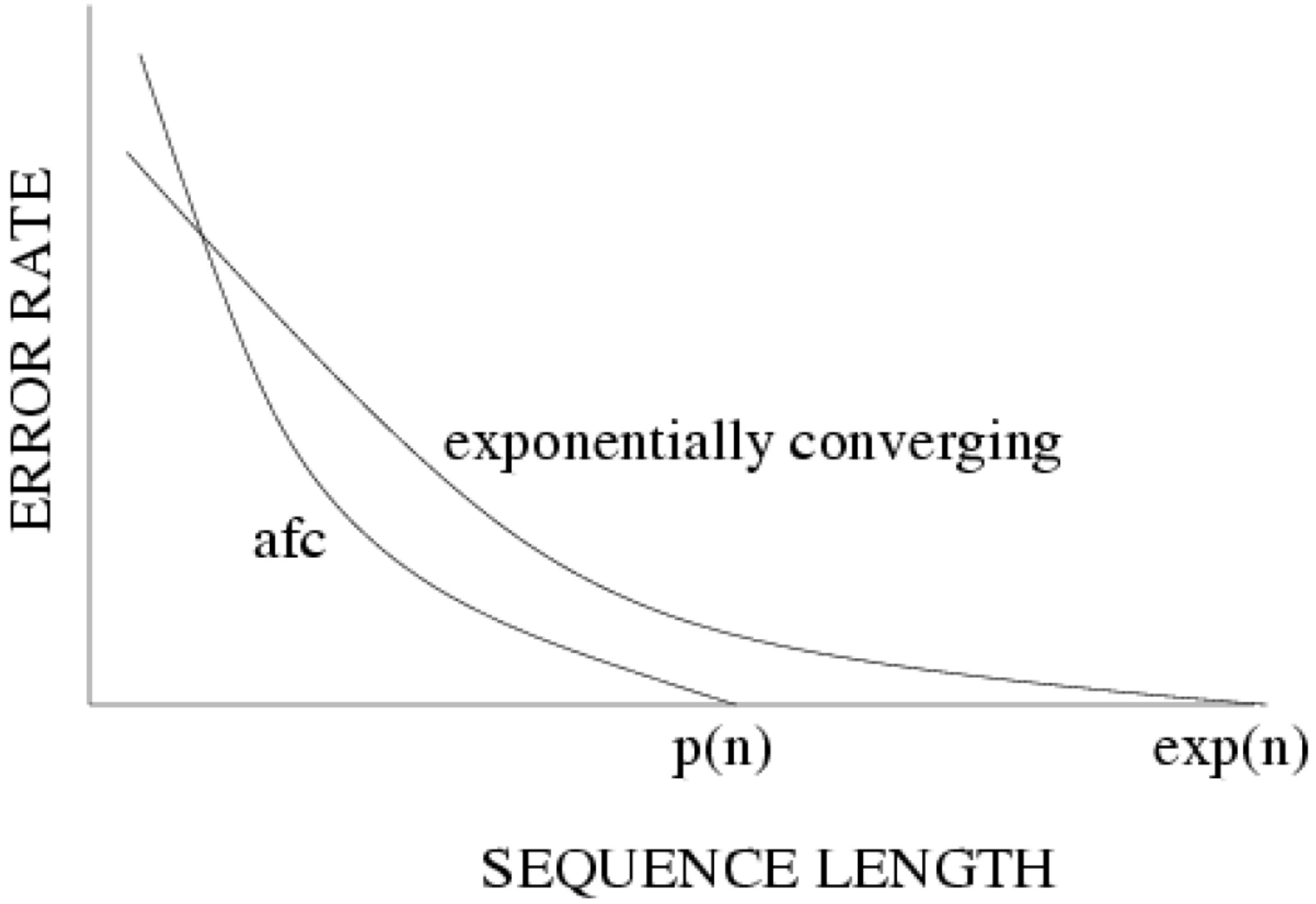


Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

Error rates reflect proportion of incorrect edges in inferred trees.

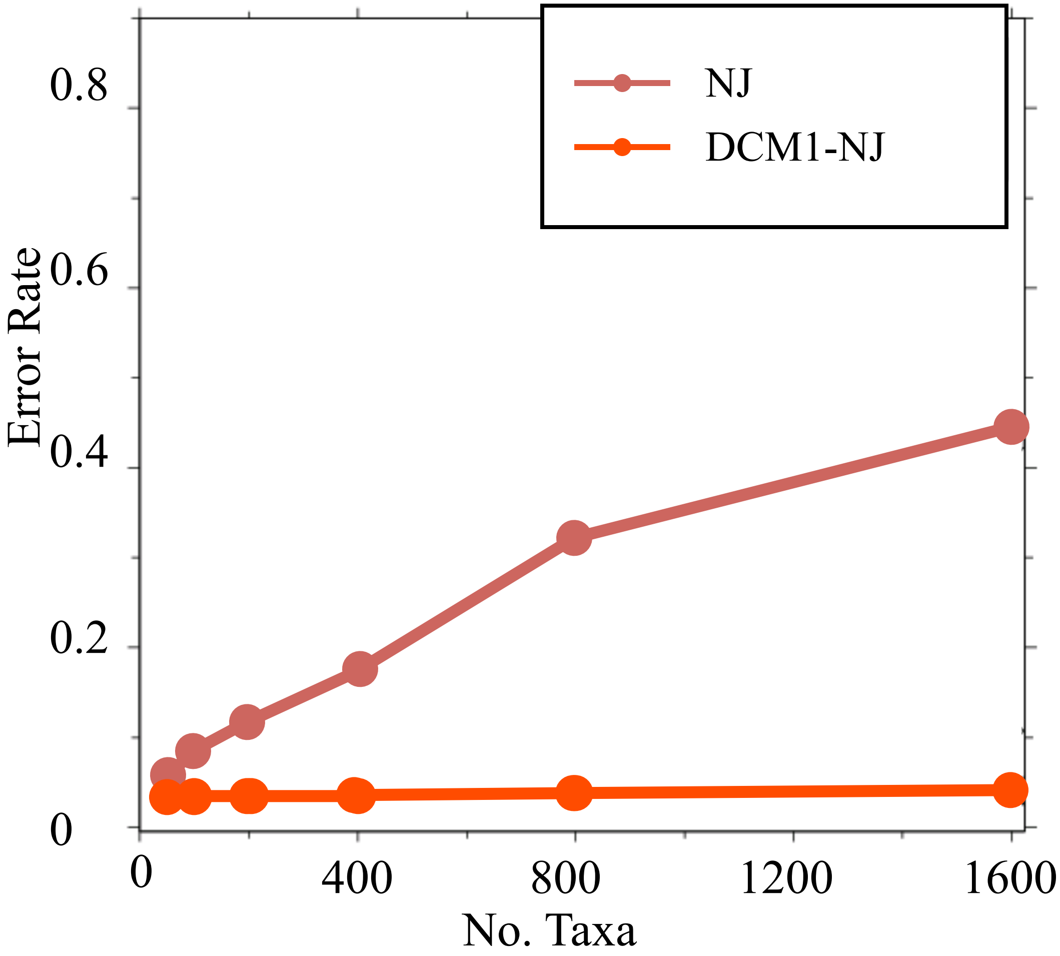
[Nakhleh et al. ISMB 2001]

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



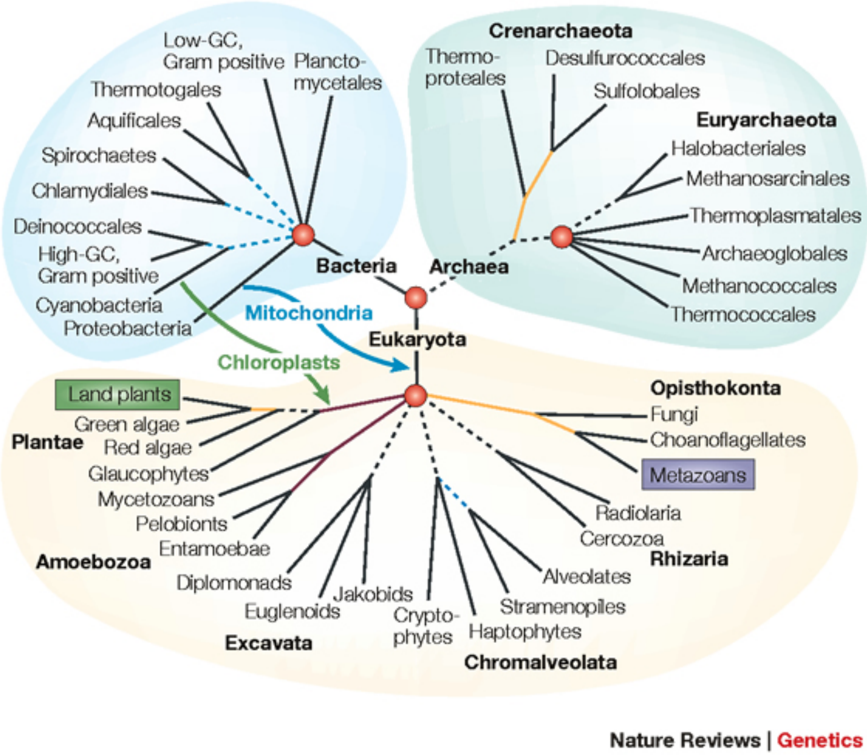
DCM1-boosting distance-based methods

[Nakhleh et al. ISMB 2001]



Theorem (Warnow et al., SODA 2001): DCM1-NJ converges to the true tree from polynomial length sequences

Large-scale Phylogeny: A grand challenge!



Nature Reviews | Genetics

Estimating phylogenies is a complex analytical task

Large datasets are very hard to analyze with high accuracy -- many sites not the same challenge as many taxa!

High Performance Computing is necessary but not sufficient

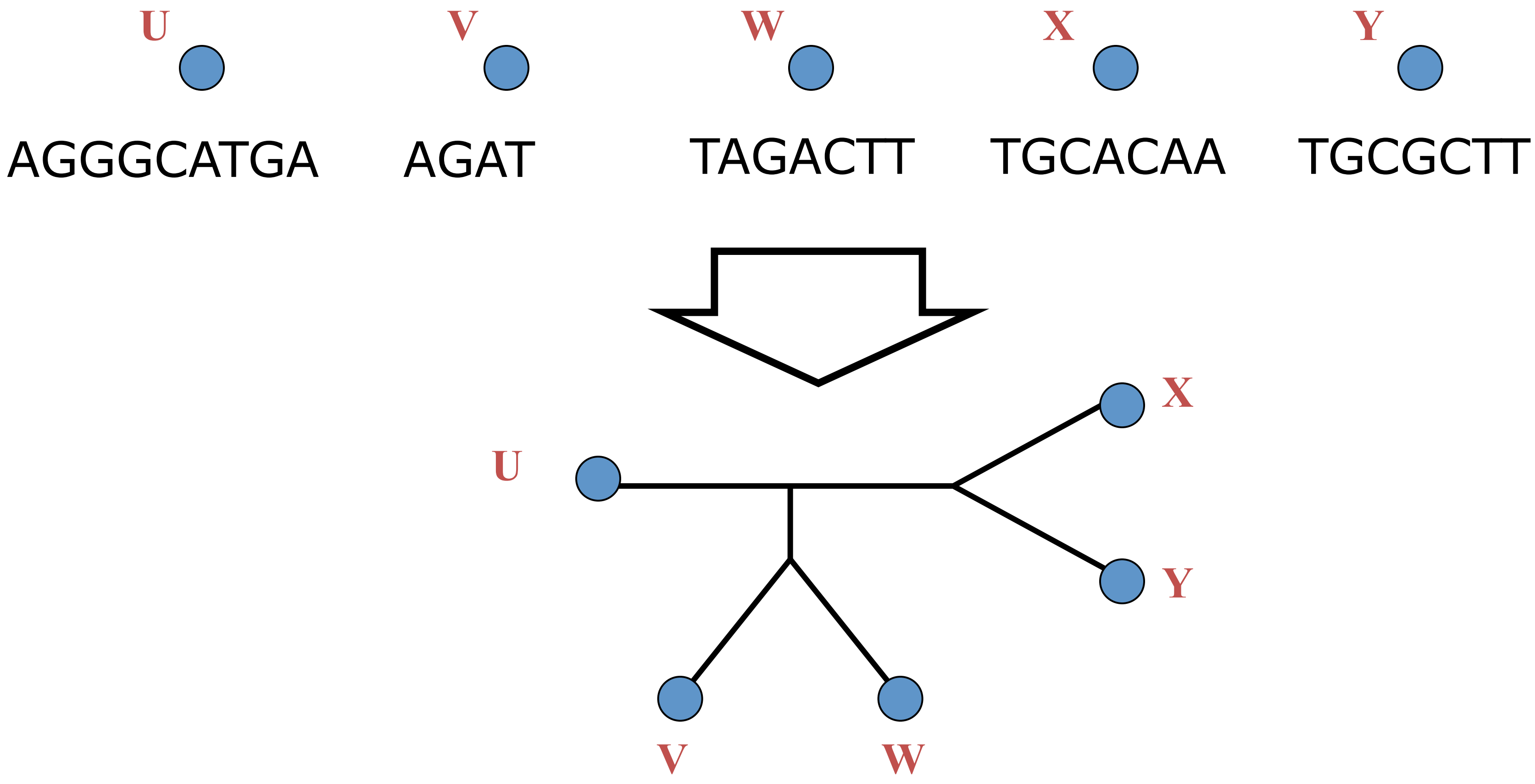
Summary

- Effective heuristics for Maximum likelihood (e.g., RAxML and FastTree) and Bayesian methods (e.g., MrBayes) have statistical guarantees and give good results, but they are slow.
- The best distance-based methods also have statistical guarantees and can give good results, but are not necessarily as accurate as maximum likelihood or Bayesian methods.
- Maximum parsimony has no guarantees, but can give good results. Some effective heuristics exist (TNT, PAUP*).

Summary

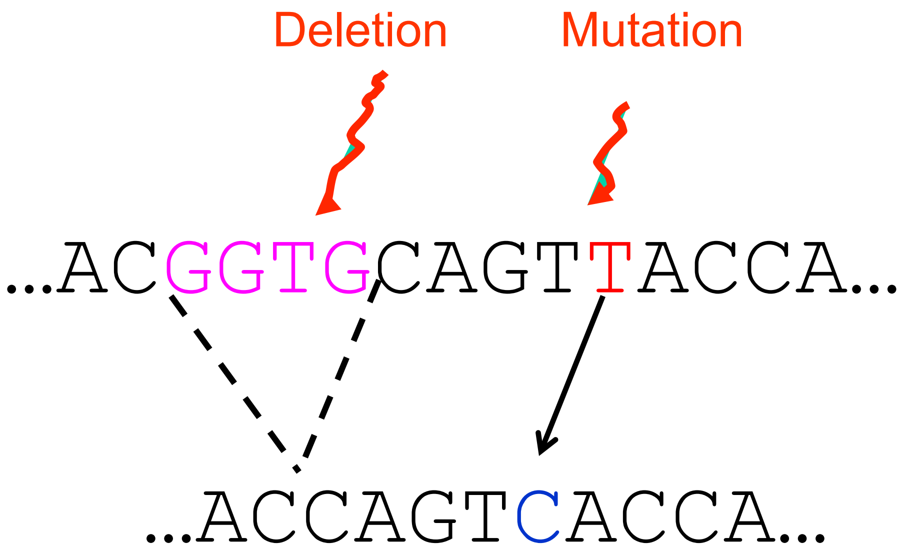
- Effective heuristics for Maximum likelihood (e.g., RAxML and FastTree) and Bayesian methods (e.g., MrBayes) have statistical guarantees and give good results, but they are slow.
- The best distance-based methods also have statistical guarantees and can give good results, but are not necessarily as accurate as maximum likelihood or Bayesian methods.
- Maximum parsimony has no guarantees, but can give good results. Some effective heuristics exist (TNT, PAUP*).
- However, all these results assume the sequences evolve **only with substitutions**.

The “real” problem



AGGGCATGA AGAT TAGACTT TGCACAA TGCGCTT

Indels (insertions and deletions)



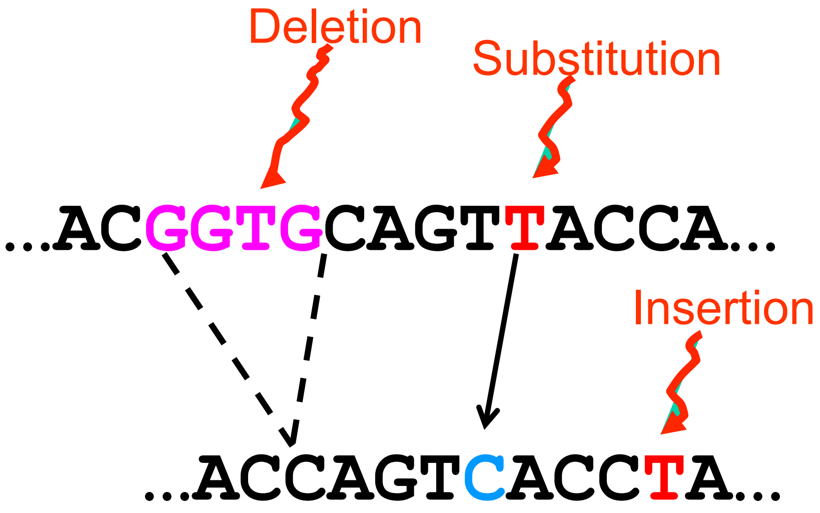
Deletion

Mutation

…ACGGTGCAGTTACCA…

…ACCAGTCACCA…

Multiple Sequence Alignment



Deletion

Substitution

...ACGGTGCAGTTACCA...

Insertion

...ACCAGTCACCTA...

```
...ACGGTGCAGTTACC-A...
...AC----CAGTCACCTA...
```

The **true multiple alignment**

- Reflects historical substitution, insertion, and deletion events
- Defined using transitive closure of pairwise alignments computed on edges of the true tree

Input: unaligned sequences

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

Phase 1: Alignment

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

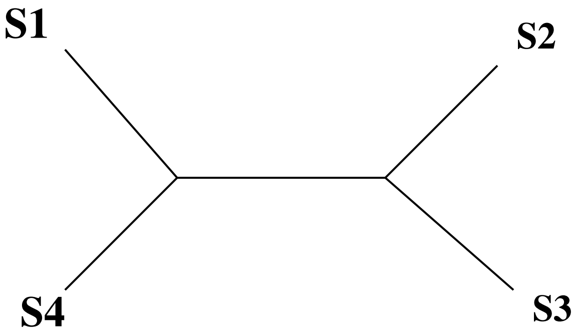
→

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

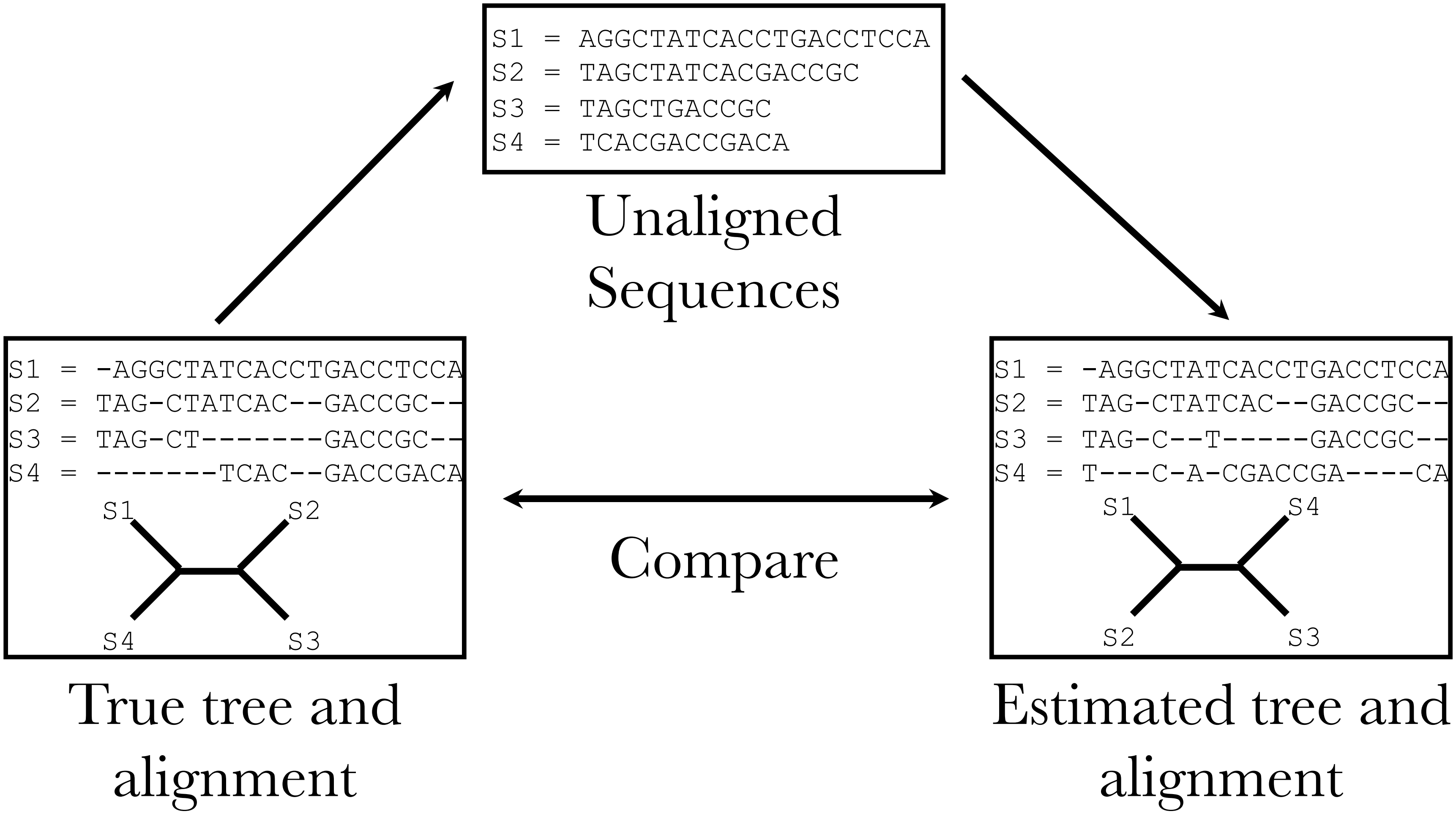
Phase 2: Construct tree

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

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



Simulation Studies



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

Unaligned Sequences

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

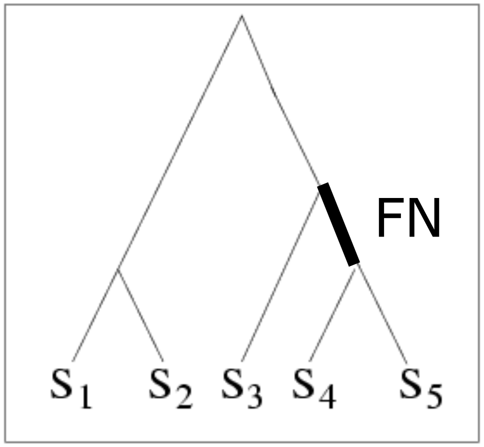
True tree and alignment

Compare

```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-C--T-----GACCGC--
S4 = T---C-A-CGACCGA----CA
```

Estimated tree and alignment

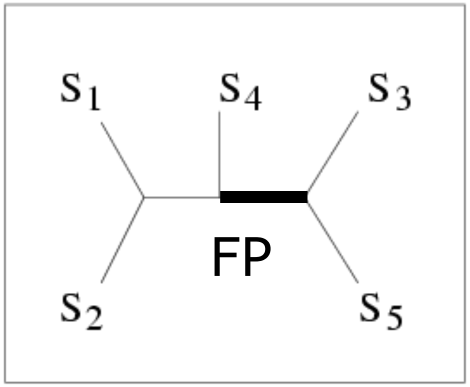
Quantifying Error



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

Two-phase estimation

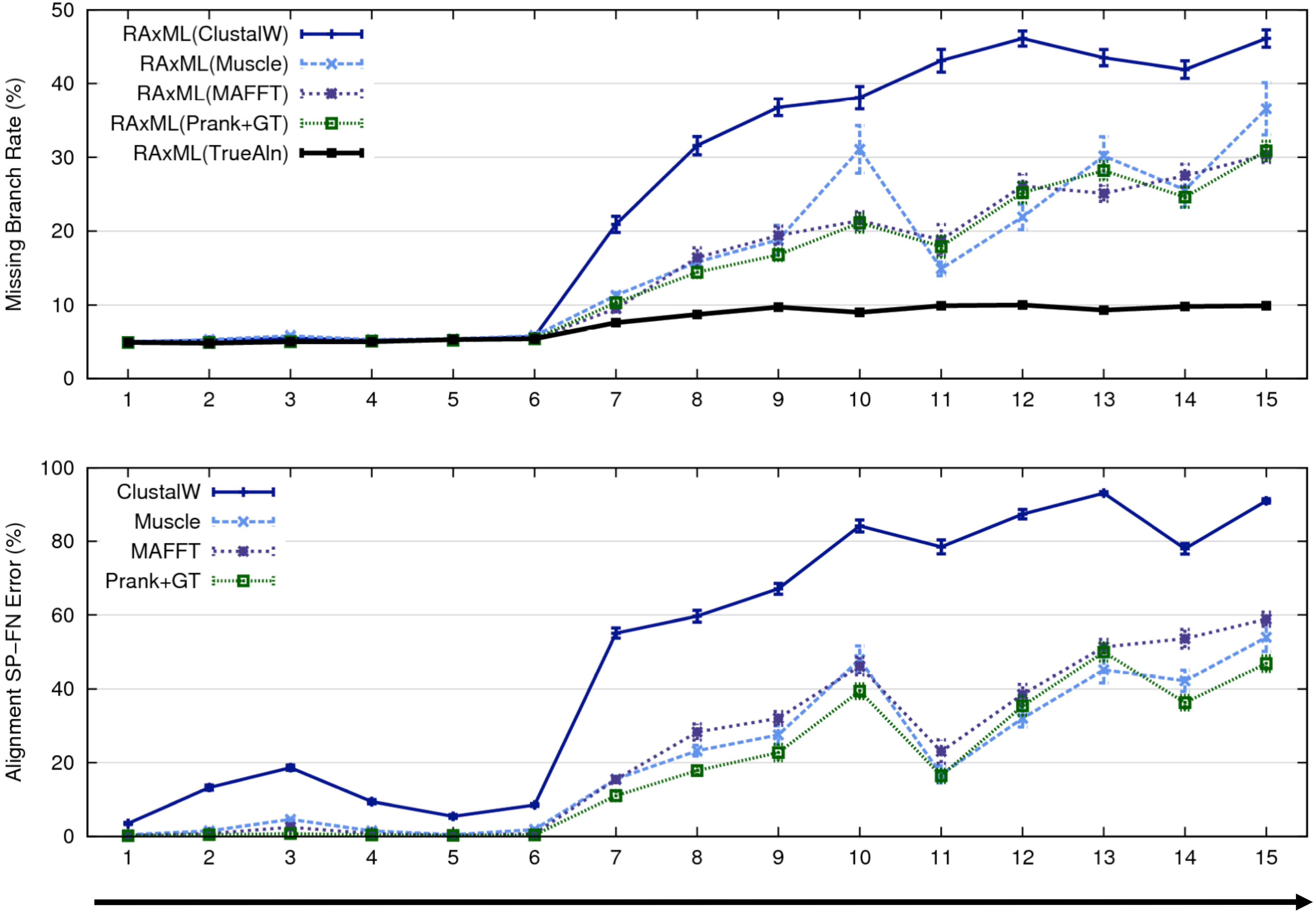
Alignment methods

- 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.

***RAxML**: heuristic for large-scale ML optimization*



1000-taxon models, ordered by difficulty (Liu et al., 2009)

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

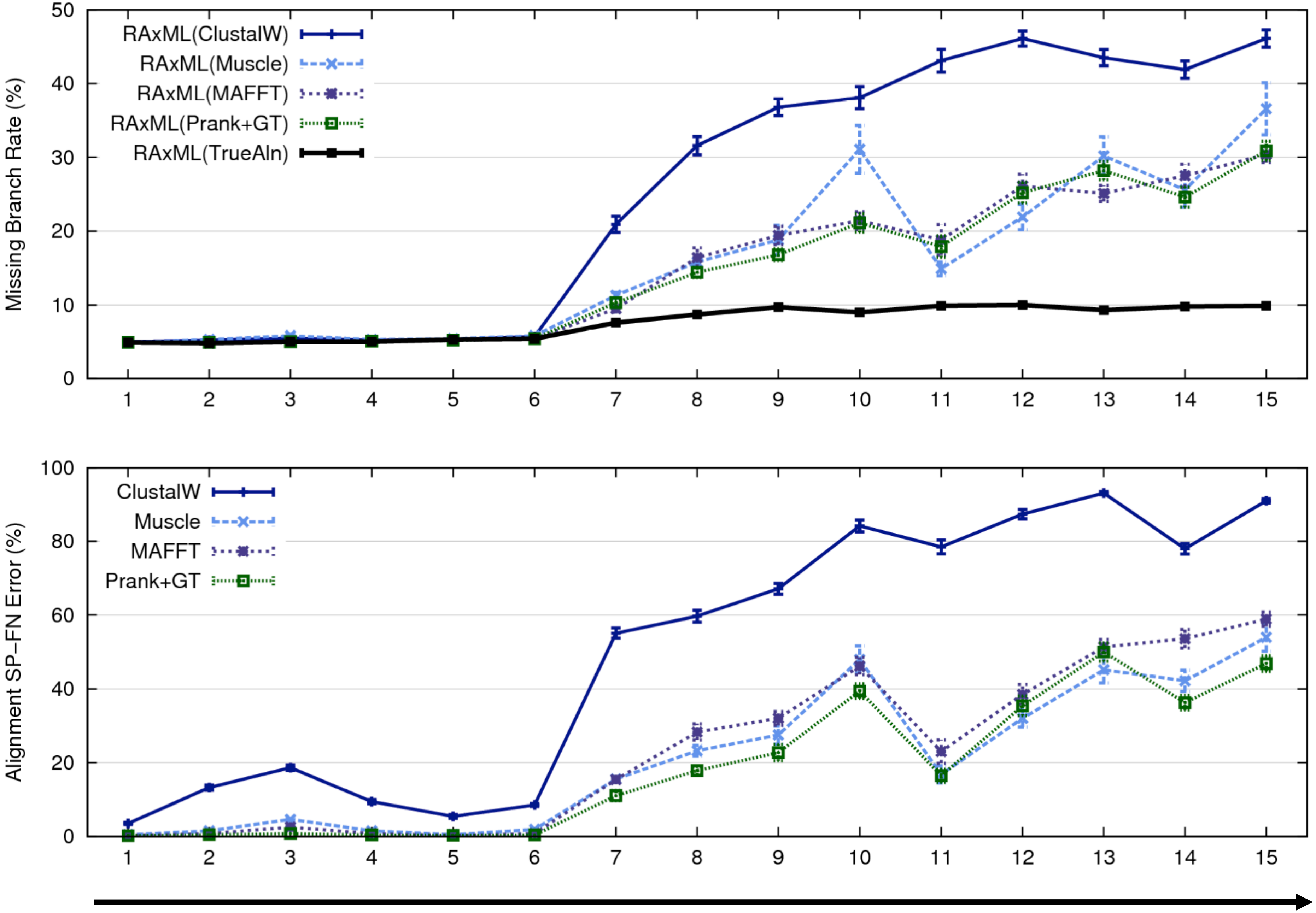
Many important applications besides phylogenetic estimation

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

SATé

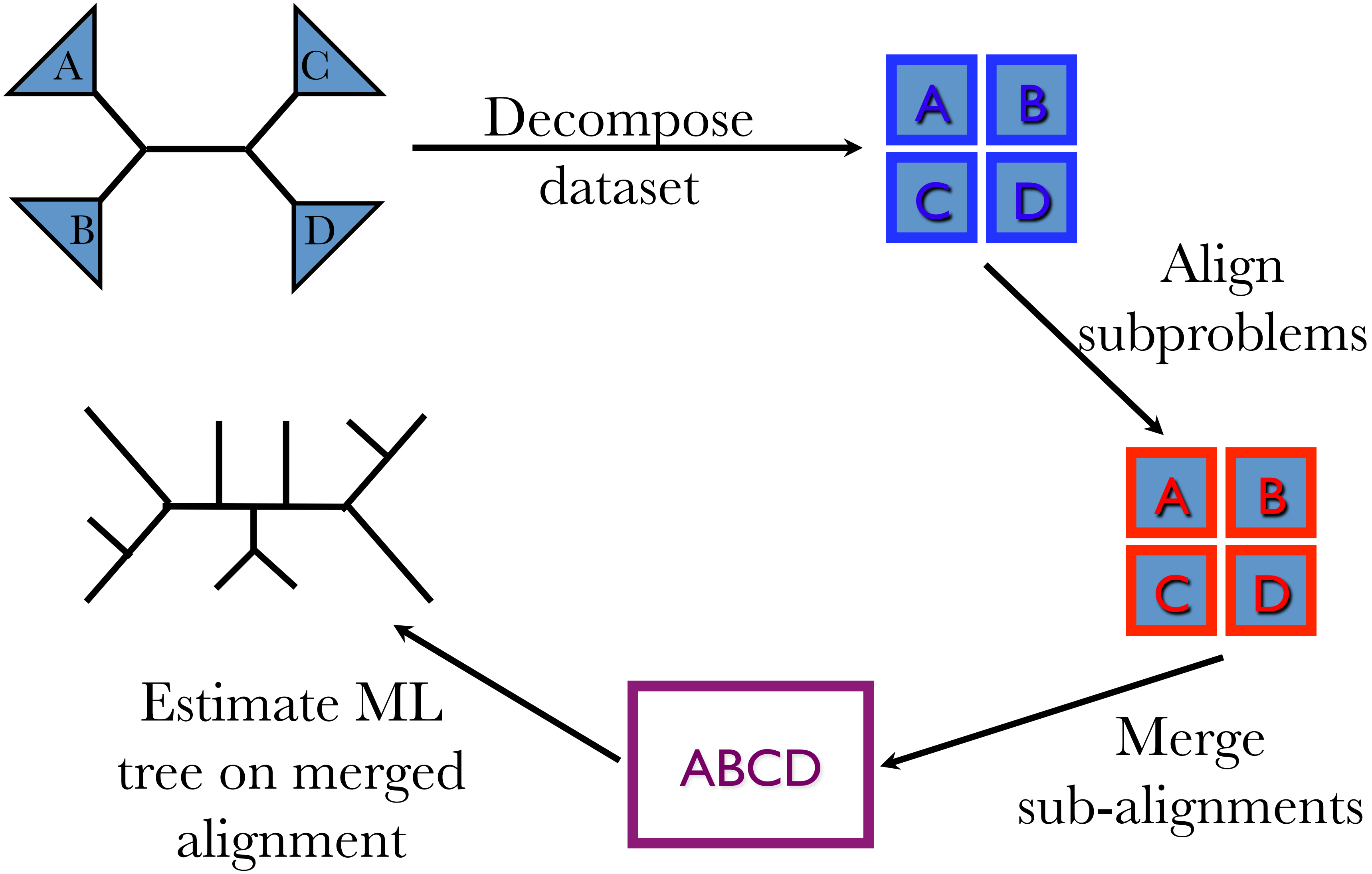
SATé (Simultaneous Alignment and Tree Estimation)

- Liu et al., Science 2009
- Liu et al., Systematic Biology 2012
- Public distribution (open source software) and user-friendly GUI



1000-taxon models, ordered by difficulty (Liu et al., 2009)

Re-aligning on a tree



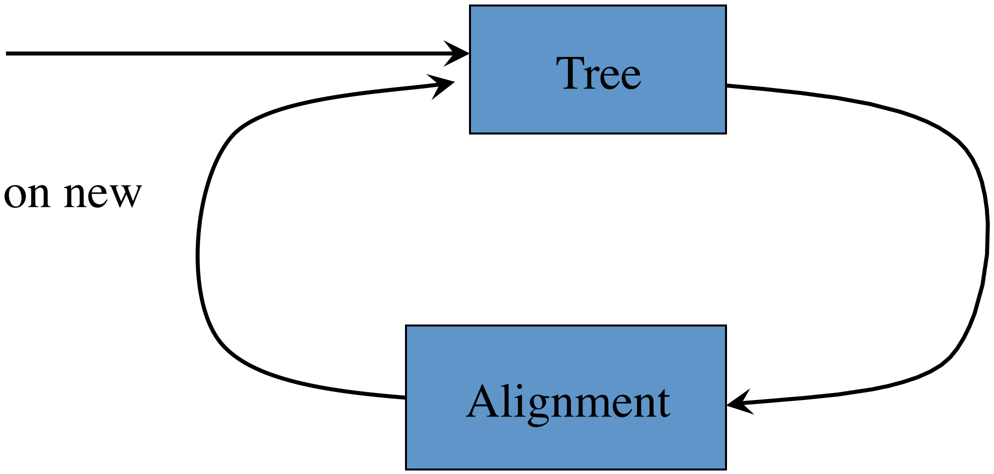
Decompose dataset

Align subproblems

Merge sub-alignments

Estimate ML tree on merged alignment

SATé Algorithm



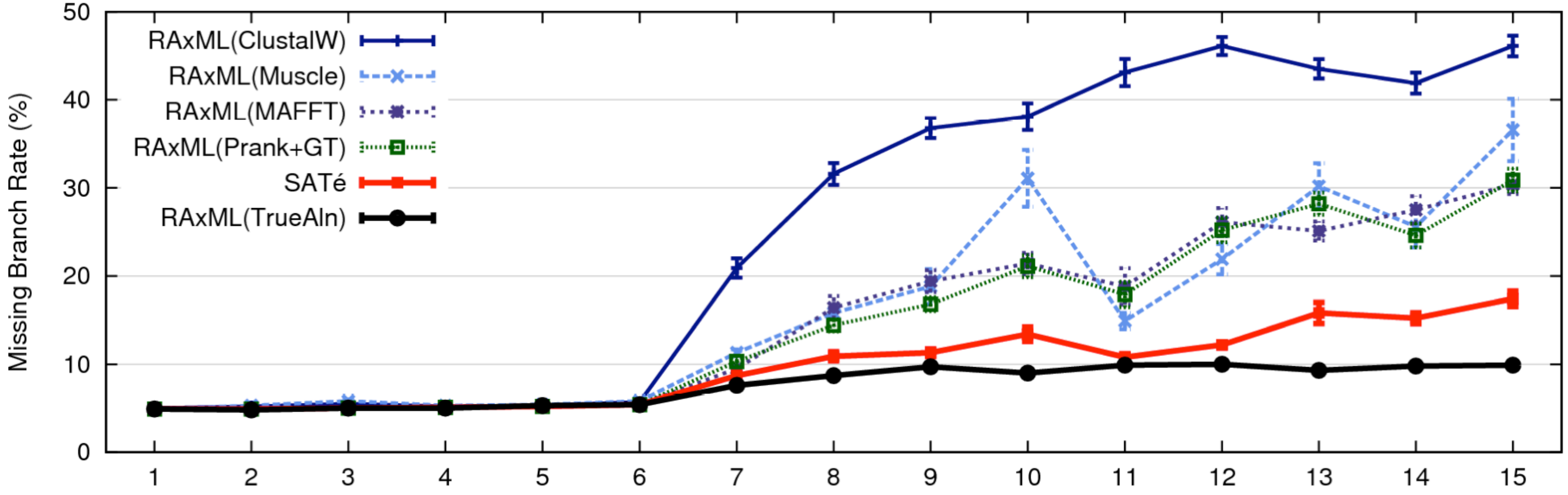
Obtain initial alignment and estimated ML tree

Use tree to compute new alignment

Estimate ML tree on new alignment

If new alignment/tree pair has worse ML score, realign using a different decomposition

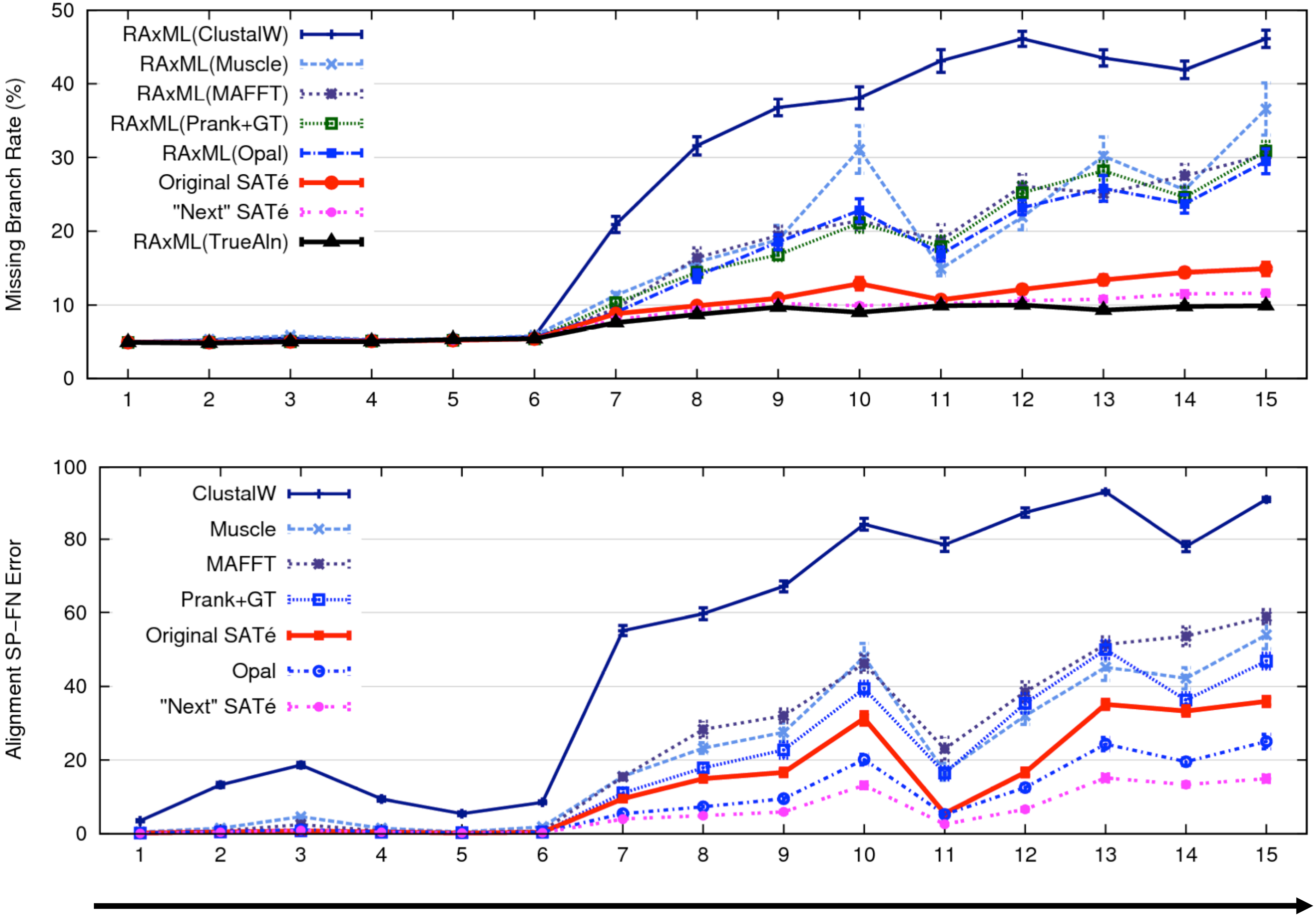
Repeat until termination condition (typically, 24 hours)



1000 taxon models, ordered by difficulty

24 hour SATé analysis, on desktop machines

(Similar improvements for biological datasets)

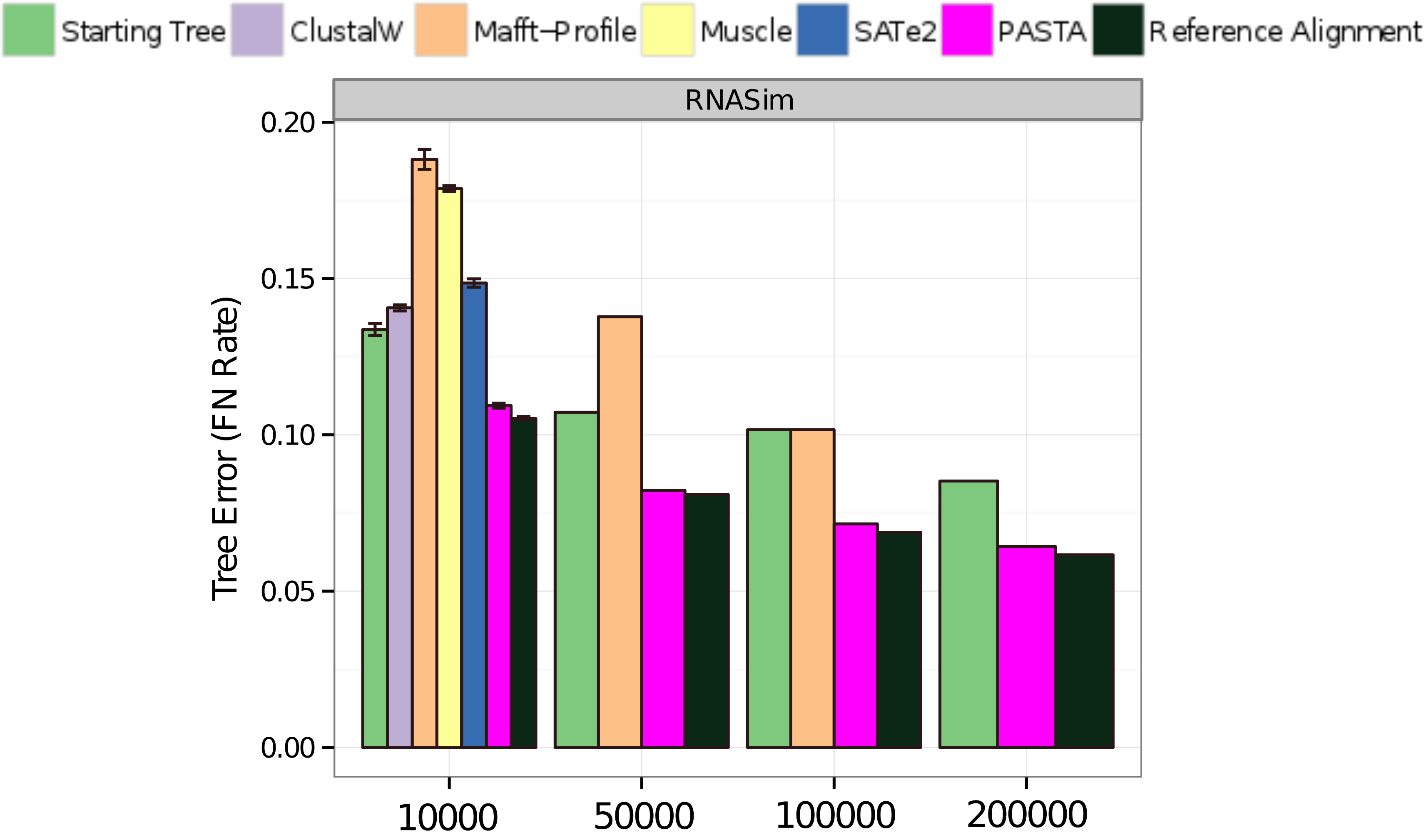


1000 taxon models ranked by difficulty

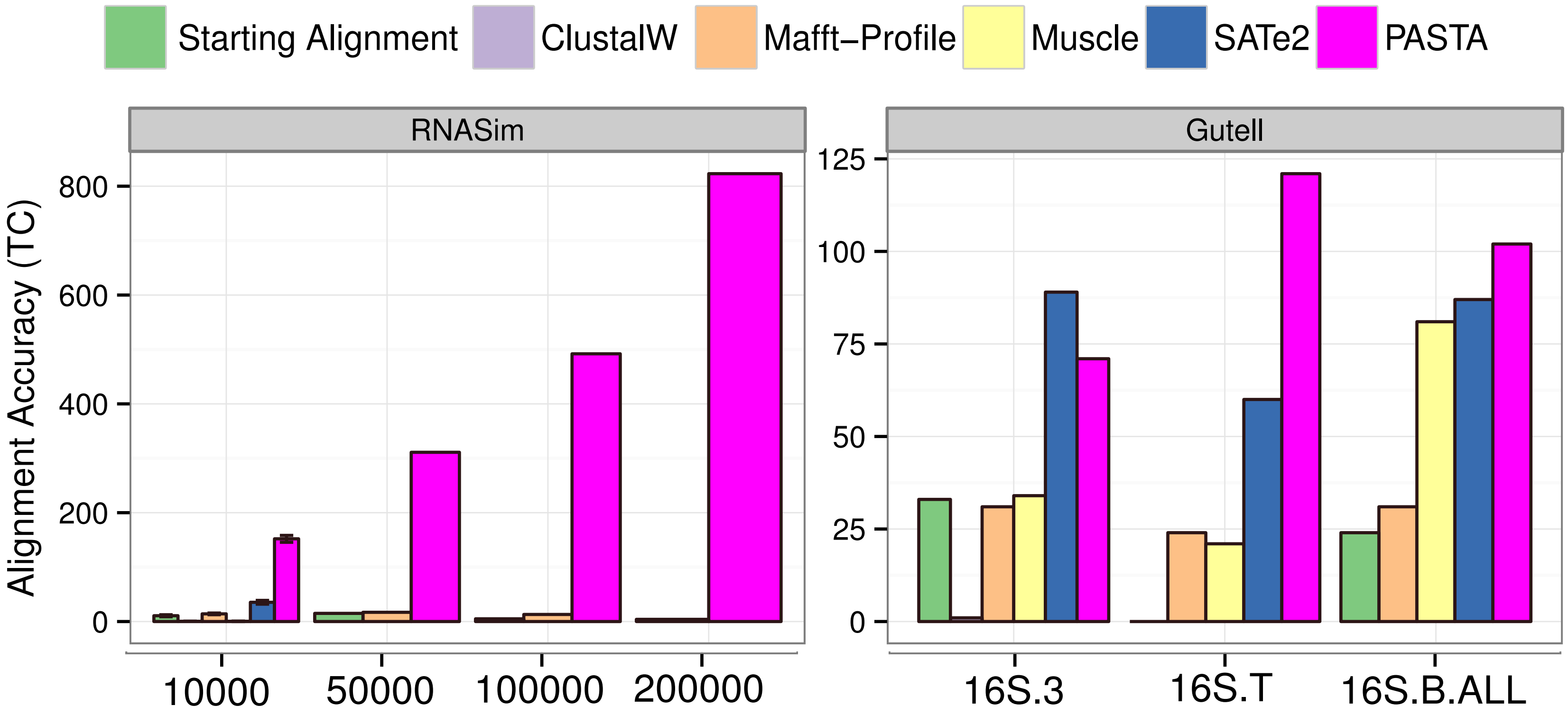
SATé and PASTA

- SATé-1 (Science 2009) can analyze 10,000 sequences
- SATé-2 (Systematic Biology 2012) can analyze 50,000 sequences, is faster and more accurate than SATé-1
- PASTA (RECOMB 2014) can analyze 200,000 sequences, and is faster and more accurate than both SATé versions.

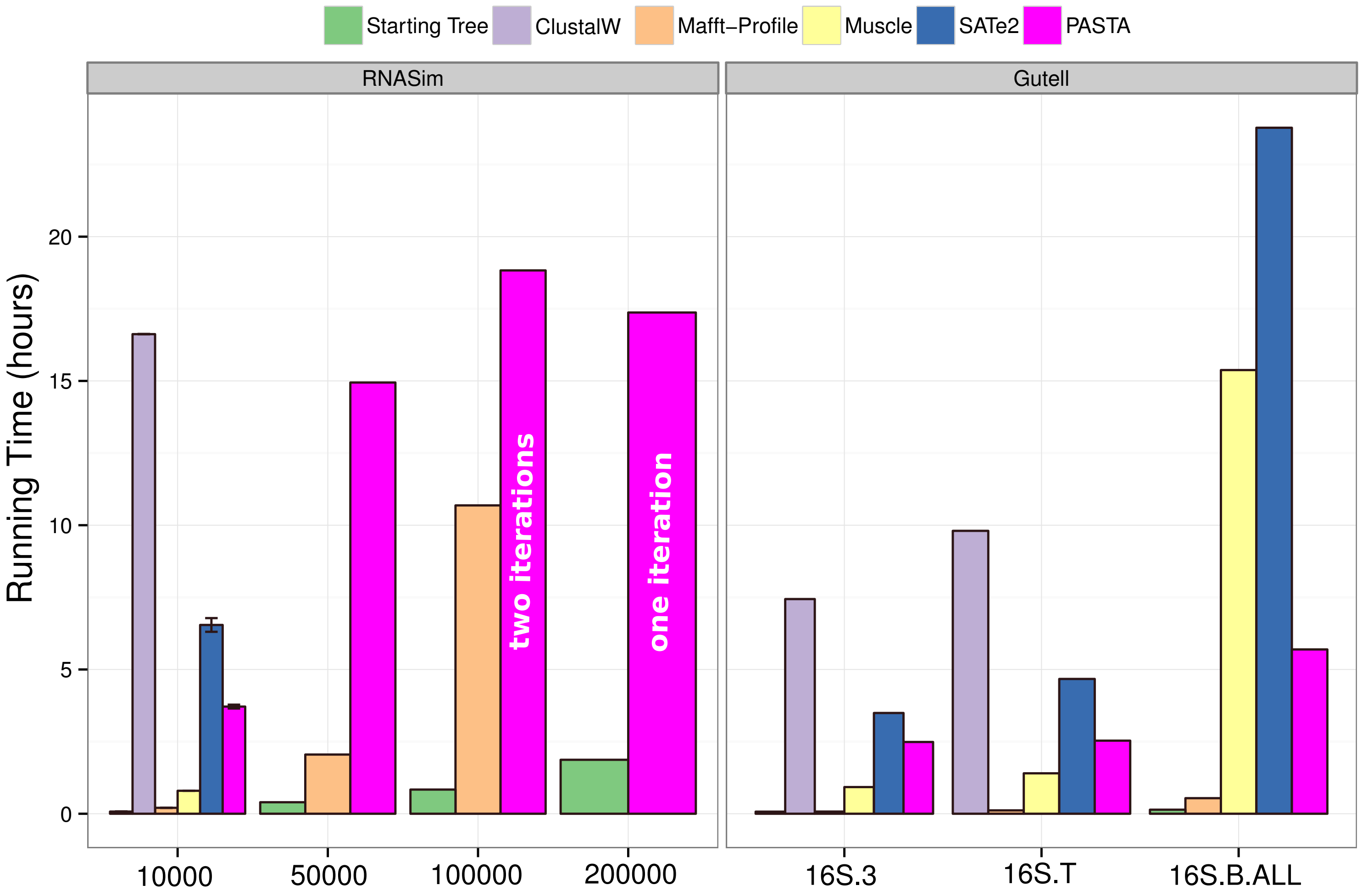
Tree Error – Simulated data



Alignment Accuracy – Correct columns



Running time



PASTA – tutorial tomorrow

- PASTA: Practical Alignments using SATé and TrAnsitivity (Published in RECOMB 2014)
- Developers: Siavash Mirarab, Nam Nguyen, and Tandy Warnow
- GOOGLE user group
- Paper online at http://www.cs.utexas.edu/~tandy/pasta-download.pdf
- Software at http://www.cs.utexas.edu/users/phylo/software/pasta/

Co-estimation

- PASTA and SATé co-estimate the multiple sequence alignment and its ML tree, but this co-estimation is not performed under a statistical model of evolution that considers indels.
- Instead, indels are treated as "missing data". This is the default for ML phylogeny estimation. (Other options exist but do not necessarily improve topological accuracy.)
- Other methods (such as SATCHMO, for proteins) also perform co-estimation, but similarly are not based on statistical models that consider indels.

Other co-estimation methods

Statistical methods:

- BAli-Phy (Redelings and Suchard): Bayesian software to co-estimate alignments and trees under a statistical model of evolution that includes indels. Can scale to about 100 sequences, but takes a very long time.
  - http://www.bali-phy.org/
- StatAlign: http://statalign.github.io/

Extensions of Parsimony

- POY (most well known software)
  - http://www.amnh.org/our-research/computational-sciences/research/projects/systematic-biology/poy
- BeeTLe (Liu and Warnow, PLoS One 2012)

1kp: Thousand Transcriptome Project

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

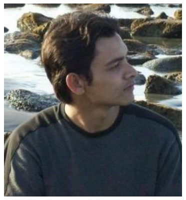
N. Matasci
iPlant

T. Warnow,
UT-Austin

S. Mirarab,
UT-Austin

N. Nguyen,
UT-Austin

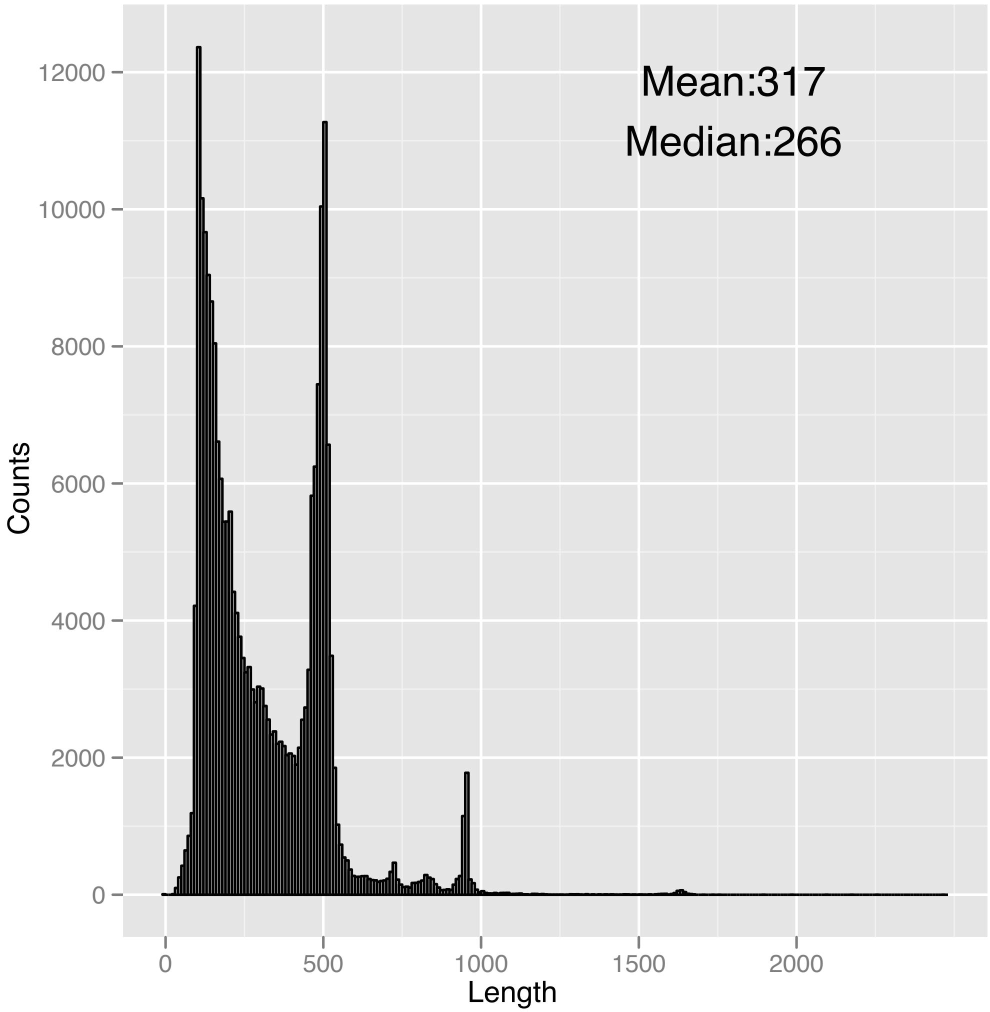
Md. S.Bayzid
UT-Austin



Plus many many other people…

- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)

Challenge:
Alignment of datasets with > 100,000 sequences with many fragmentary sequences



1KP dataset:

More than 100,000 sequences
Lots of fragmentary sequences

Mixed Datasets

- Some sequences are very short – much shorter than the full-length sequences – and some are full-length (so mixture of lengths)

- Estimating a multiple sequence alignment on datasets with some fragments is very difficult (research area)

- Trees based on MSAs computed on datasets with fragments have high error

- Occurs in transcriptome datasets, or in metagenomic analyses

Phylogenies from “mixed” datasets

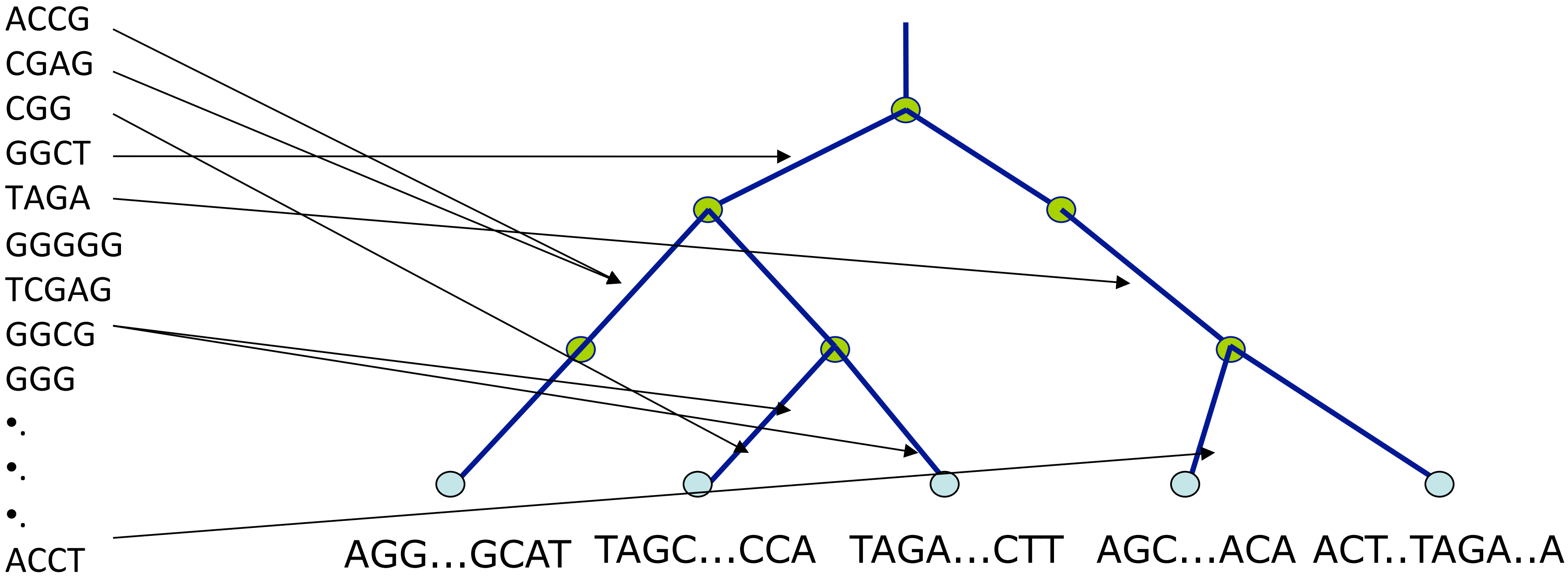
Challenge: Given set of sequences, some full length and some fragmentary, how do we estimate a tree?

- Step 1: Extract the full-length sequences, and get MSA and tree
- Step 2: Add the remaining sequences (short ones) into the tree.

Phylogenetic Placement

Fragmentary sequences from some gene

Full-length sequences for same gene, and an alignment and a tree



ACCG
CGAG
CGG
GGCT
TAGA
GGGGG
TCGAG
GGCG
GGG
•.
•.
•.
ACCT

AGG…GCAT TAGC…CCA TAGA…CTT AGC…ACA ACT..TAGA..A

Phylogenetic Placement

- Input: Tree and MSA on full-length sequences (called the “backbone tree and backbone MSA”) and a set of “query sequences” (that can be very short)
- Output: placement of each query sequence into the “backbone” tree

Several methods for Phylogenetic Placement developed in the last few years

Phylogenetic Placement

Step 1: Align each query sequence to backbone alignment

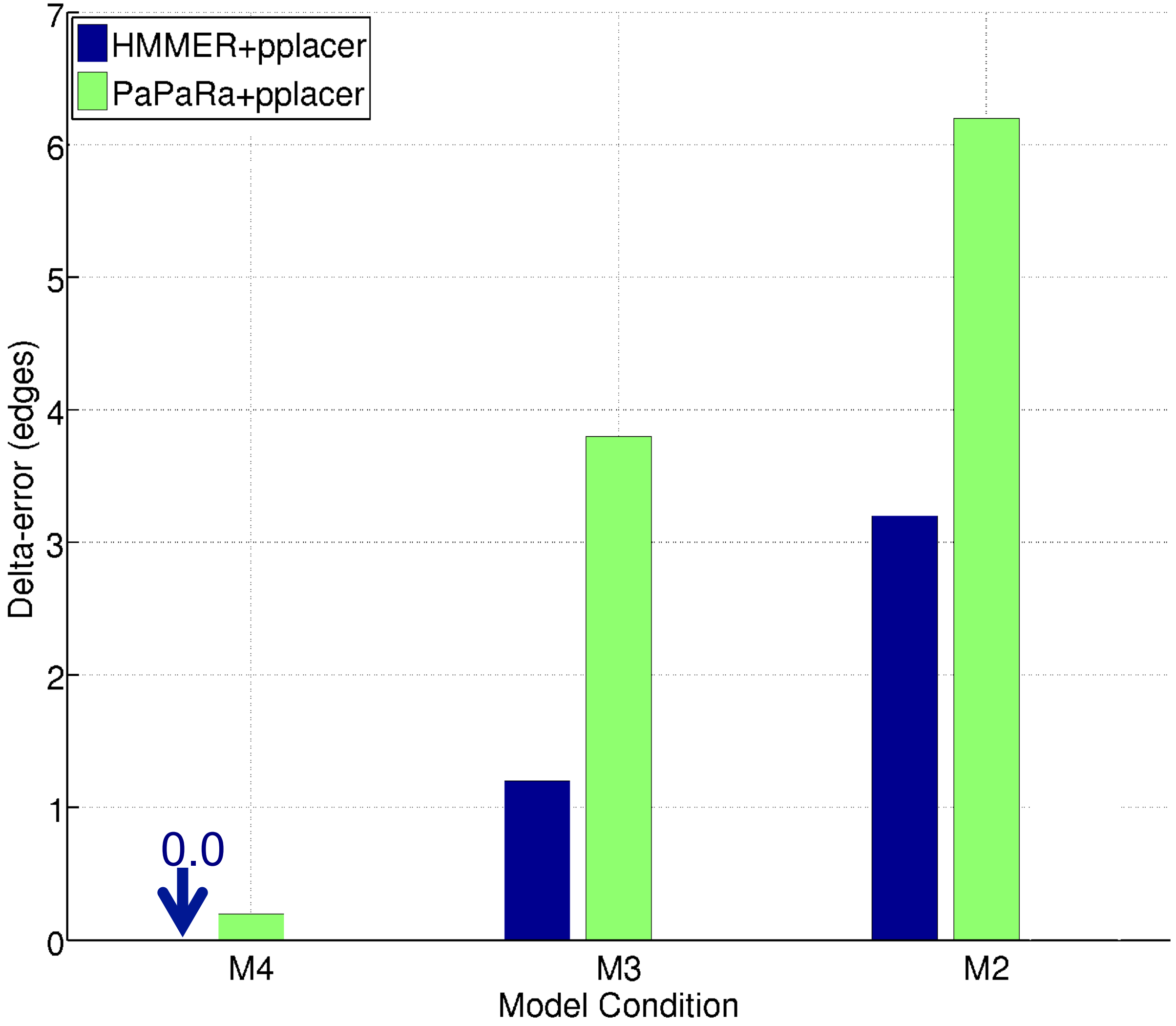
Step 2: Place each query sequence into backbone tree, using extended alignment

Phylogenetic Placement

- Align each query sequence to backbone alignment
  - HMMALIGN (Eddy, Bioinformatics 1998)
  - PaPaRa (Berger and Stamatakis, Bioinformatics 2011)
- Place each query sequence into backbone tree
  - Pplacer (Matsen et al., BMC Bioinformatics, 2011)
  - EPA (Berger and Stamatakis, Systematic Biology 2011)

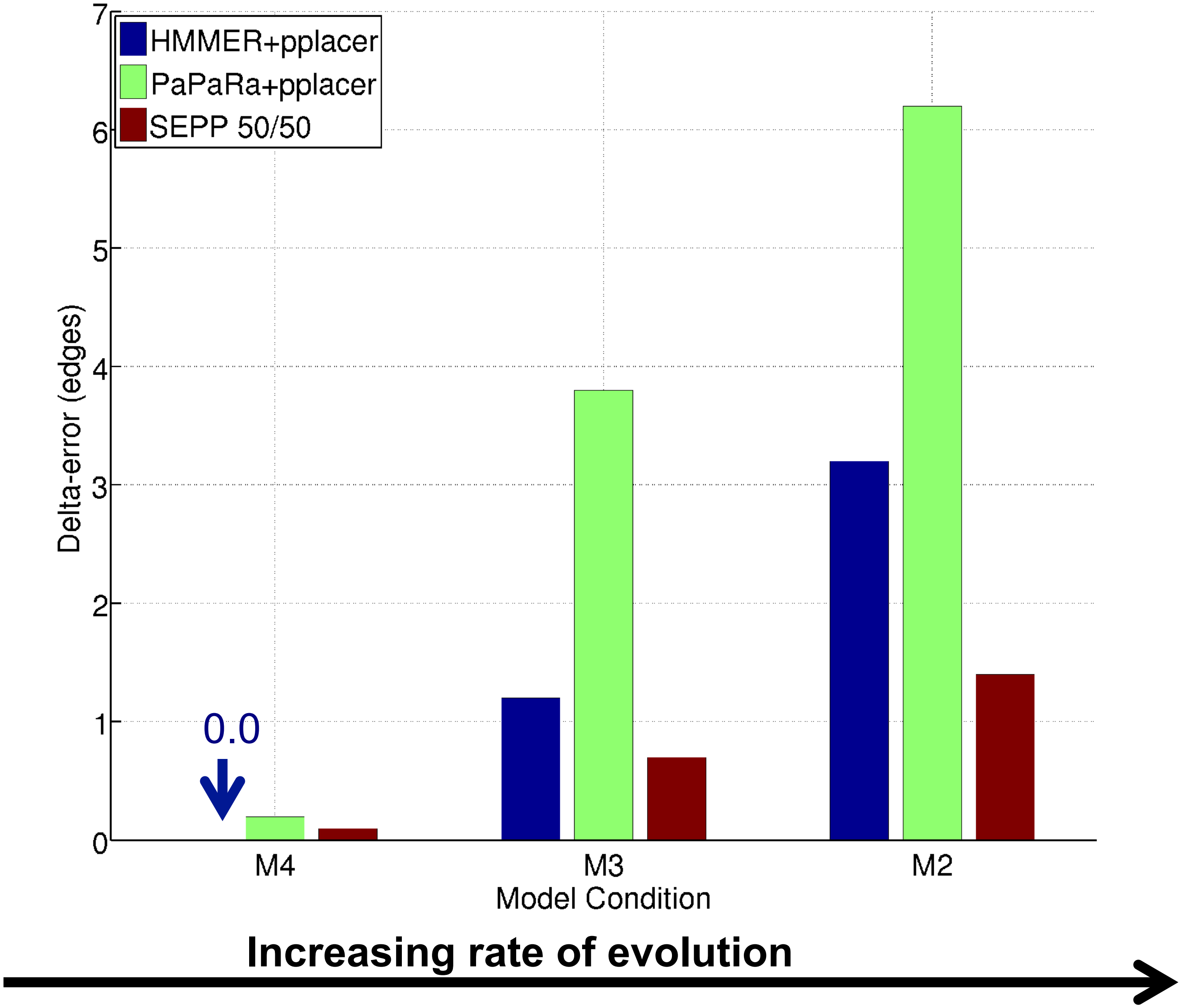
Note: pplacer and EPA use maximum likelihood, and are reported to have the same accuracy.

HMMER vs. PaPaRa placement error



Increasing rate of evolution

SEPP(10%), based on ~10 HMMs



SEPP

- SEPP = SATé-enabled Phylogenetic Placement
- Developers: Nam Nguyen, Siavash Mirarab, and Tandy Warnow
- Software available at https://github.com/smirarab/sepp
- Paper available at http://psb.stanford.edu/psb-online/proceedings/psb12/mirarab.pdf
- Tutorial on Thursday

Summary so far

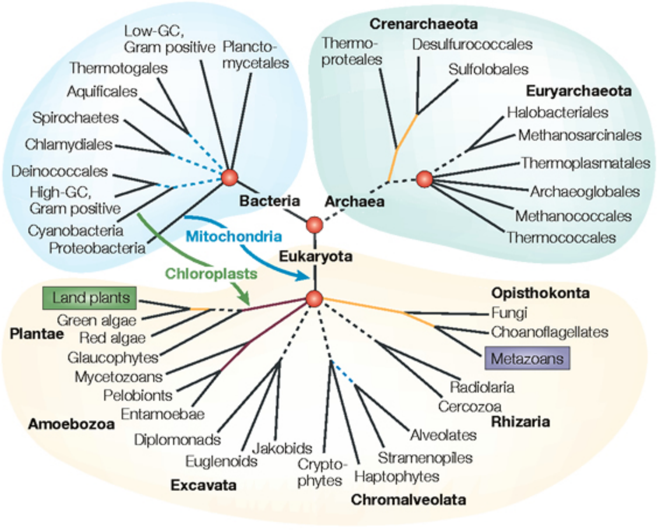
- Great progress in multiple sequence alignment, even for very large datasets with high rates of evolution – provided all sequences are full-length.
- Trees based on good MSA methods (e.g., MAFFT for small enough datasets, PASTA for large datasets) can be highly accurate – but sequence length limitations reduces tree accuracy.
- Handling fragmentary sequences is challenging, but phylogenetic placement is helpful.
- However, all of this is just for a single gene (more generally, a single location in the genome) – no rearrangements, duplications, etc.

Summary so far

- Great progress in multiple sequence alignment, even for very large datasets with high rates of evolution – provided all sequences are full-length.
- Trees based on good MSA methods (e.g., MAFFT for small enough datasets, PASTA for large datasets) can be highly accurate – but sequence length limitations reduces tree accuracy.
- Handling fragmentary sequences is challenging, but phylogenetic placement is helpful.
- However, all of this is just for a single gene (more generally, a single location in the genome) – no rearrangements, duplications, etc.

Phylogenomics

(Phylogenetic estimation from whole genomes)



Nature Reviews | Genetics

Species Tree Estimation

Not all genes present in all species

	gene 1
S_1	TCTAATGGAA
S_2	GCTAAGGGAA
S_3	TCTAAGGGAA
S_4	TCTAACGGAA
S_7	TCTAATGGAC
S_8	TATAACGGAA

	gene 2
S_4	GGTAACCCTC
S_5	GCTAAACCTC
S_6	GGTGACCATC
S_7	GCTAAACCTC

	gene 3
S_1	TATTGATACA
S_3	TCTTGATACC
S_4	TAGTGATGCA
S_7	TAGTGATGCA
S_8	CATTCATACC

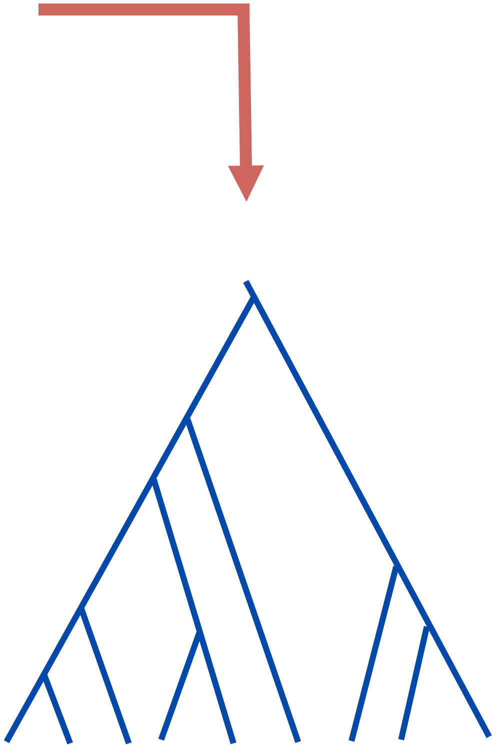
Two basic approaches for species tree estimation

- Concatenate (“combine”) sequence alignments for different genes, and run phylogeny estimation methods

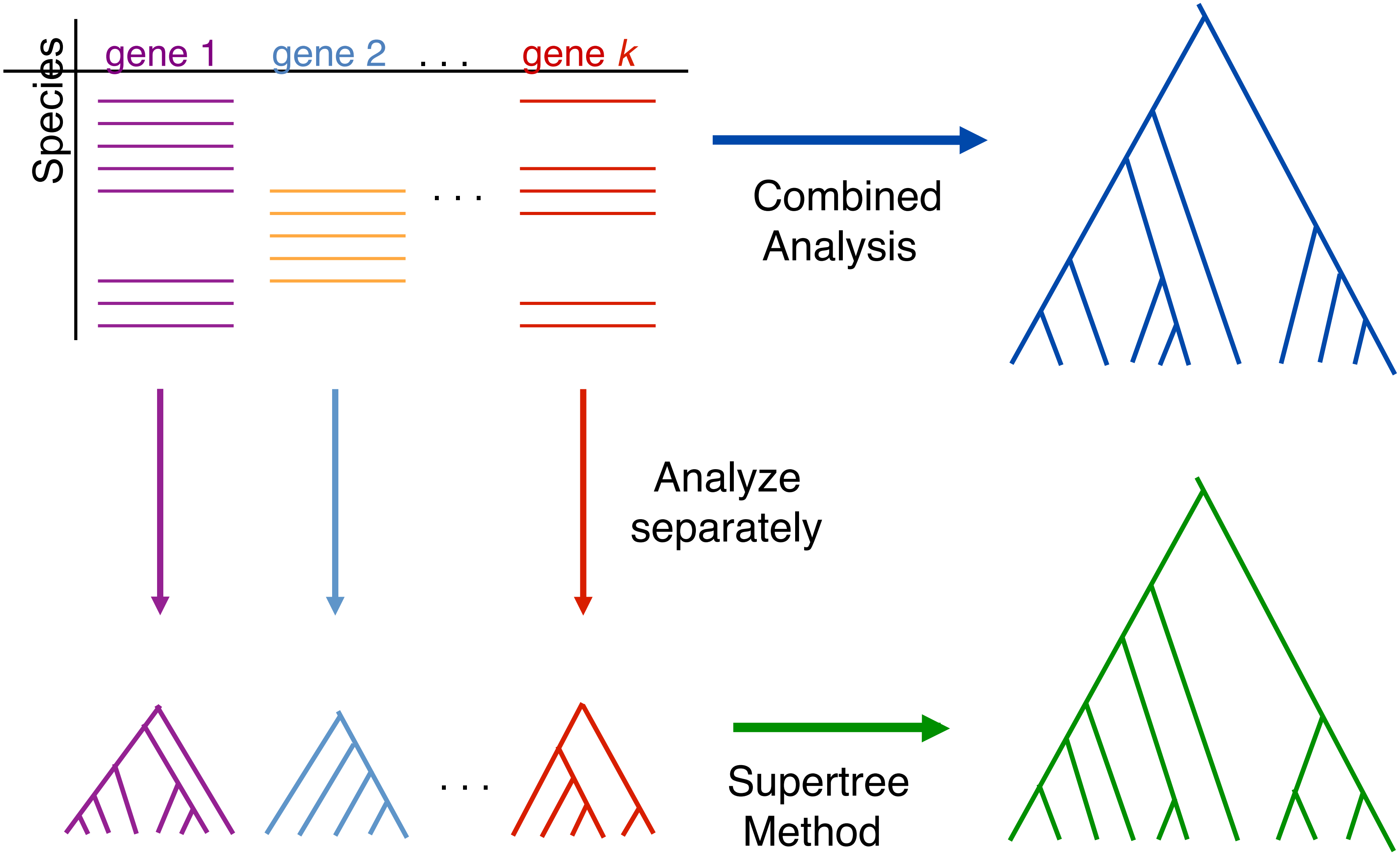
- Compute trees on individual genes and combine gene trees

Combined analysis

	gene 1	gene 2	gene 3
S_1	TCTAATGGAA	??????????	TATTGATACA
S_2	GCTAAGGGAA	??????????	??????????
S_3	TCTAAGGGAA	??????????	TCTTGATACC
S_4	TCTAACGGAA	GGTAACCCTC	TAGTGATGCA
S_5	??????????	GCTAAACCTC	??????????
S_6	??????????	GGTGACCATC	??????????
S_7	TCTAATGGAC	GCTAAACCTC	TAGTGATGCA
S_8	TATAACGGAA	??????????	CATTCATACC



Two competing approaches



Species

gene 1

gene 2

. . .

gene *k*

. . .

Combined Analysis

Analyze separately

. . .

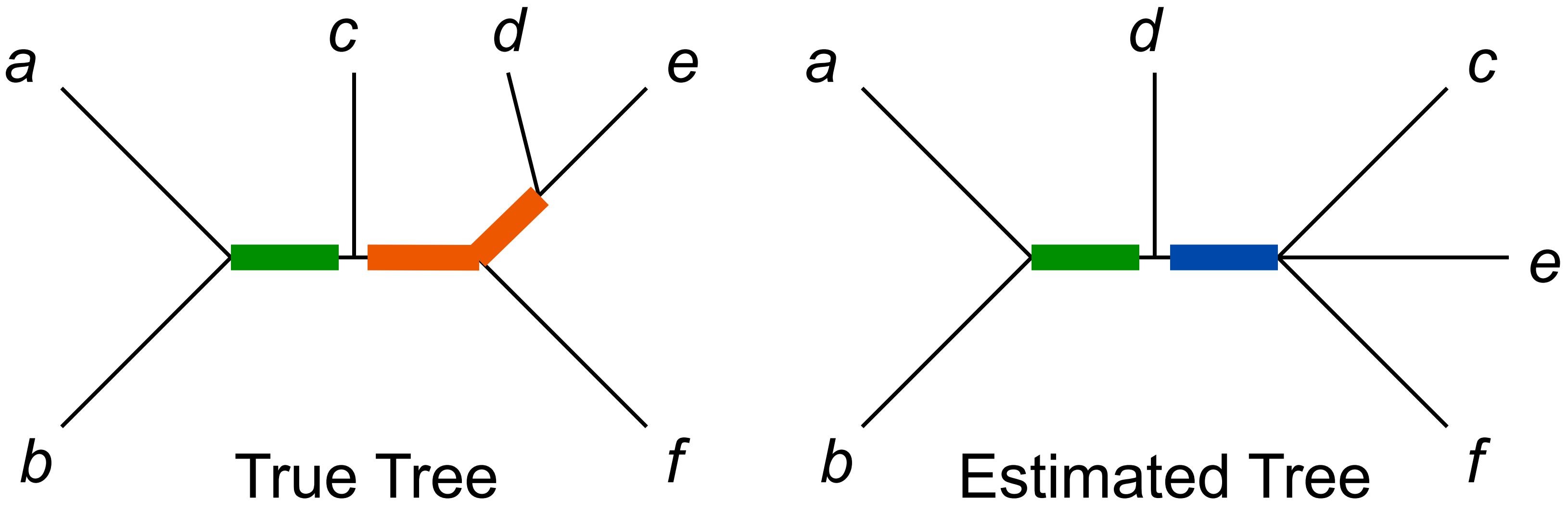
Supertree Method

Many Supertree Methods

Matrix Representation with Parsimony
(Most commonly used and most accurate)

- MRP
- weighted MRP
- MRF
- MRD
- Robinson-Foulds Supertrees
- Min-Cut
- Modified Min-Cut
- Semi-strict Supertree
- QMC
- Q-imputation
- SDM
- PhySIC
- Majority-Rule Supertrees
- Maximum Likelihood Supertrees
- and many more ...

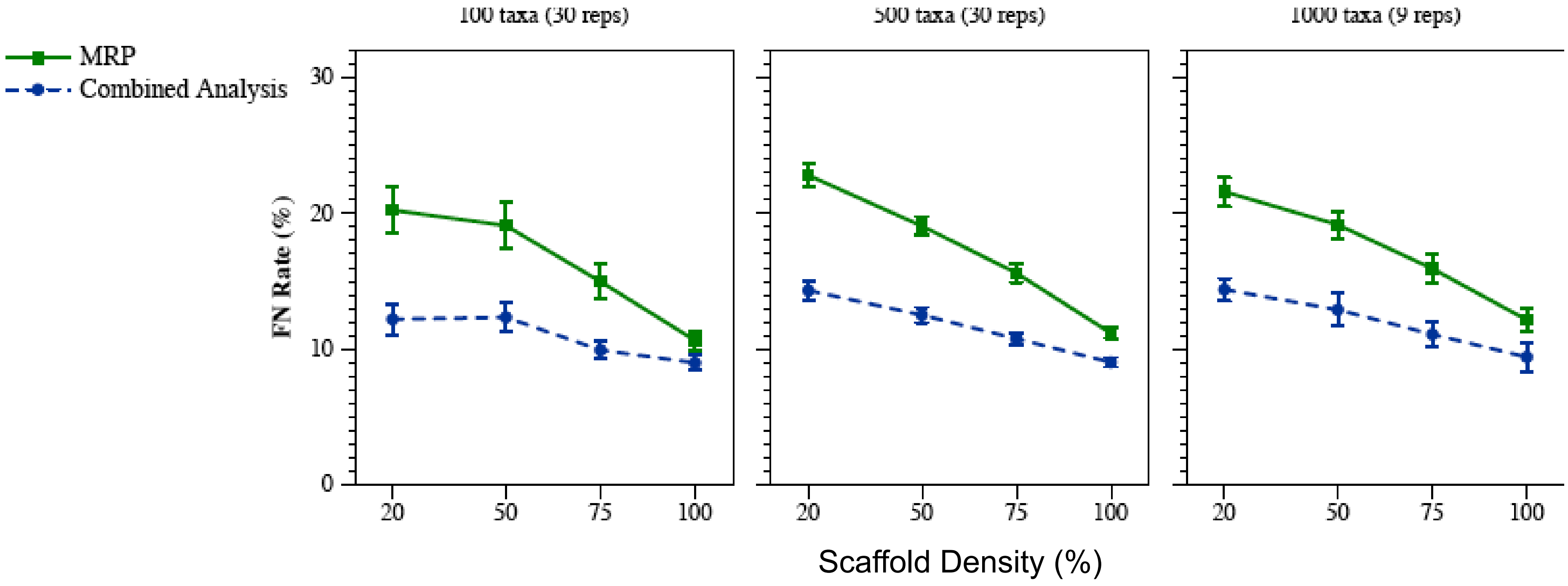
Quantifying topological error



- False negative (FN): $b \in B(T_{\mathrm{true}})\text{-}B(T_{\mathrm{est.}})$

- False positive (FP): $b \in B(T_{\mathrm{est.}})\text{-}B(T_{\mathrm{true}})$

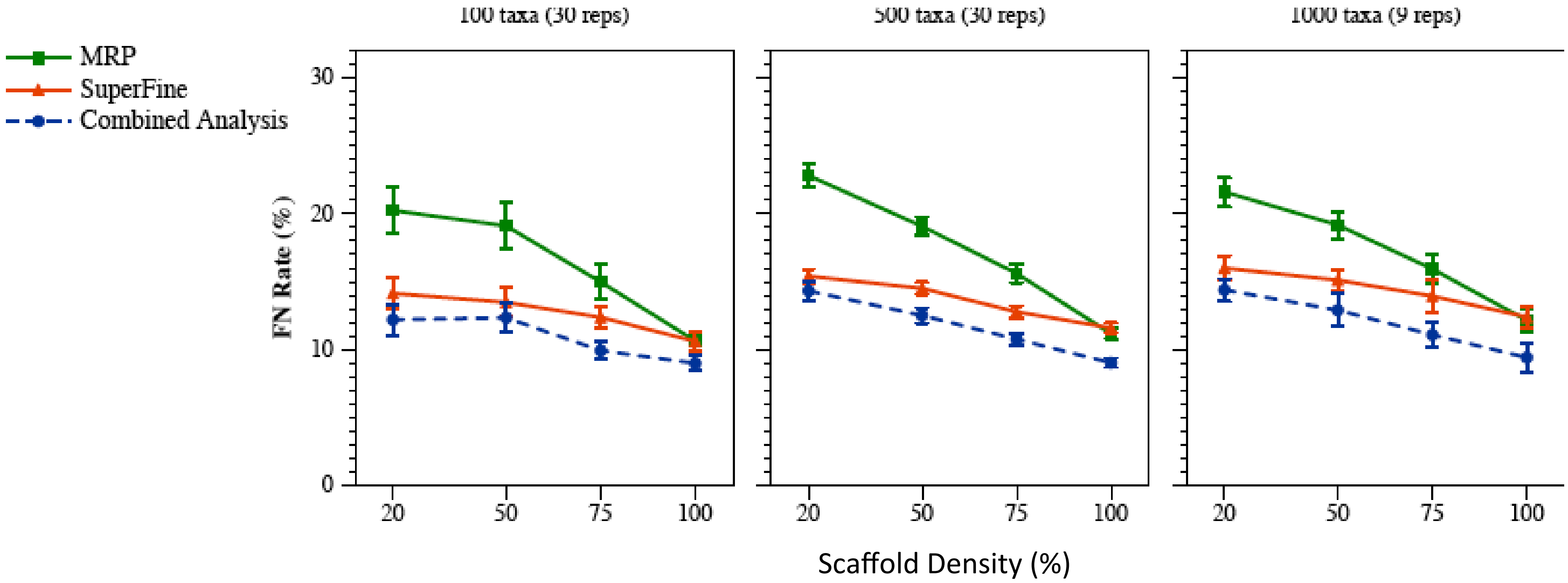
FN rate of MRP vs. combined analysis



SuperFine

- SuperFine: Fast and Accurate Supertree Estimation
- Systematic Biology 2012
- Authors: Shel Swenson, Rahul Suri, Randy Linder, and Tandy Warnow
- Software available at http://www.cs.utexas.edu/~phylo/software/superfine/

SuperFine-boosting: improves MRP



(Swenson et al., Syst. Biol. 2012)

Summary

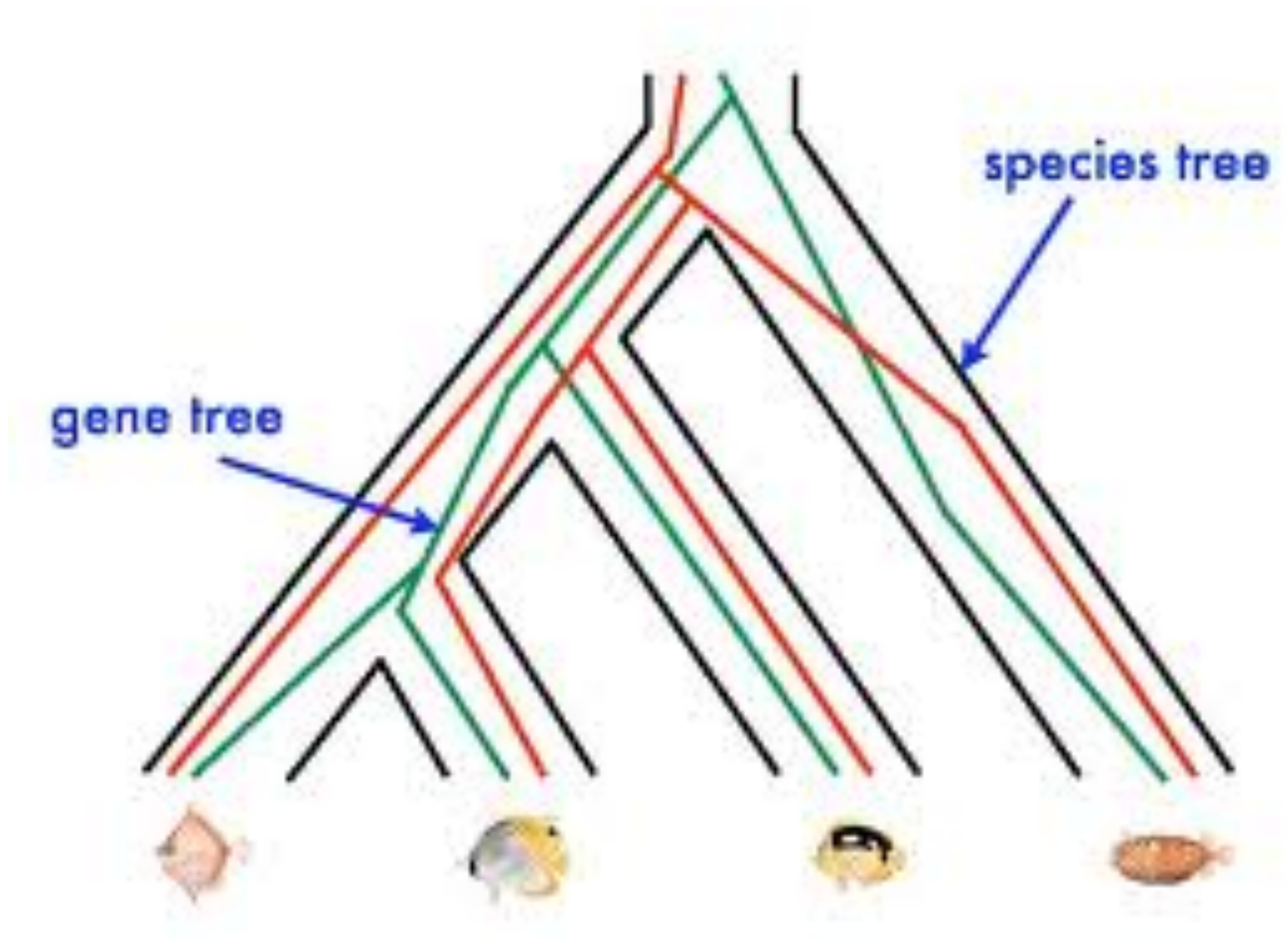
- Supertree methods approach the accuracy of concatenation (“combined analysis”)
- Supertree methods can be much faster than concatenation, especially for whole genome analyses (thousands of genes with millions of sites).
- But…

But…

- Gene trees may not be identical to species trees:
  - Incomplete Lineage Sorting (deep coalescence)
  - Gene duplication and loss
  - Horizontal gene transfer

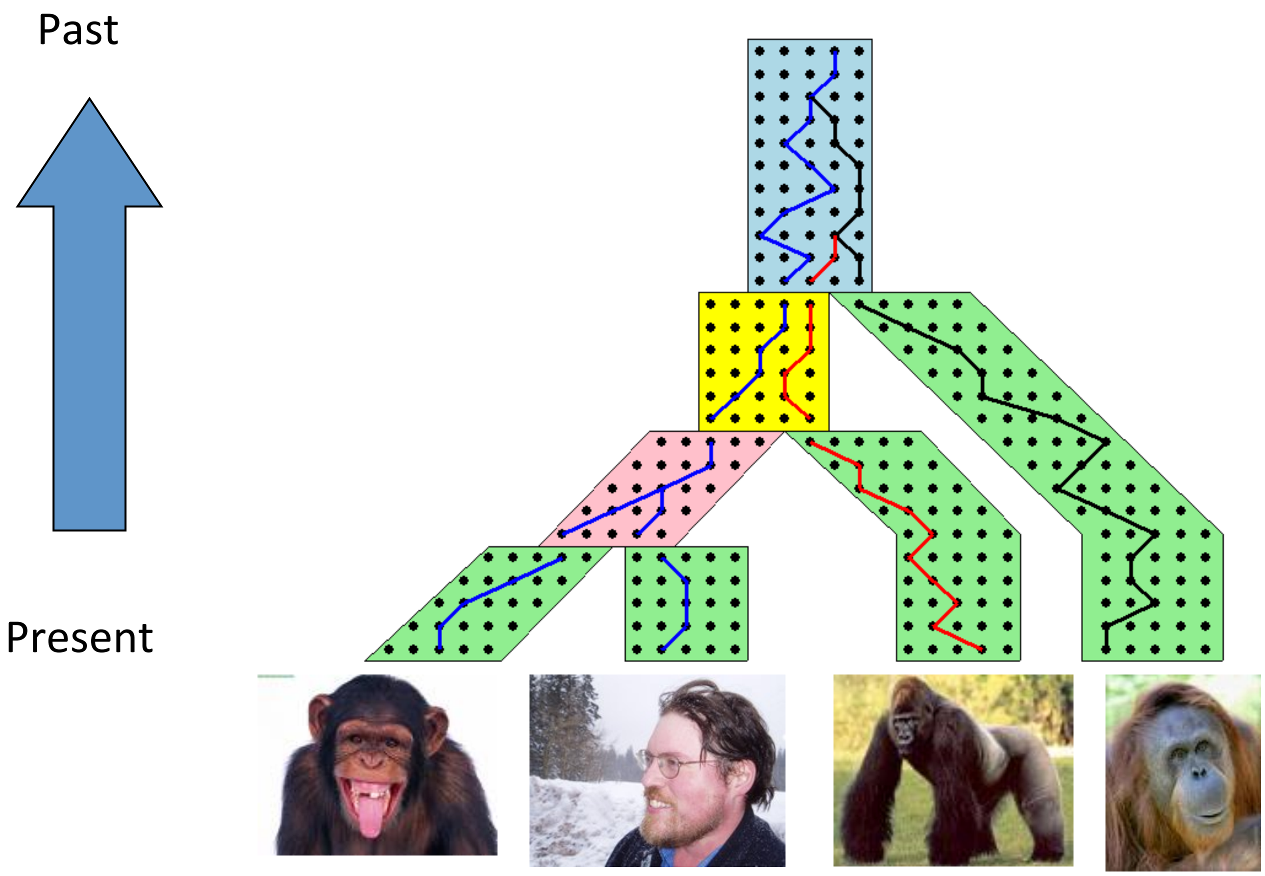
- This makes combined analysis and standard supertree analyses inappropriate

Red gene tree ≠ species tree (green gene tree okay)



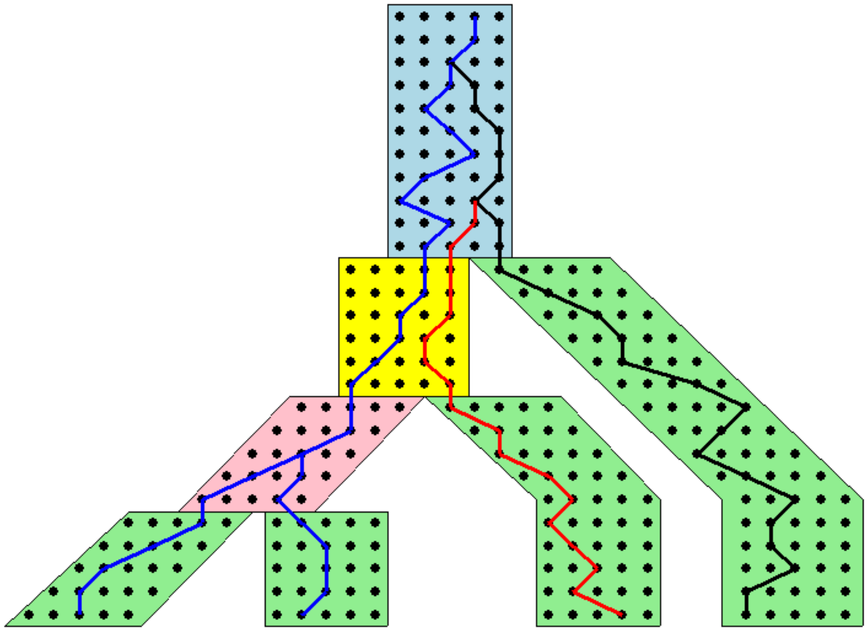
The Coalescent

Courtesy James Degnan



Gene tree in a species tree

Courtesy James Degnan



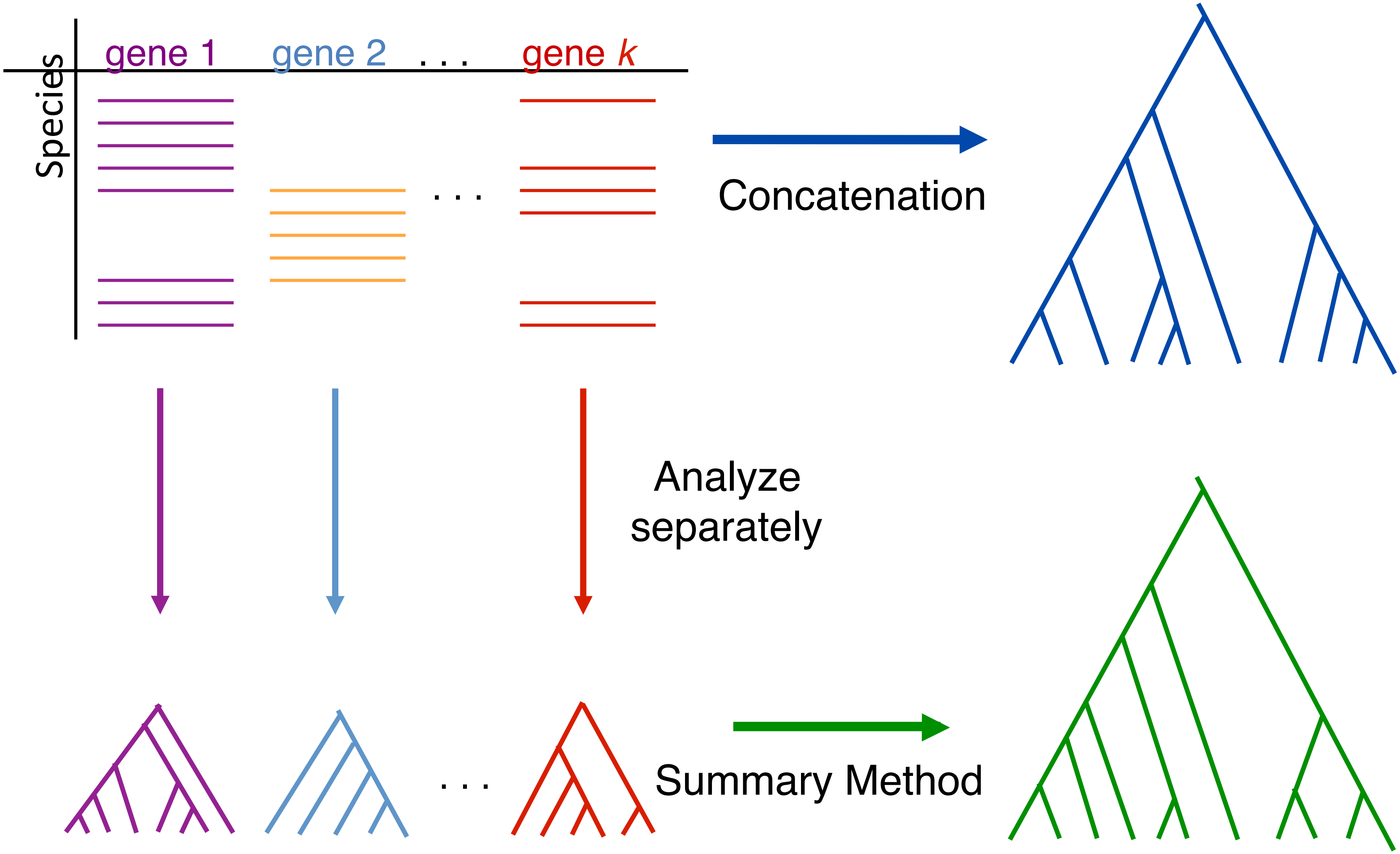
Deep coalescence

- Population-level process
- Gene trees can differ from species trees due to short times between speciation events

Incomplete Lineage Sorting (ILS)

- 2000+ papers in 2013 alone
- Confounds phylogenetic analysis for many groups:
  - Hominids
  - Birds
  - Yeast
  - Animals
  - Toads
  - Fish
  - Fungi
- There is substantial debate about how to analyze phylogenomic datasets in the presence of ILS.

Two competing approaches



Species

gene 1 gene 2 . . . gene *k*

. . .

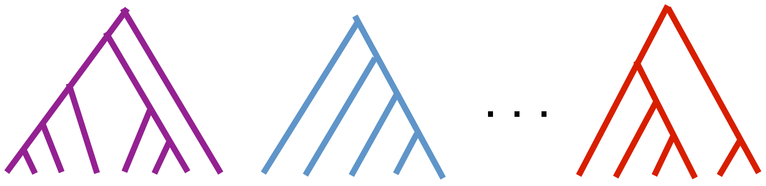
Concatenation

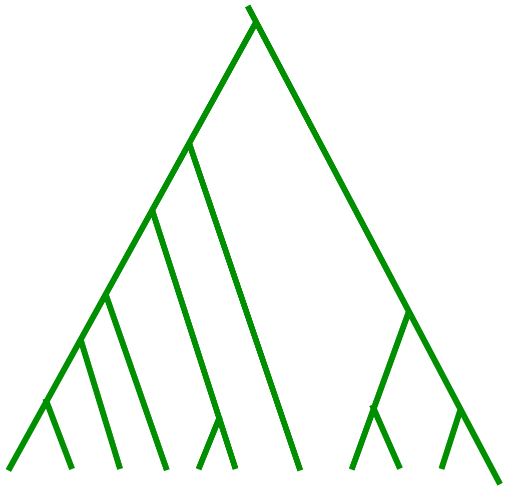
Analyze separately

. . .

Summary Method

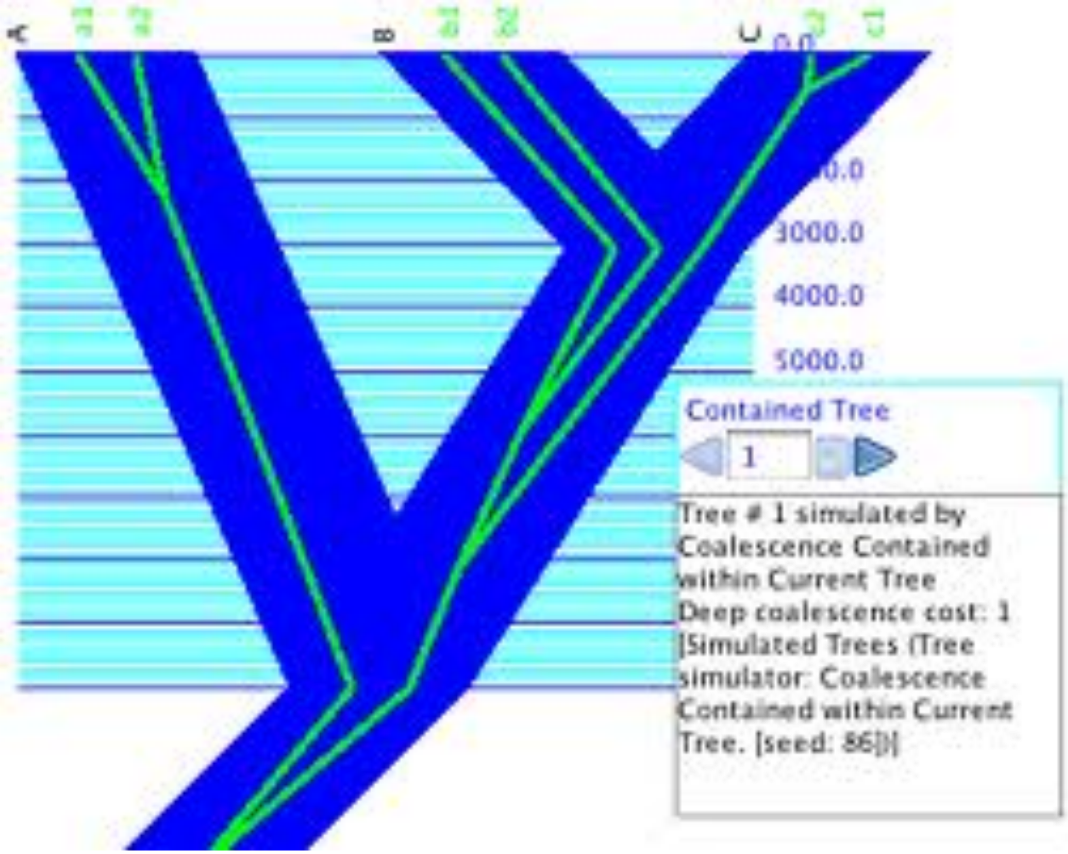
How to compute a species tree?



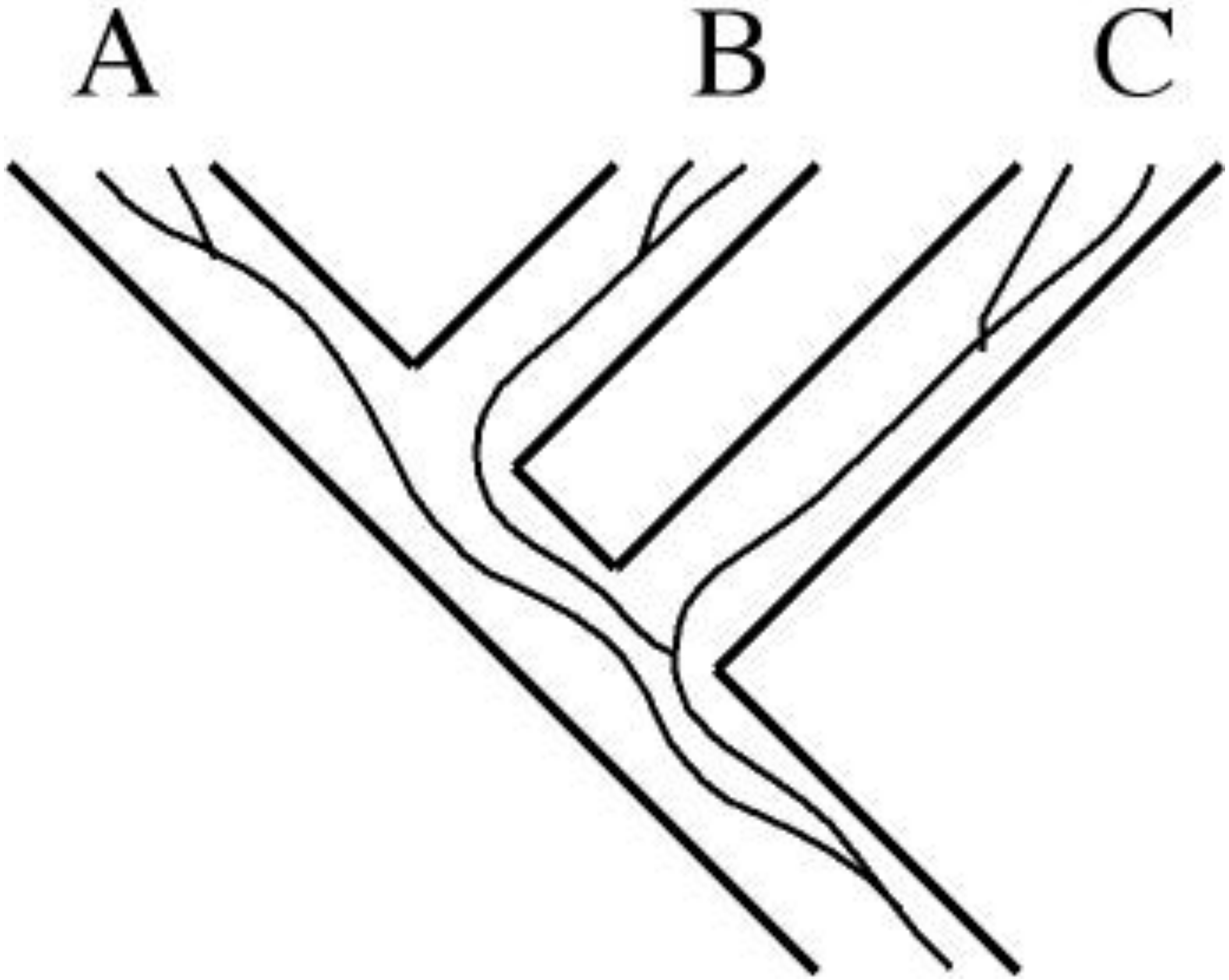


MDC: count # extra lineages

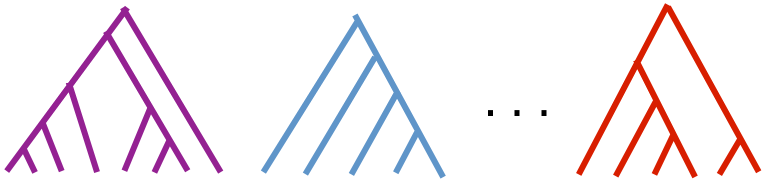
- Wayne Maddison proposed the MDC (minimize deep coalescence) problem: given set of true gene trees, find the species tree that implies the fewest deep coalescence events
- (Really amounts to counting the number of extra lineages)



Tree # 1 simulated by Coalescence Contained within Current Tree
Deep coalescence cost: 1
[Simulated Trees (Tree simulator: Coalescence Contained within Current Tree. [seed: 86])]

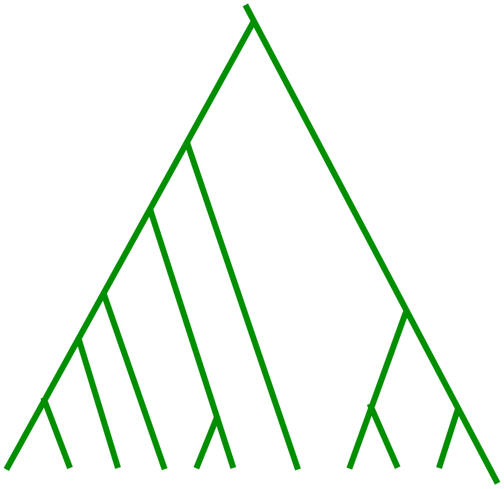

A
B
C

How to compute a species tree?



Techniques:

- MDC?
- Most frequent gene tree?
- Consensus of gene trees?
- Other?



Statistically consistent under ILS?

- MP-EST (Liu et al. 2010): maximum likelihood estimation of rooted species tree – YES

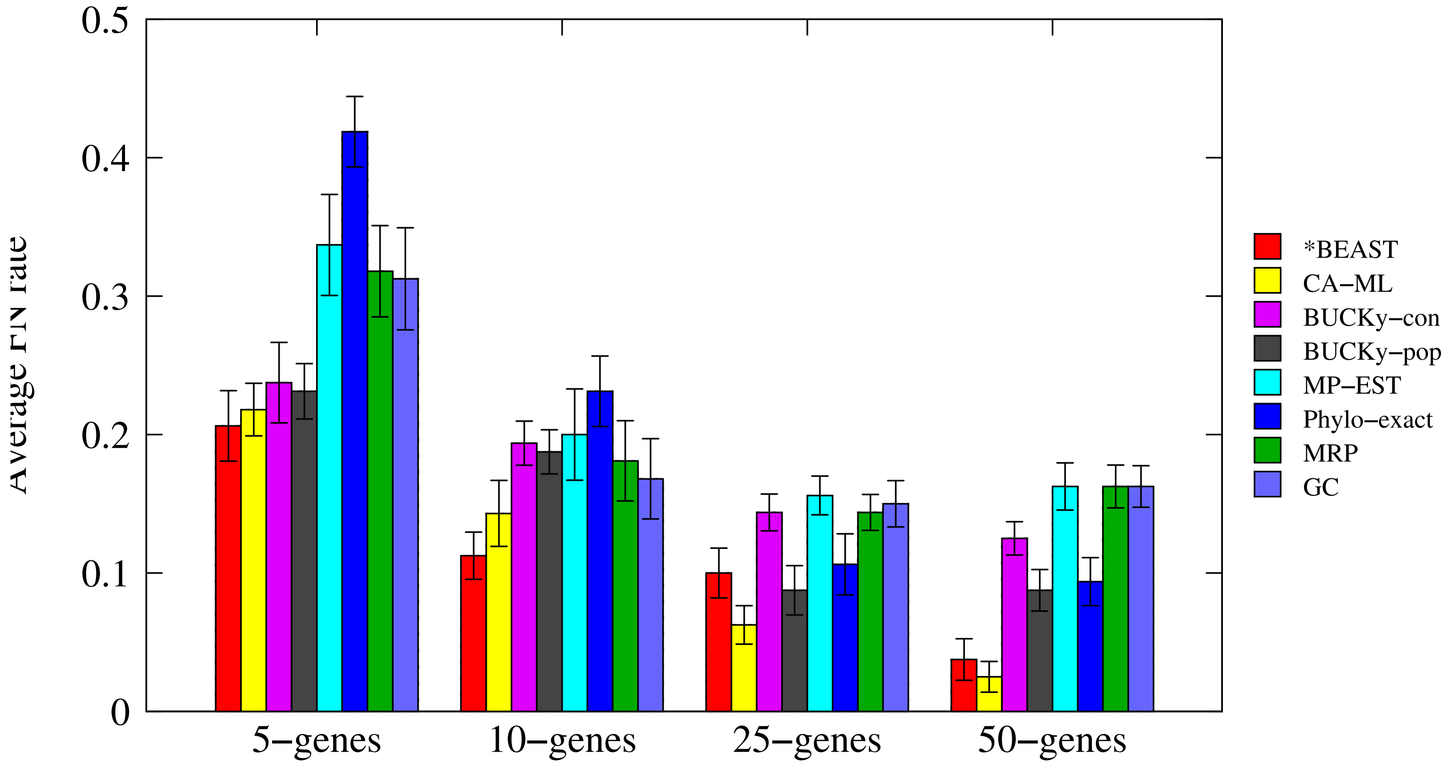
- BUCKy-pop (Ané and Larget 2010): quartet-based Bayesian species tree estimation –YES
- MDC – NO
- Greedy – NO
- Concatenation under maximum likelihood – open
- MRP (supertree method) – open

The Debate: Concatenation vs. Coalescent Estimation

- In favor of coalescent-based estimation
  - Statistical consistency guarantees
  - Addresses gene tree incongruence resulting from ILS
  - Some evidence that concatenation can be positively misleading

- In favor of concatenation
  - Reasonable results on data
  - High bootstrap support
  - Summary methods (that combine gene trees) can have poor support or miss well-established clades entirely
  - Some methods (such as *BEAST) are computationally too intensive to use

Results on 11-taxon datasets with strongILS



***BEAST** more accurate than summary methods (MP-EST, BUCKy, etc)

CA-ML: (concatenated analysis) also very accurate

Datasets from Chung and Ané, 2011
Bayzid & Warnow, Bioinformatics 2013

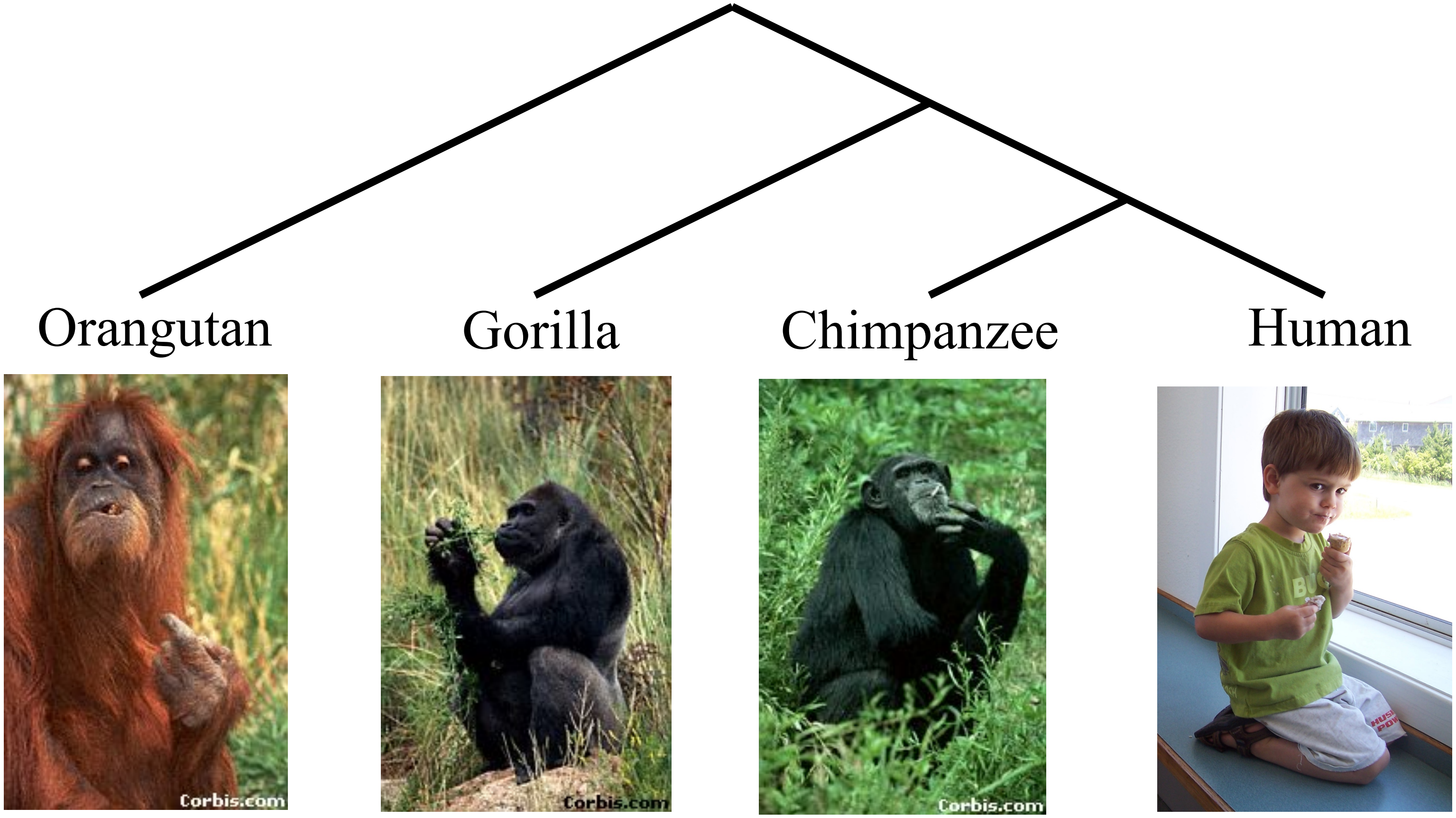
Is Concatenation Evil?

- Joseph Heled:
  - YES

- John Gatesy
  - No

- Data needed to held understand existing methods and their limitations
- Better methods are needed

Species tree estimation: difficult, even for small datasets



Orangutan

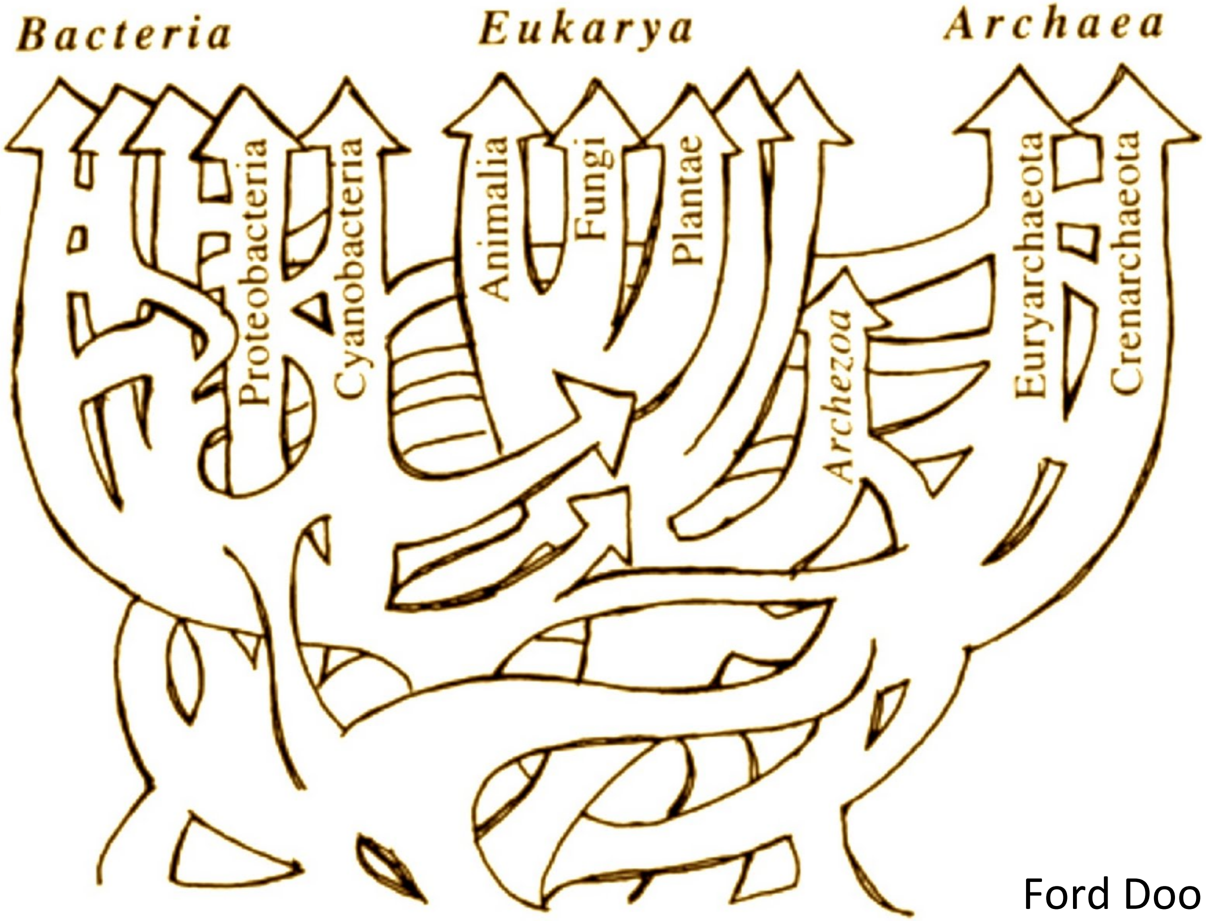
Gorilla

Chimpanzee

Human

From the Tree of the Life Website,
University of Arizona

Horizontal Gene Transfer – Phylogenetic Networks



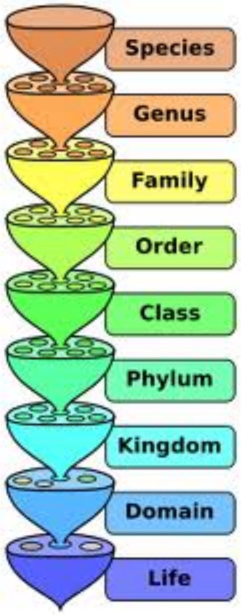
Ford Doolittle

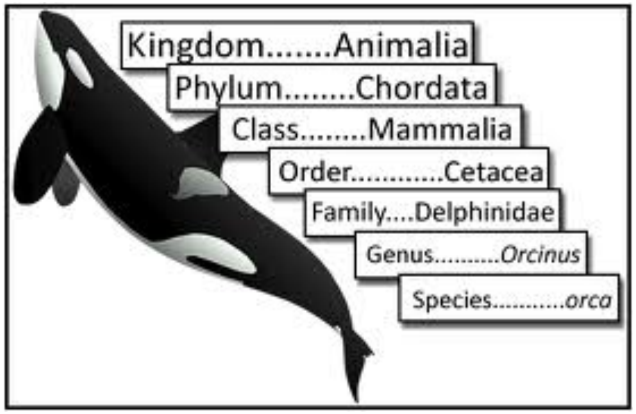
Species tree/network estimation

- Methods have been developed to estimate species phylogenies (trees or networks!) from gene trees, when gene trees can conflict from each other (e.g., due to ILS, gene duplication and loss, and horizontal gene transfer).
- Phylonet (software suite), has effective methods for many optimization problems – including MDC and maximum likelihood.
- Tutorial on Wednesday.
- Software available at http://bioinfo.cs.rice.edu/phylonet?destination=node/3

Metagenomic Taxon Identification

Objective: classify short reads in a metagenomic sample





Two Basic Questions

1. What is this fragment? (Classify each fragment as well as possible.)

2. What is the taxonomic distribution in the dataset? (Note: helpful to use marker genes.)

SEPP

- **SEPP: SATé-enabled Phylogenetic Placement**, by Mirarab, Nguyen, and Warnow

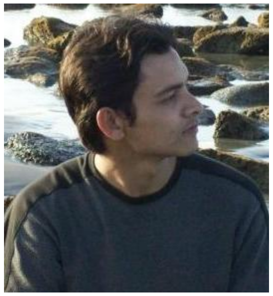
- Pacific Symposium on Biocomputing, 2012 (special session on the Human Microbiome)

- Tutorial on Thursday.

Other problems

- Genomic MSA estimation:
  - Multiple sequence alignment of very long sequences
  - Multiple sequence alignment of sequences that evolve with rearrangement events
- Phylogeny estimation under more complex models
  - Heterotachy
  - Violation of the rates-across-sites assumption
  - Rearrangements
- Estimating branch support on very large datasets

Warnow Laboratory



PhD students: Siavash Mirarab*, Nam Nguyen, and Md. S. Bayzid**
Undergrad: Keerthana Kumar
Lab Website: http://www.cs.utexas.edu/users/phylo

Funding: Guggenheim Foundation, Packard, NSF, Microsoft Research New England, David Bruton Jr. Centennial Professorship, TACC (Texas Advanced Computing Center), and the University of Alberta (Canada)

TACC and UTCS computational resources

* Supported by HHMI Predoctoral Fellowship
** Supported by Fulbright Foundation Predoctoral Fellowship