Complexity and The Tree of Life

Tandy Warnow
The University of Texas at Austin

CIPRES

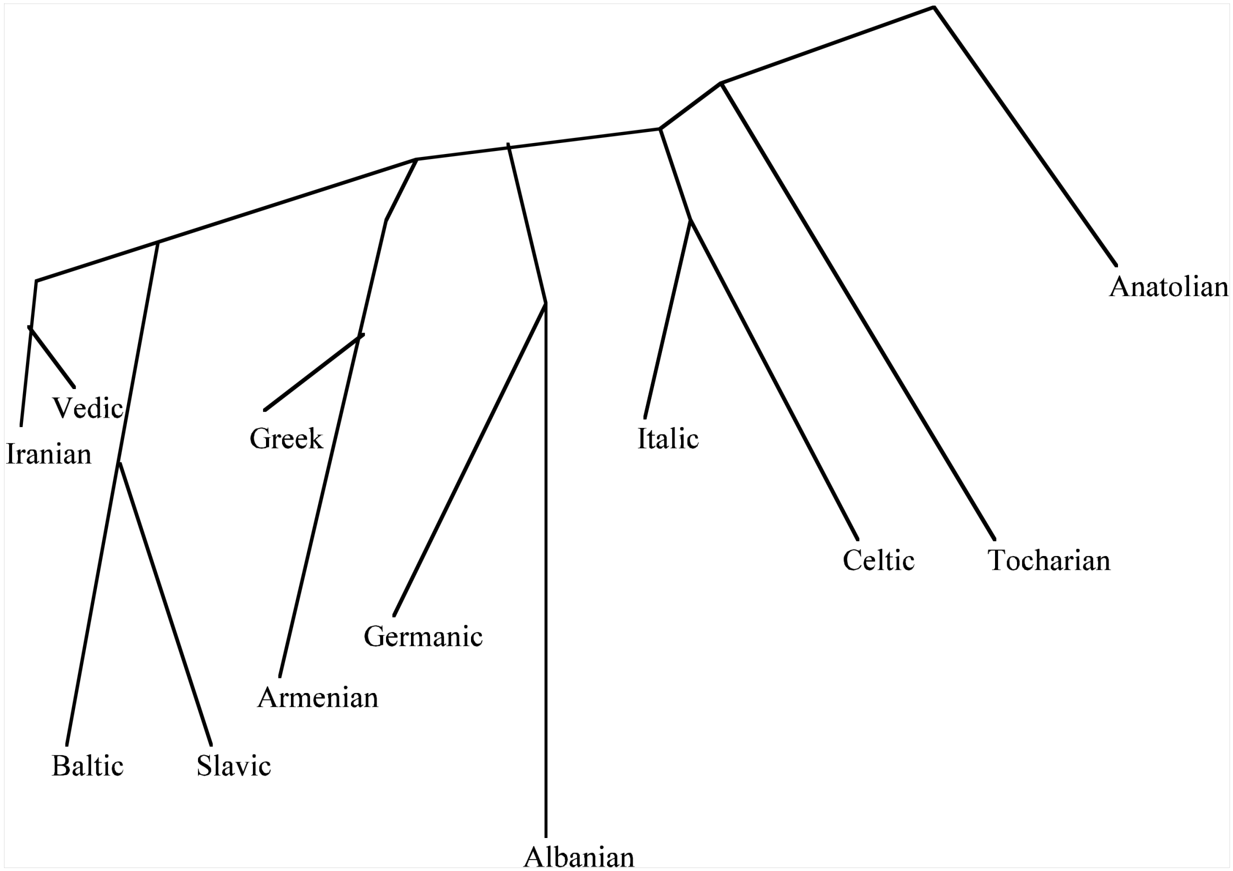
How did life evolve on earth?

- Courtesy of the Tree of Life project

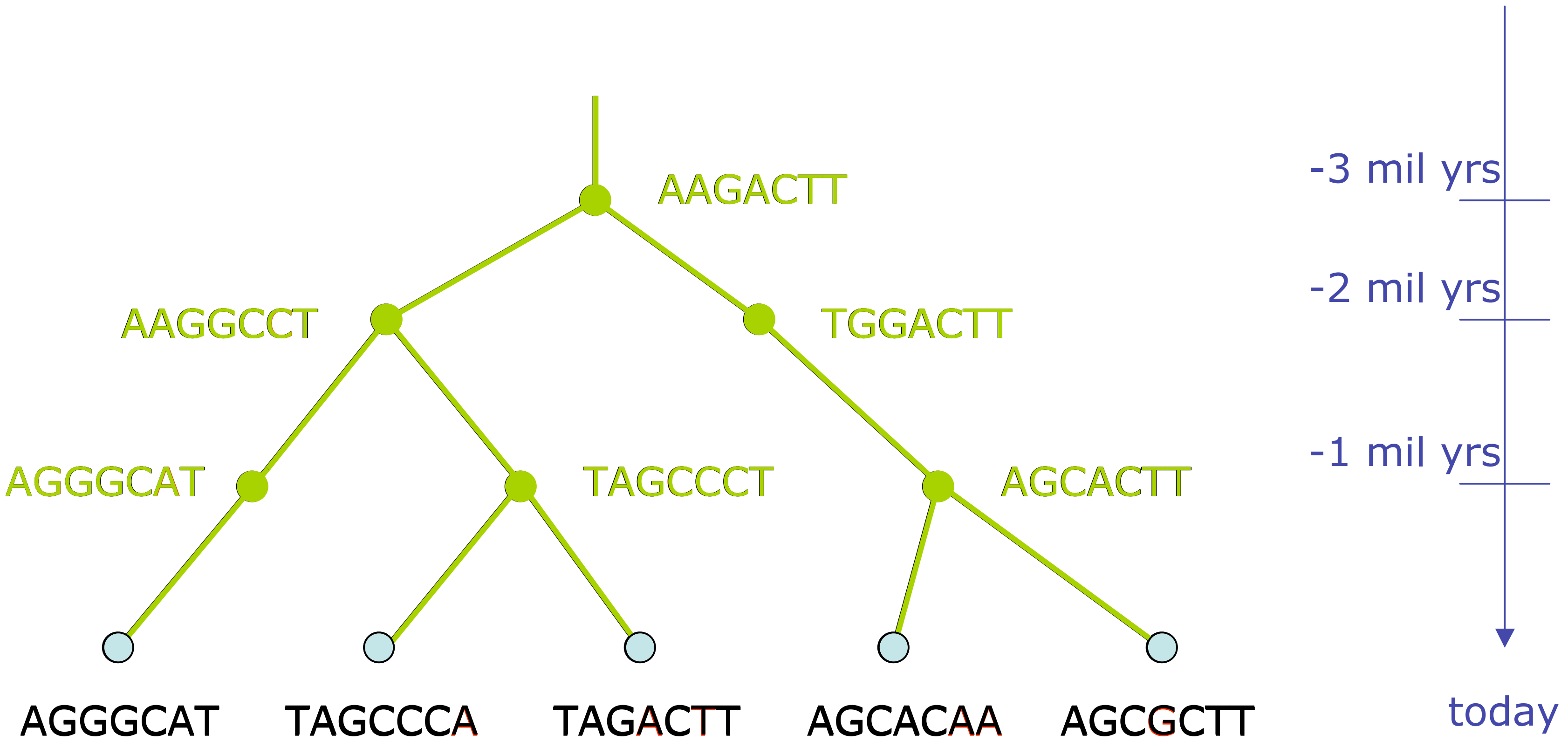
An international effort to understand how life evolved on earth

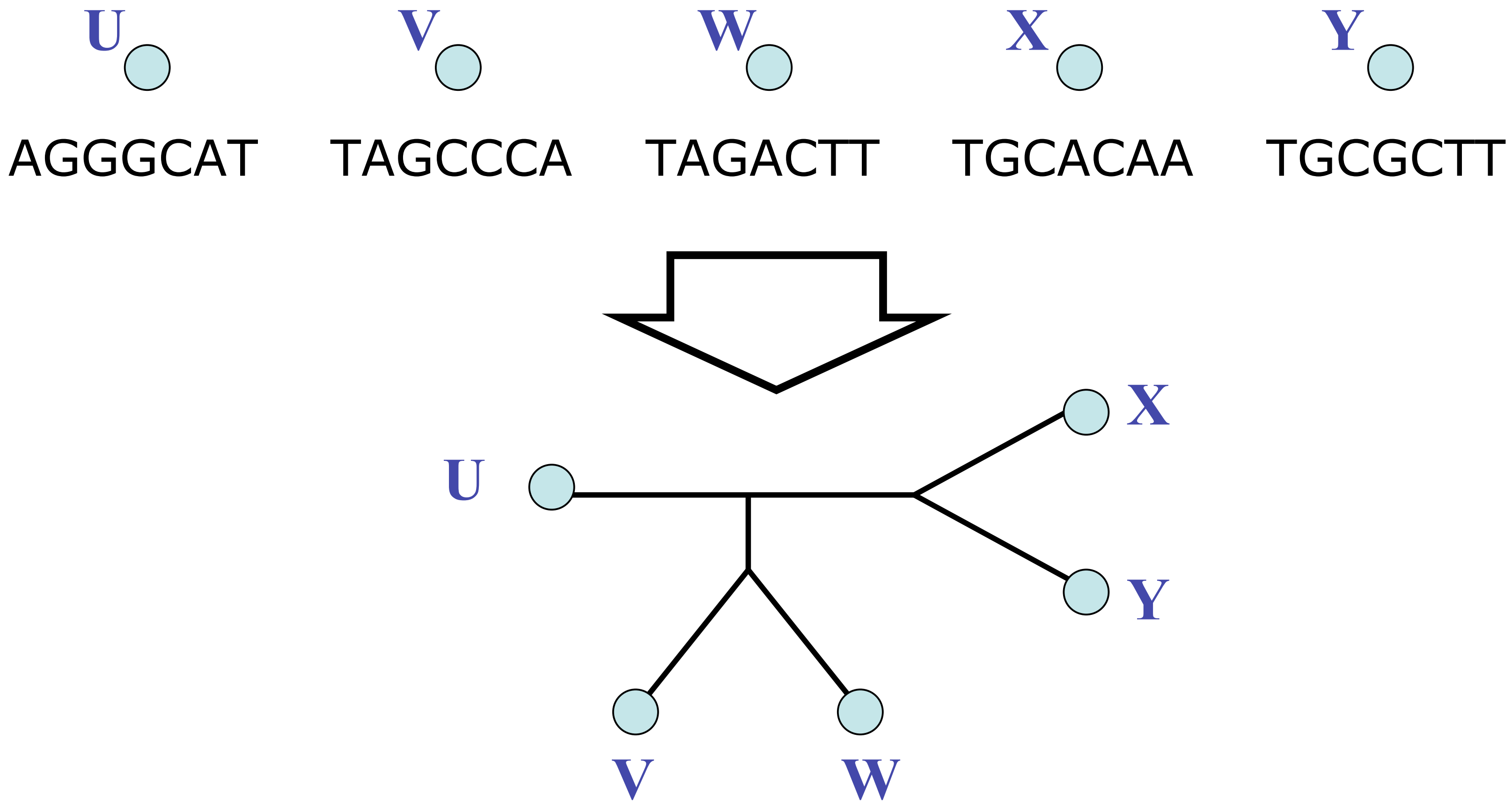
Biomedical applications: drug design, protein structure and function prediction, biodiversity.

How did human languages evolve? (Possible Indo-European tree, Ringe, Warnow and Taylor 2000)



DNA Sequence Evolution





U	V	W	X	Y
AGGGCAT	TAGCCCA	TAGACTT	TGCACAA	TGCGCTT

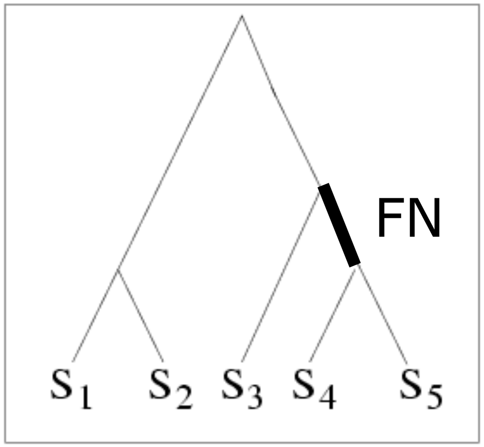
Standard Markov models

- Sequences evolve just with substitutions
- Sites (i.e., positions) evolve identically and independently, and have “rates of evolution” that are drawn from a common distribution (typically gamma)
- Numerical parameters describe the probability of substitutions of each type on each edge of the tree

Questions

- *Statistical consistency*: Is the given phylogeny reconstruction method guaranteed to reconstruct the model tree when infinitely long sequences are available?
- *Convergence rate* (sample size complexity): How long do the sequences need to be for the method to be accurate with high probability?

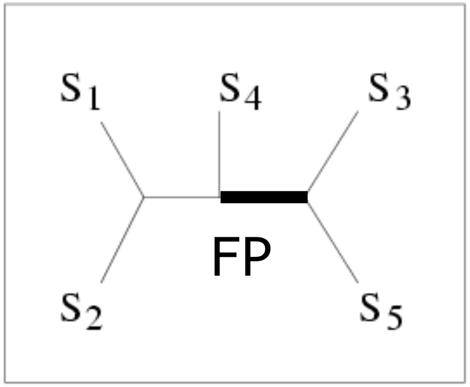
Quantifying Error



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

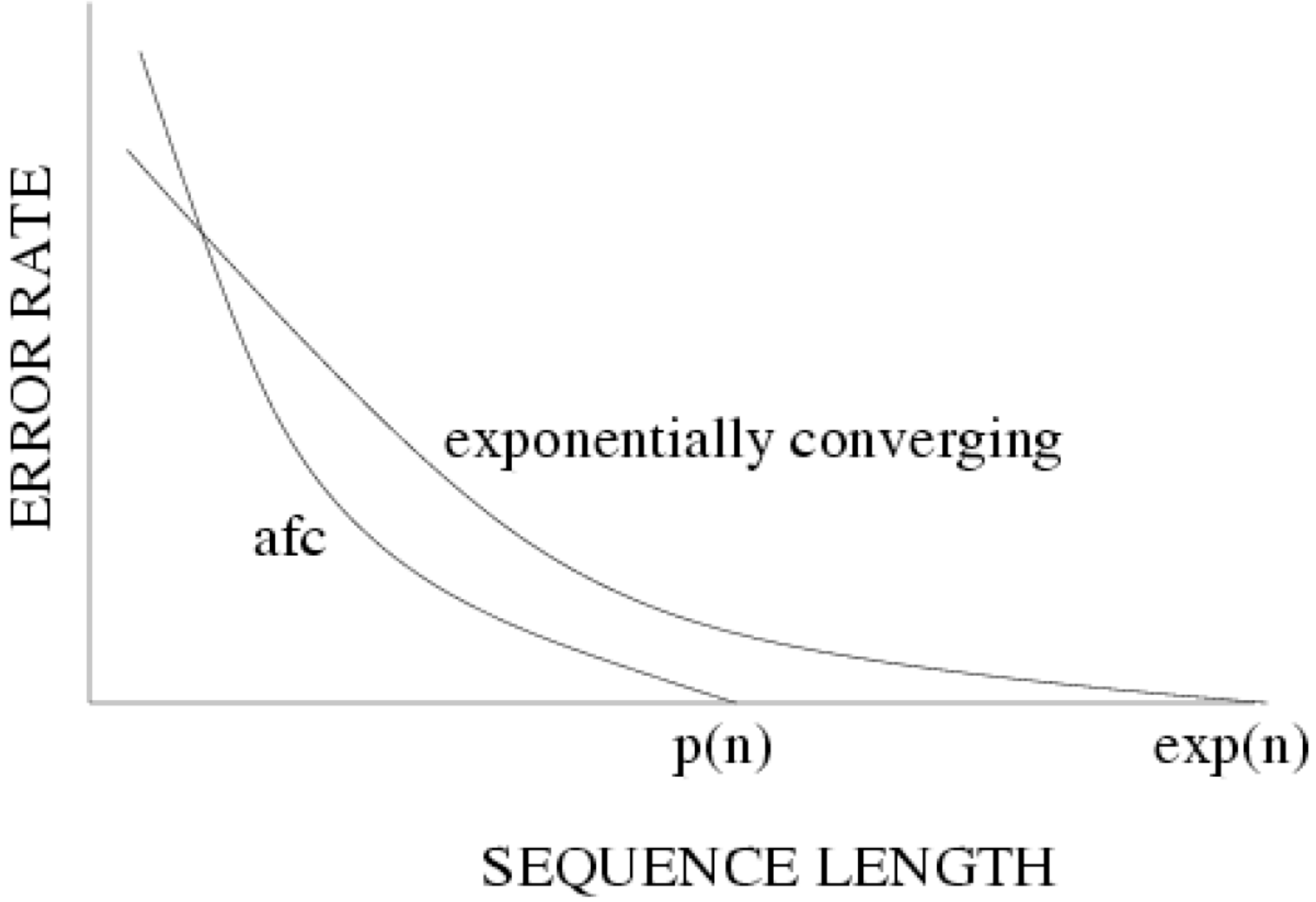


INFERRED TREE

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

50% error rate

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



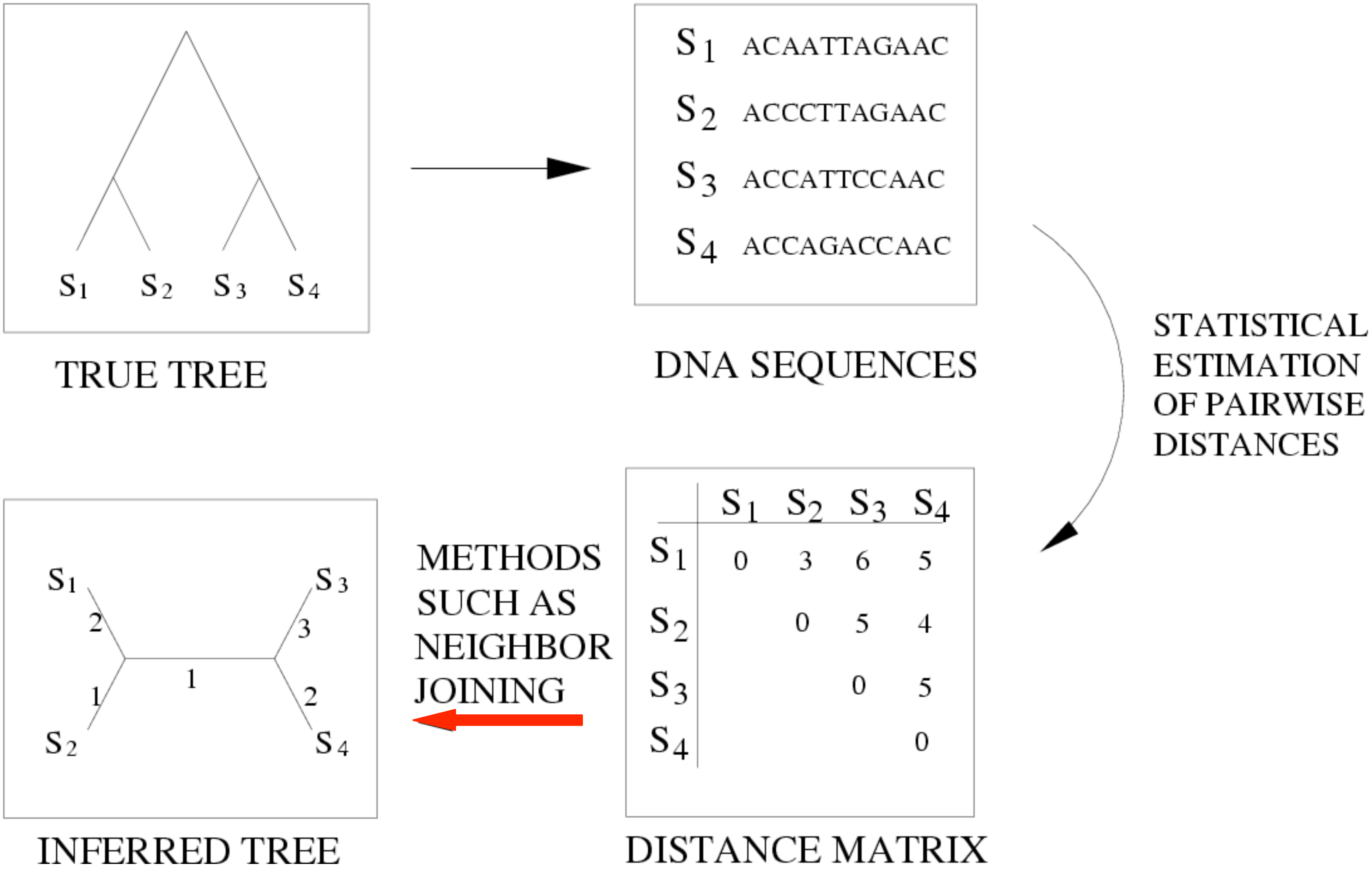
Complexity *viz.* The Tree of Life

- **Algorithmic complexity** (e.g., running time and NP-hardness)
- **Sample size complexity** (e.g. how long do the sequences need to be to obtain a highly accurate reconstruction with high probability?)
- **Stochastic model complexity** (i.e., how realistic are the models of evolution, and what are the consequences of making the models more realistic?)

Current state of knowledge

- We have established much of the statistical performance (consistency and convergence rates) of the major methods for phylogeny estimation.
- We have developed "fast converging" methods (guaranteed to reconstruct the true tree from polynomial length sequences) with excellent performance in practice.
- We have very fast methods for solving maximum likelihood and maximum parsimony, the major optimization problems, even for large datasets.

Distance-based Phylogenetic Methods (polynomial time)



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

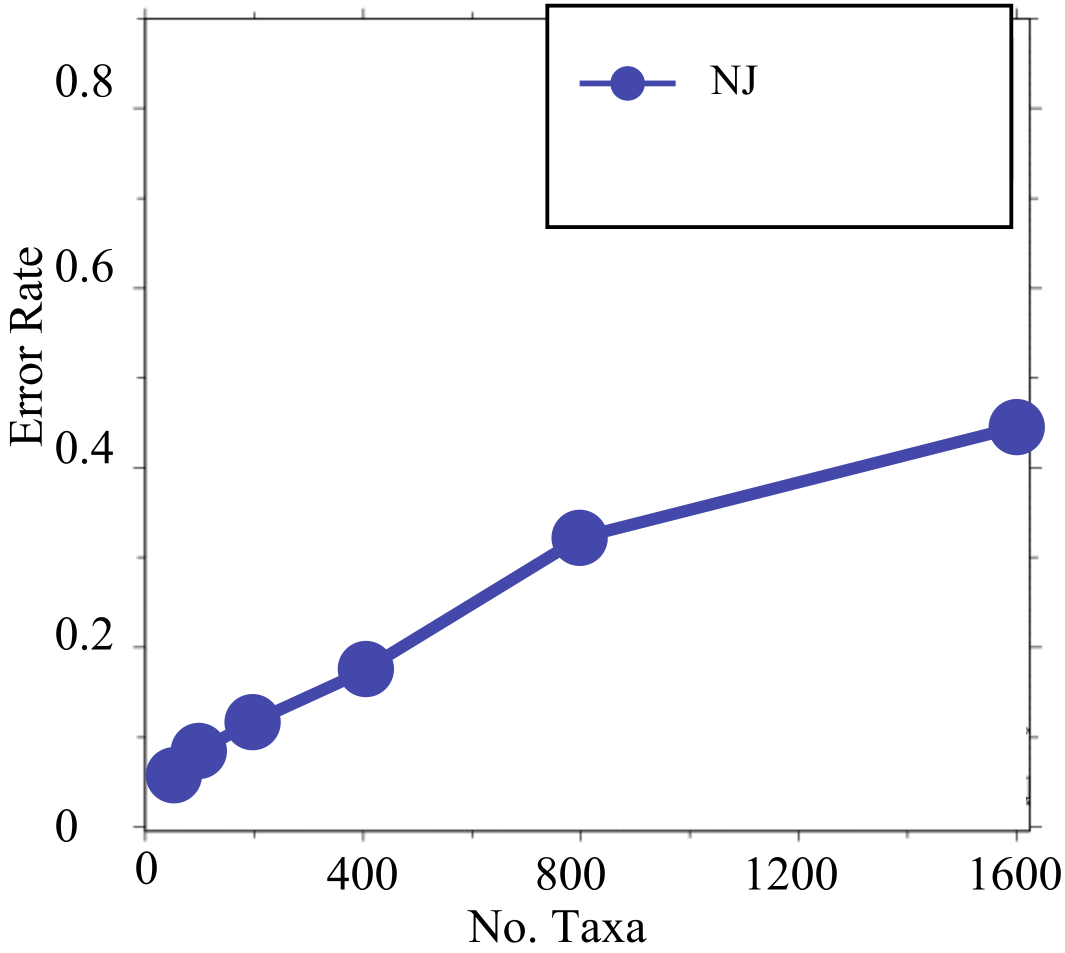
METHODS SUCH AS NEIGHBOR JOINING

INFERRED TREE

Neighbor Joining's sequence length requirement is exponential!

- Atteson: Let T be a General Markov model tree defining distance matrix D. Then Neighbor Joining will reconstruct the true tree with high probability from sequences that are of length at least **$O(\lg n\ e^{\max D_{ij}})$,** where n is the number of leaves in T.

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

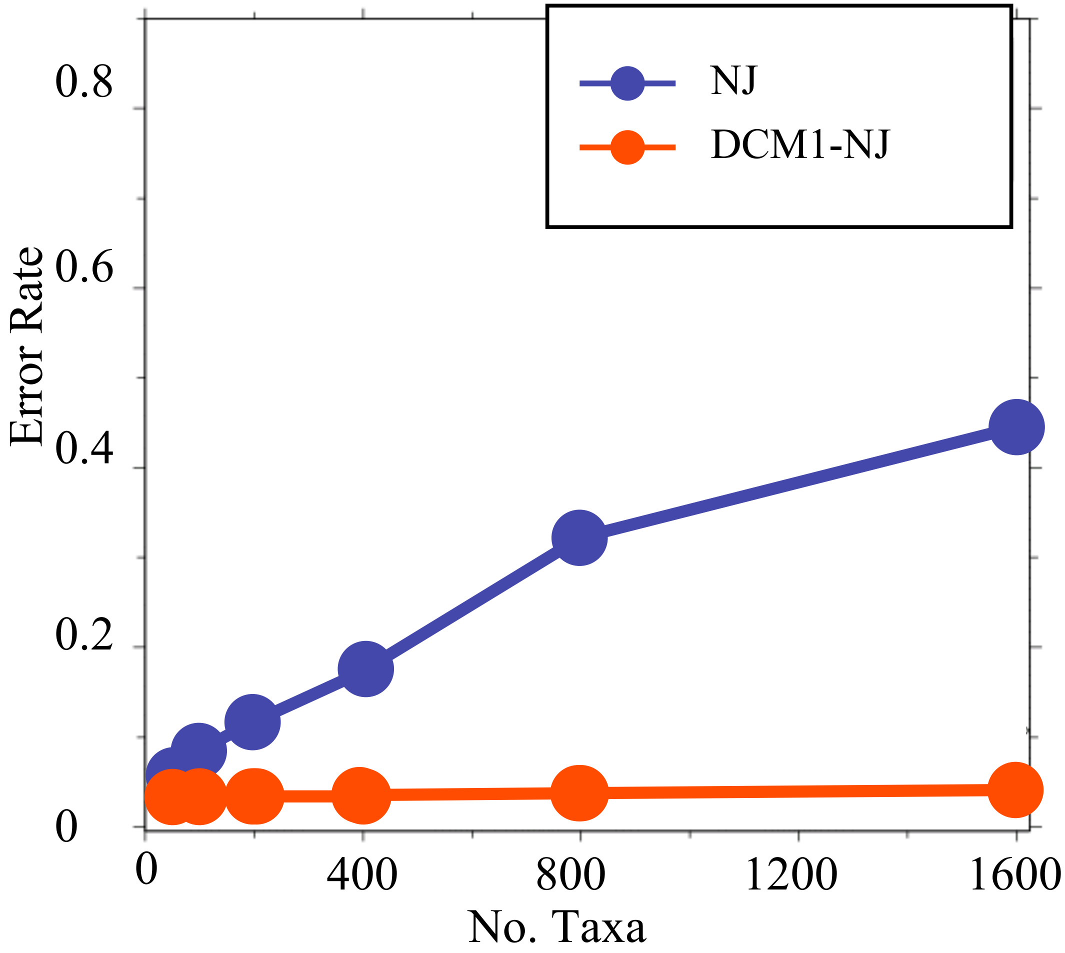


Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

Error rates reflect proportion of incorrect edges in inferred trees.

DCM1-boosting distance-based methods

[Nakhleh et al. ISMB 2001]



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

Other “fast-converging” methods

- The “short quartet” methods (Erdös, Steel, Székéley and Warnow 1997) were the first fast-converging methods, published in RSA 1999 and TCS 1999.
- Csüros and Kao (SODA 1999)
- Cryan, Goldberg, and Goldberg (SICOMP 2001)
- Csüros (J Comp Bio 2002)
- Daskalakis et al. (RECOMB 2006)
- Daskalakis, Mossel and Roch (STOC 2006)
- Gronau, Moran and Snir (SODA 2008)

Maximum Likelihood (ML)

- Given: Set S of aligned DNA sequences, and a parametric model of sequence evolution
- Objective: Find tree T and numerical parameter values (e.g, substitution probabilities) so as to maximize the probability of the data.

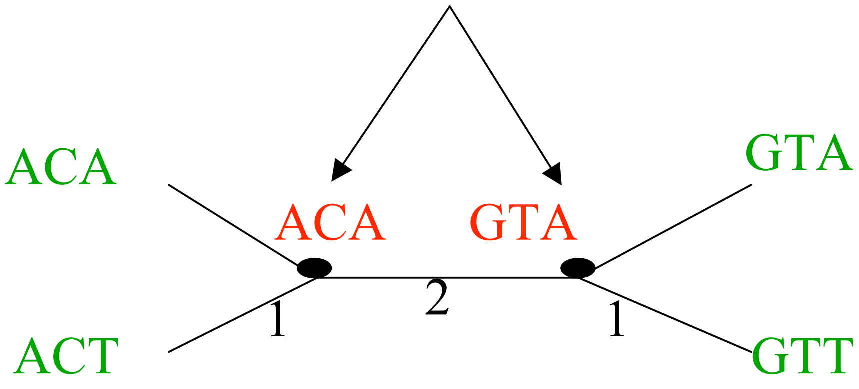
NP-hard

Statistically consistent for standard models if solved exactly

Maximum Parsimony
(Hamming distance Steiner Tree problem)

Input: set of aligned sequences

Output: tree with minimum total length (“MP score”)



MP score = 4

Not statistically consistent (even under simple models)

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

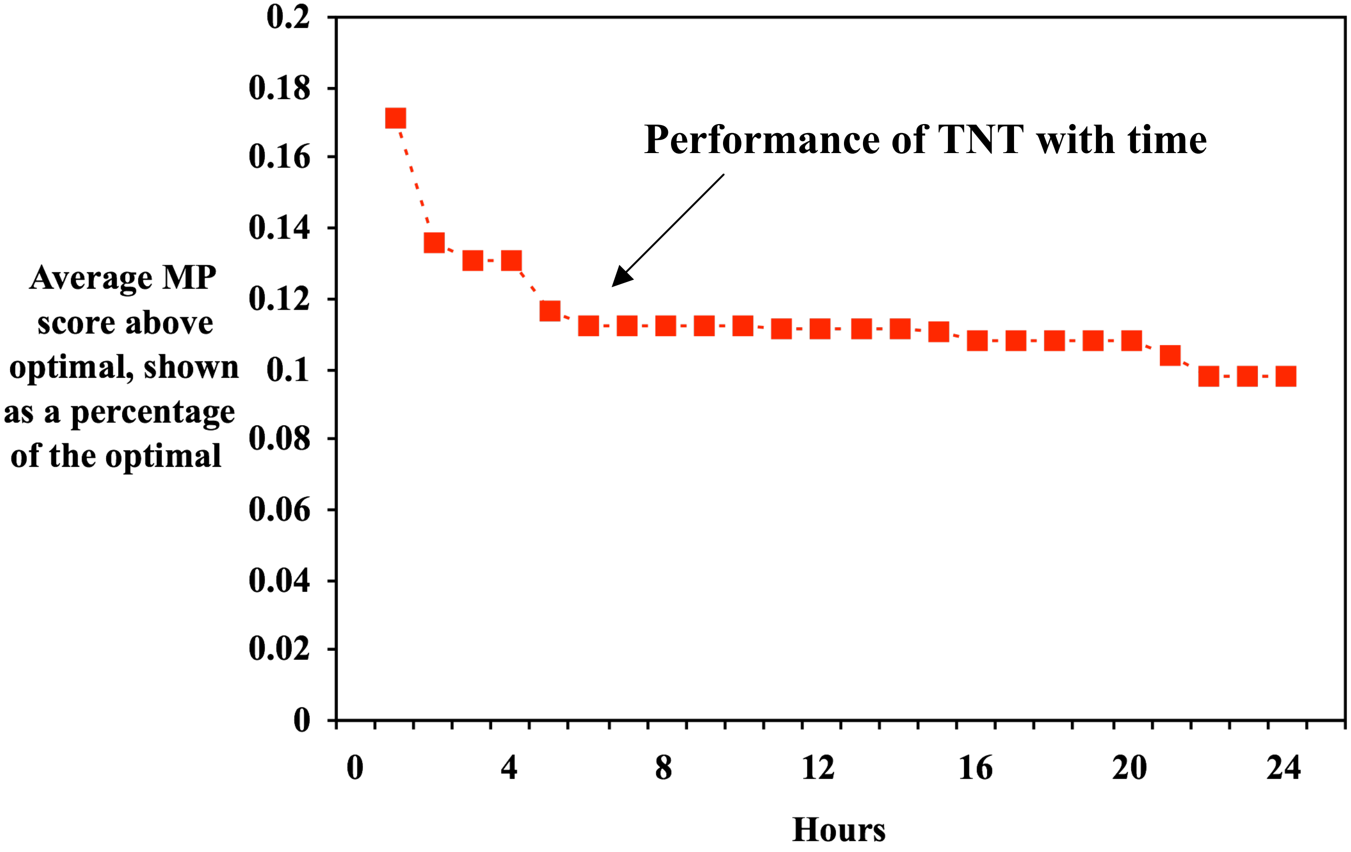
Solving NP-hard problems exactly is … unlikely

- Number of (unrooted) binary trees on *n* leaves is (2n-5)!!
- If each tree on **1000** taxa could be analyzed in **0.001** seconds, we would find the best tree in **2890 millennia**

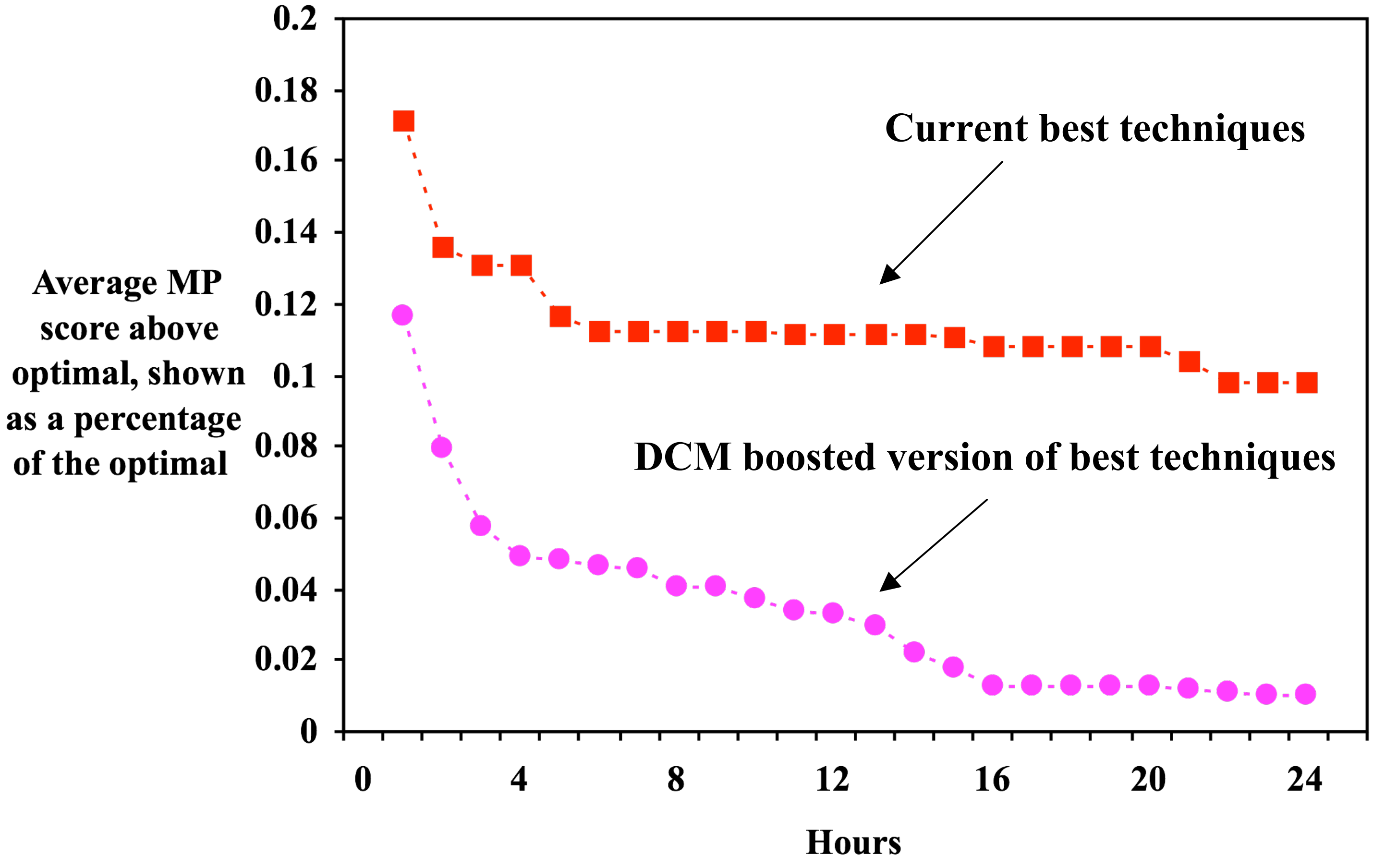
#leaves	#trees
4	3
5	15
6	105
7	945
8	10395
9	135135
10	2027025
20	2.2×10^{20}
100	4.5×10^{190}
1000	2.7×10^{2900}

Problems with techniques for Maximum Parsimony

Shown here is the performance of a very good heuristic (TNT) for 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 significantly improves performance (Roshan et al. CSB 2004)



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

Similar improvements obtained for RAxML (maximum likelihood).

Current state of knowledge

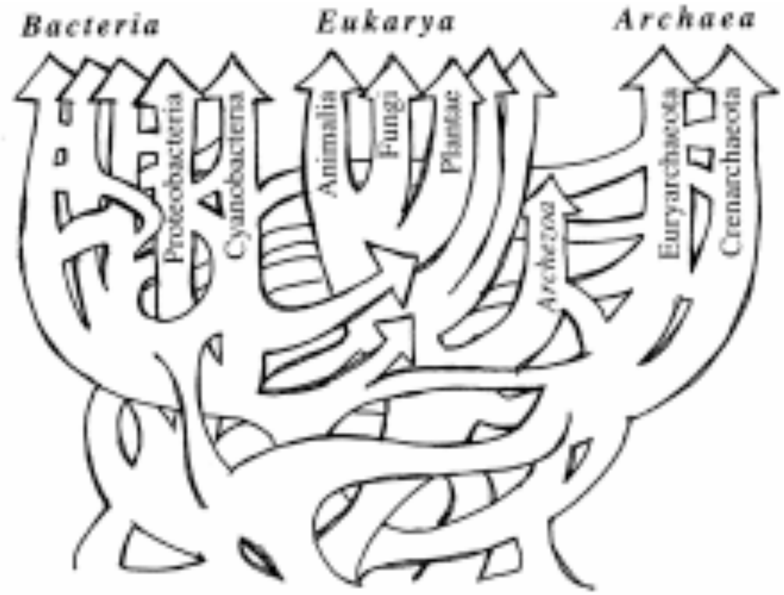
- We have established much of the statistical performance (consistency and convergence rates) of the major methods for phylogeny estimation.
- We have developed "fast converging" methods (guaranteed to reconstruct the true tree from polynomial length sequences) with excellent performance in practice.
- We have very fast methods for solving maximum likelihood and maximum parsimony, the major optimization problems, even for large datasets.

But the Standard Markov models *are too simple!*

- Sequences evolve just with substitutions
- Sites (i.e., positions) evolve identically and independently, and have “rates of evolution” that are drawn from a common distribution (typically gamma)
- Numerical parameters describe the probability of substitutions of each type on each edge of the tree

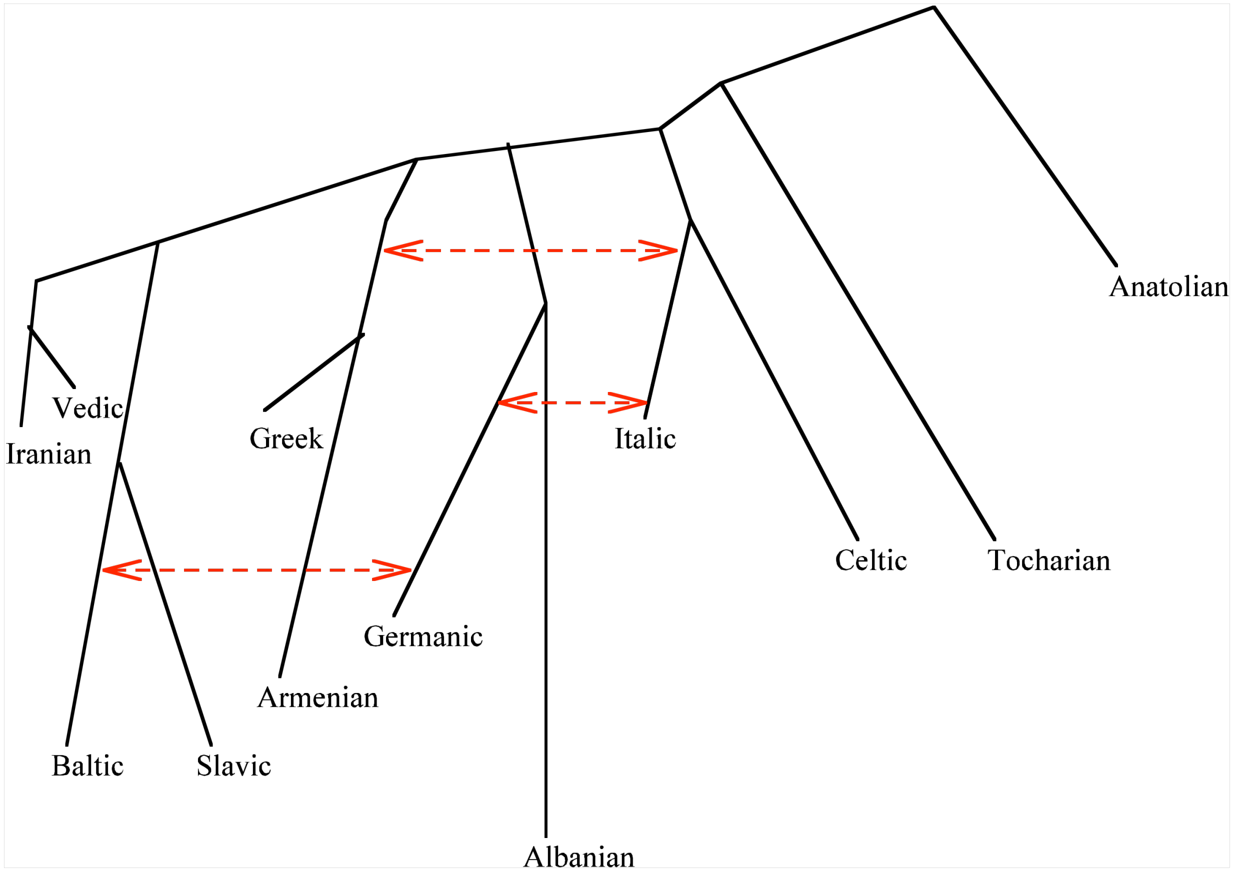
And all the positive results we’ve shown disappear under more realistic models

The “tree of life” is not a tree



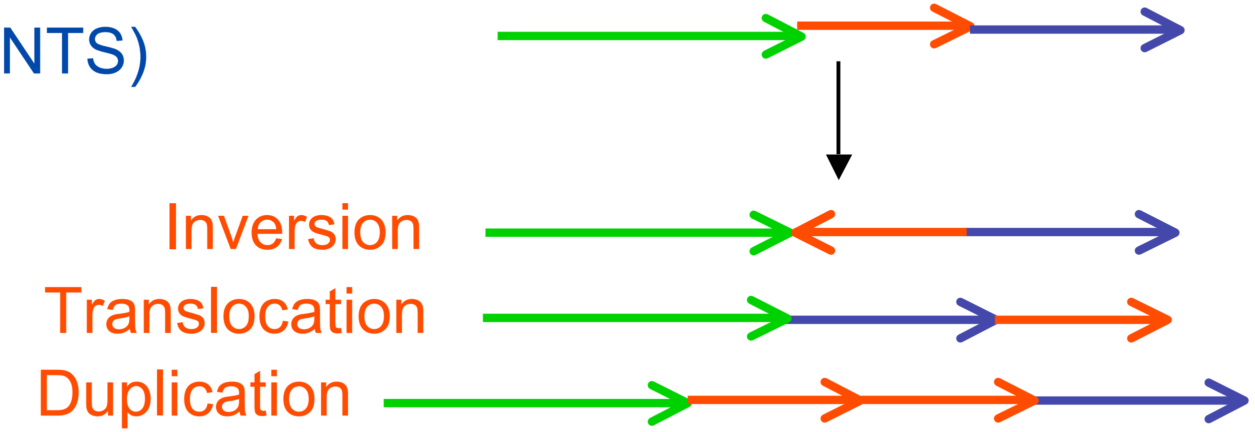
Reticulate evolution (horizontal gene transfer and hybridization) is also a problem

Languages also evolve with reticulation (Nakhleh et al., 2005)



Genome-scale evolution

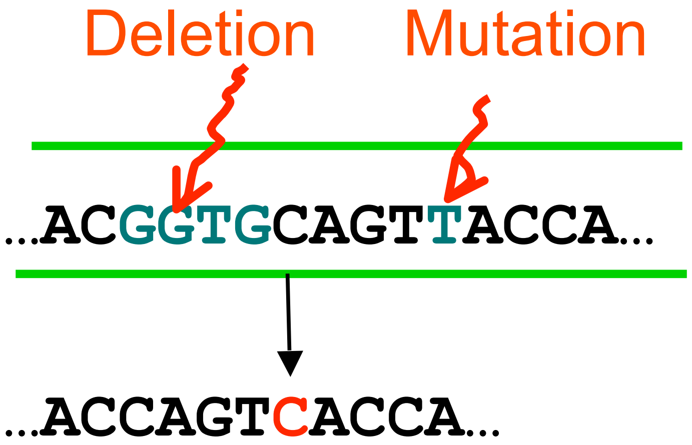
(REARRANGEMENTS)



Inversion

Translocation

Duplication

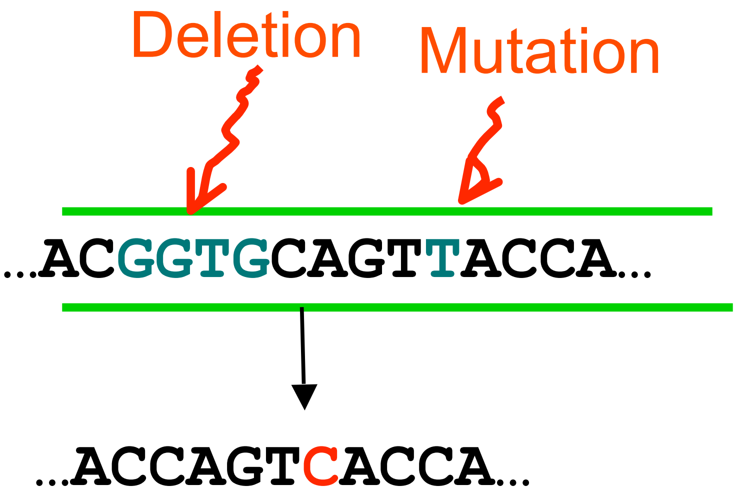


Deletion Mutation

...ACGGTGCAGTTACCA...

...ACCAGTCACCA...

indels (insertions and deletions) also occur!



Deletion

Mutation

...ACGGTGCAGTTACCA...

...ACCAGTCACCA...

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.

Input: unaligned sequences

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

Phase 1: Multiple Sequence Alignment

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

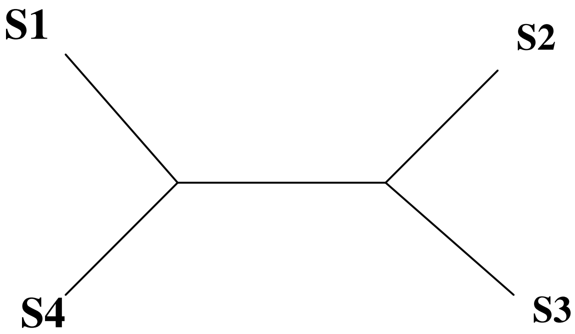
→

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

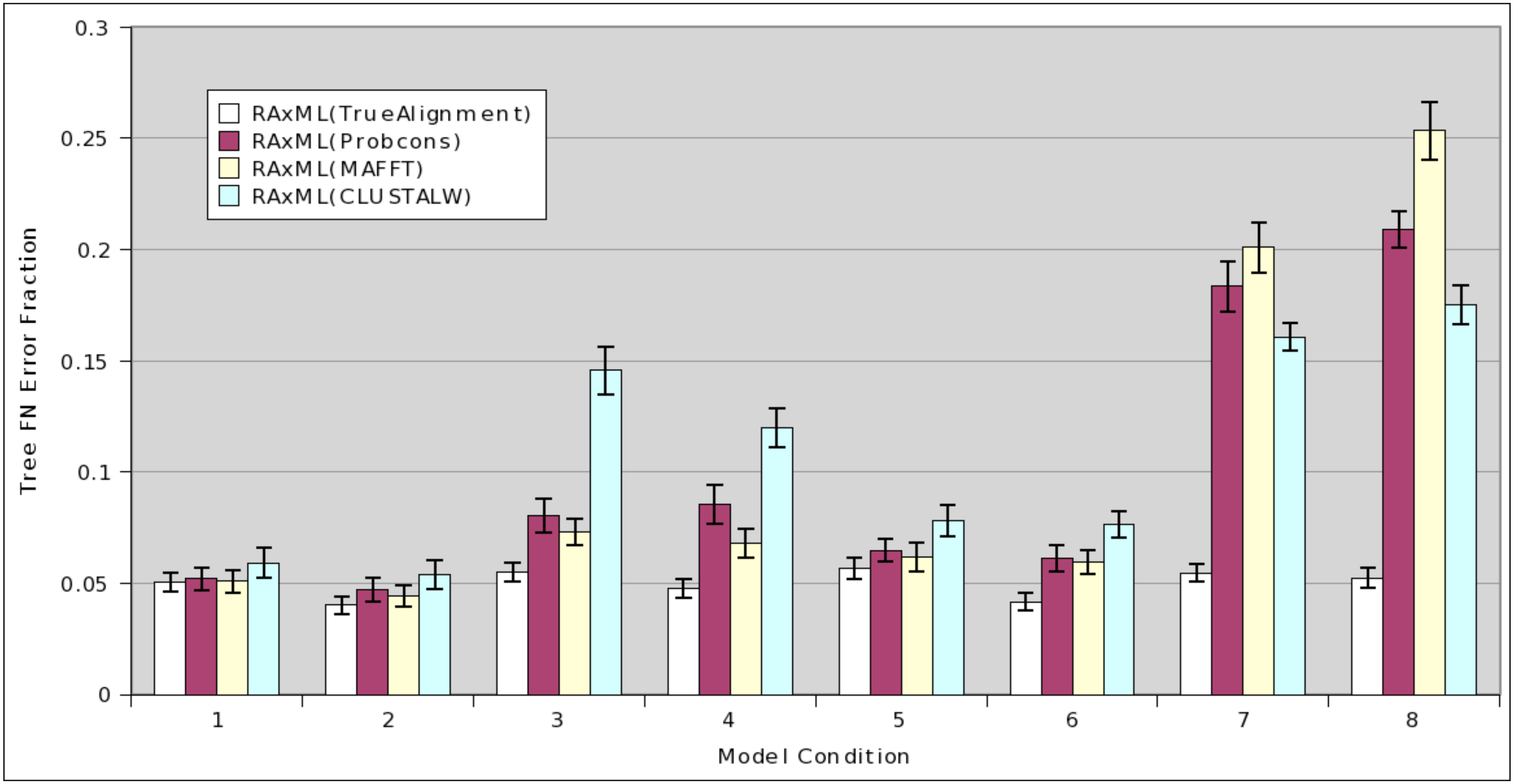
Phase 2: Construct tree

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

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



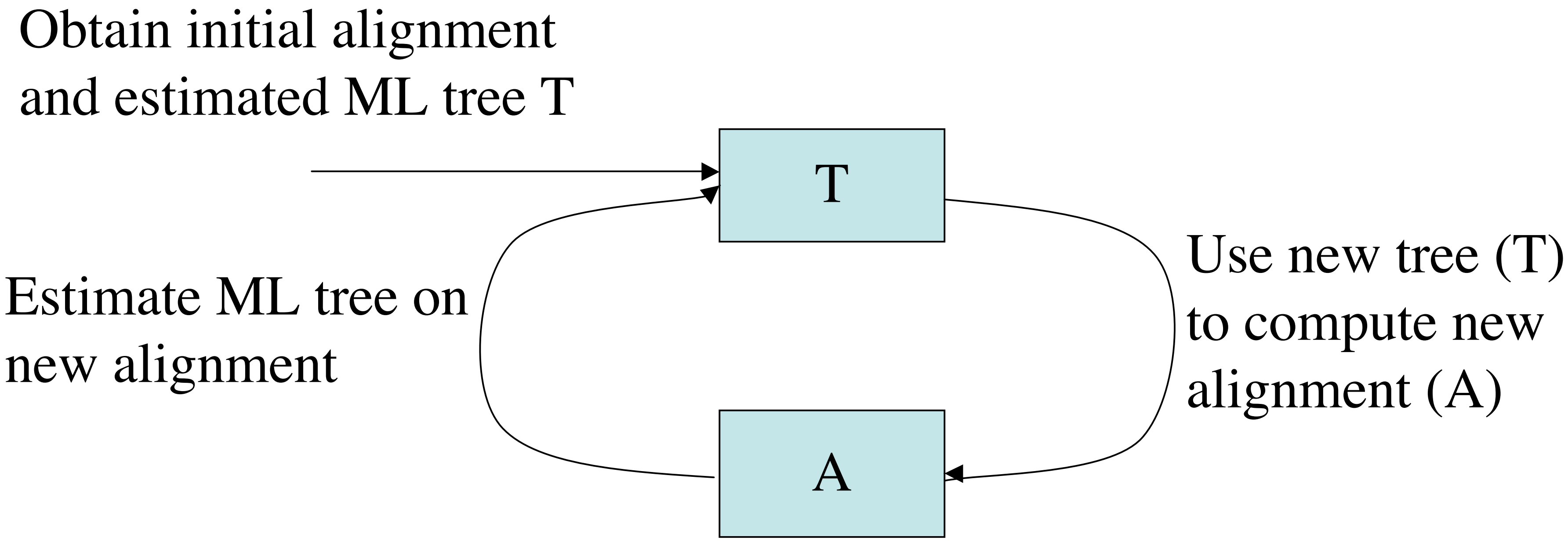
DNA sequence evolution



Simulation using ROSE: 100 taxon model trees, models 1-4 have “long gaps”, and 5-8 have “short gaps”, site substitution is HKY+Gamma

SATé Algorithm (unpublished)

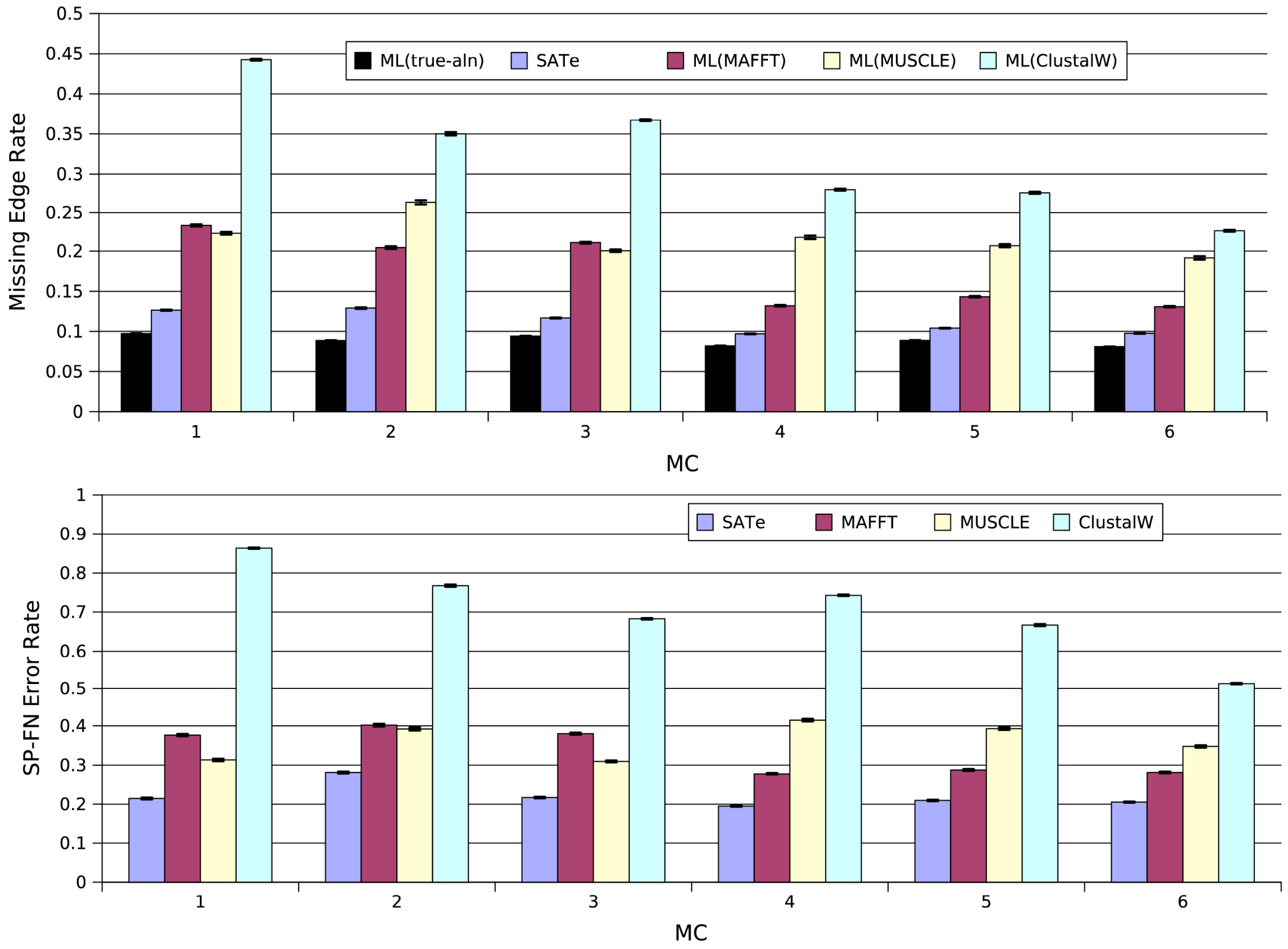
SATé keeps track of the maximum likelihood scores of the tree/alignment pairs it generates, and returns the best pair it finds



Obtain initial alignment and estimated ML tree T

Use new tree (T) to compute new alignment (A)

Estimate ML tree on new alignment



Models 1-3 have 1000 taxa, Models 4-6 have 500 taxa

(gap length distributions: long, medium, short)

Complexity *viz.* The Tree of Life

- **Algorithmic complexity** (e.g., running time and NP-hardness)
- **Sample size complexity** (e.g. how long do the sequences need to be to obtain a highly accurate reconstruction with high probability?)
- ***Stochastic model complexity*** *(i.e., how realistic are the models of evolution, and what are the consequences of making the models more realistic?)*

Thoughts

- Current models of sequence evolution are clearly too simple, and more realistic ones are not identifiable.
- The relative performance between methods can change as the models become more complex or as the number of taxa increases.
- We do not know how methods perform under realistic conditions (nor how long we need to let computationally intensive methods run).
- Therefore, *simulations should be done under very realistic (sufficiently complex) models, even if estimations are done under simpler models (and it is likely that estimations are best done under more realistic models, too).*

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:
  - Fast-converging methods: Peter Erdös, Daniel Huson, Bernard Moret, Luay Nakhleh, Usman Roshan, Katherine St. John, Michael Steel, and Laszlo Székély
  - Rec-I-DCM3: Usman Roshan, Bernard Moret, and Tiffani Williams
  - SATé: Randy Linder, Kevin Liu, Serita Nelesen, and Sindhu Raghavan

Simulated Model Conditions

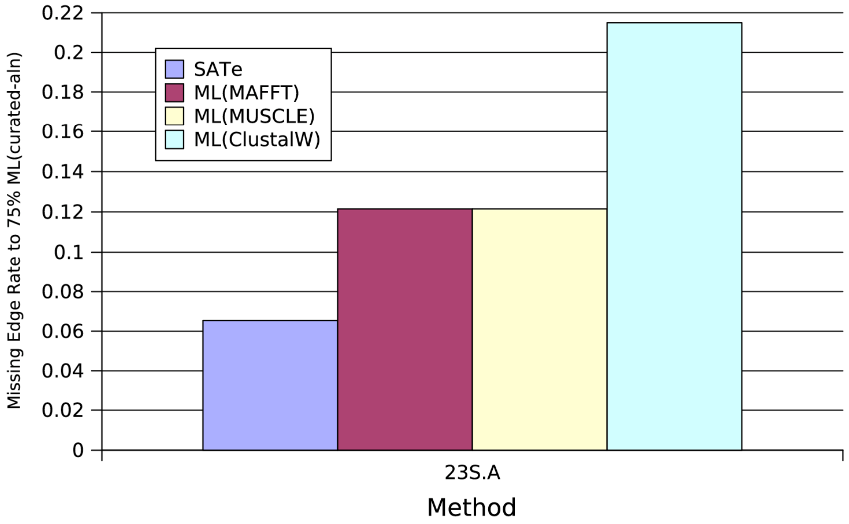
Model Condition	Taxa	Average gap length	ANHD	MNHD	Percent gaps
1	1000	9.2 (7.2)	69.2 (.01)	76.7 (.01)	72.1 (.19)
2	1000	5 (4.4)	68.0 (.02)	75.7 (.02)	70.4 (.10)
3	1000	2 (1.2)	69.1 (.01)	76.6 (.01)	41.7 (.14)
4	500	9.2 (7.2)	66.1 (.02)	74.3 (.01)	76.7 (.11)
5	500	5 (4.4)	66.3 (.02)	74.2 (.01)	64.7 (.14)
6	500	2 (1.2)	66.1 (.02)	74.2 (.02)	42.8 (.14)

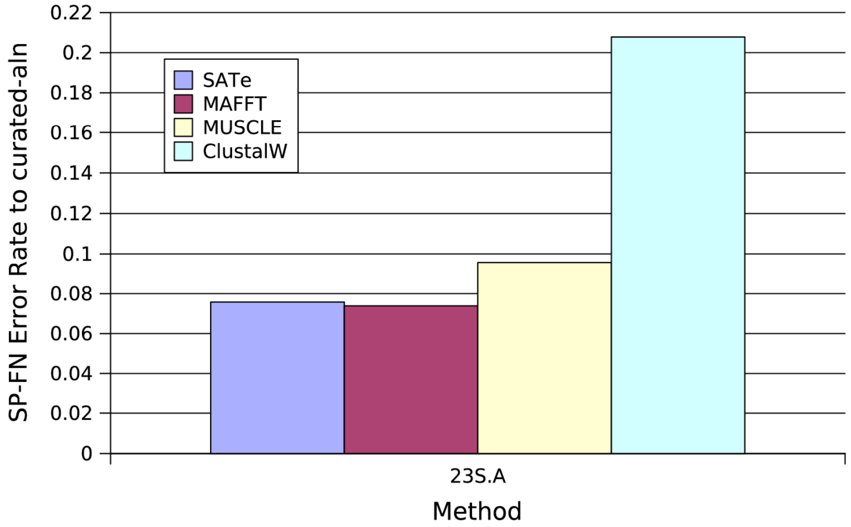
- ANHD is the average normalized Hamming distance. MNHD is the maximum normalized Hamming distance. (Normalized Hamming distances are also known as *p-distances.)*
- Standard deviations are given parenthetically for average gap length, and standard errors are given parenthetically for all other statistics.

Biological datasets

- We used 8 different biological datasets with curated alignments (produced by Robin Gutell (UT-Austin)) based upon secondary structures.
- We computed various alignments, and maximum likelihood trees on each alignment.
- We ran SATé for 24 hours, producing an alignment/tree pair.
- We evaluated alignments and trees in comparison to the curated alignment and to the reference tree (the 75% bootstrap maximum likelihood tree on the curated alignment), respectively.

Results for 23S rRNA dataset


Missing Edge Rate to 75% ML(curated-aln)
SATe
ML(MAFFT)
ML(MUSCLE)
ML(ClustalW)
23S.A
Method


SP-FN Error Rate to curated-aln
SATe
MAFFT
MUSCLE
ClustalW
23S.A
Method