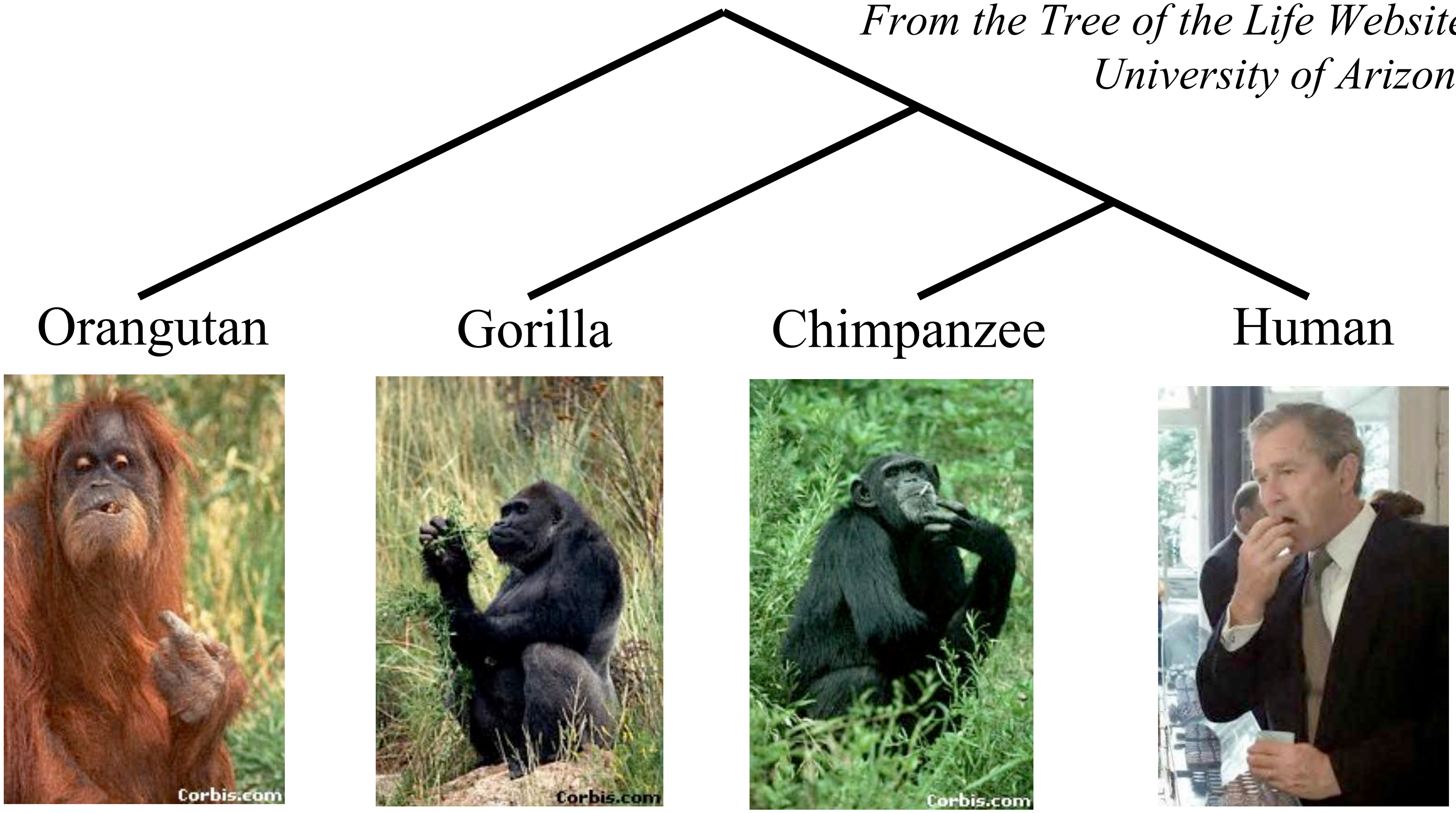
Introduction to Phylogenetic Estimation Algorithms

Tandy Warnow

Questions

- What is a phylogeny?
- What data are used?
- What is involved in a phylogenetic analysis?
- What are the most popular methods?
- What is meant by “accuracy”, and how is it measured?

Phylogeny



From the Tree of the Life Website, University of Arizona

Orangutan

Gorilla

Chimpanzee

Human

Data

- Biomolecular sequences: DNA, RNA, amino acid, in a multiple alignment
- Molecular markers (e.g., SNPs, RFLPs, etc.)
- Morphology
- Gene order and content

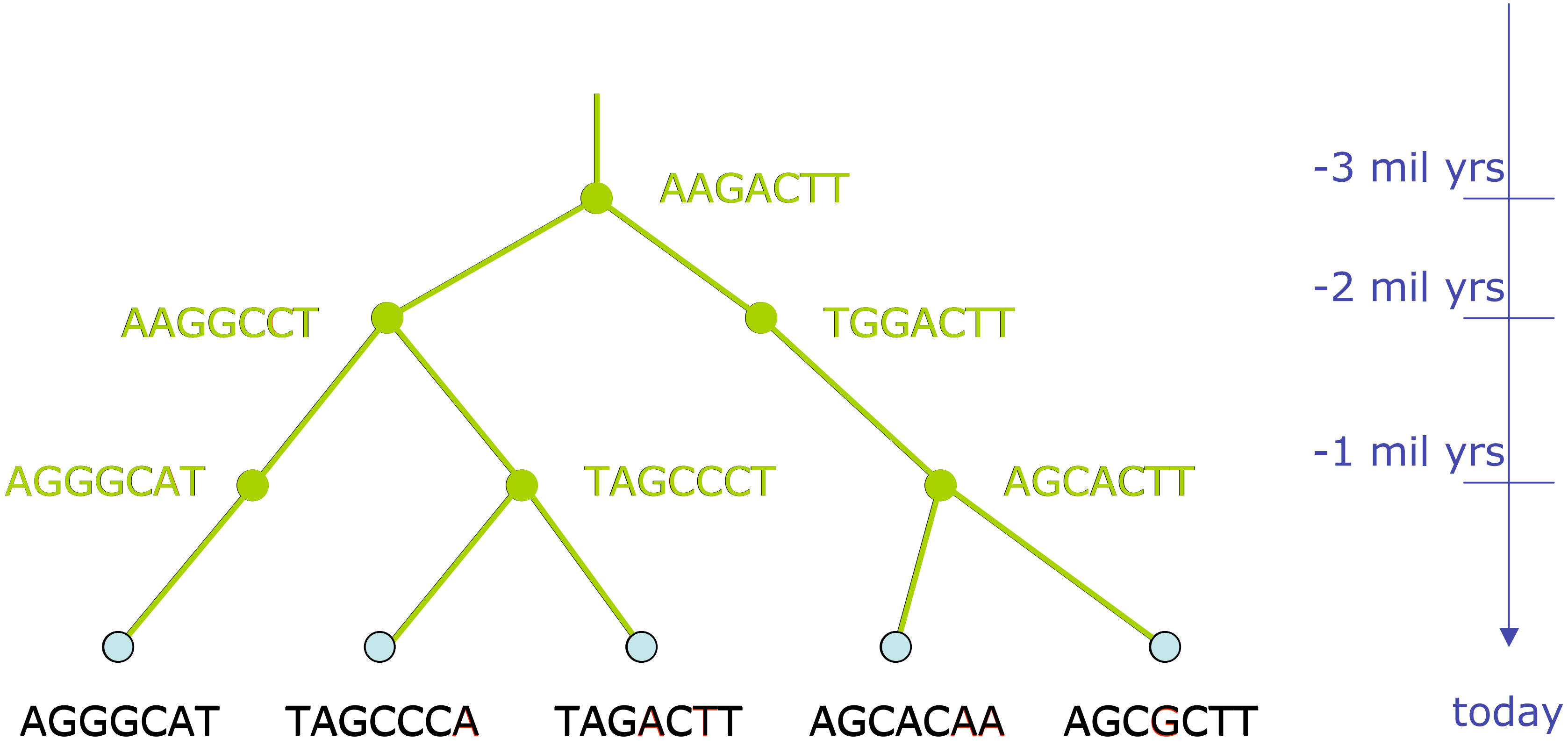
These are “character data”: each character is a function mapping the set of taxa to distinct states (equivalence classes), with evolution modelled as a process that changes the state of a character

Data

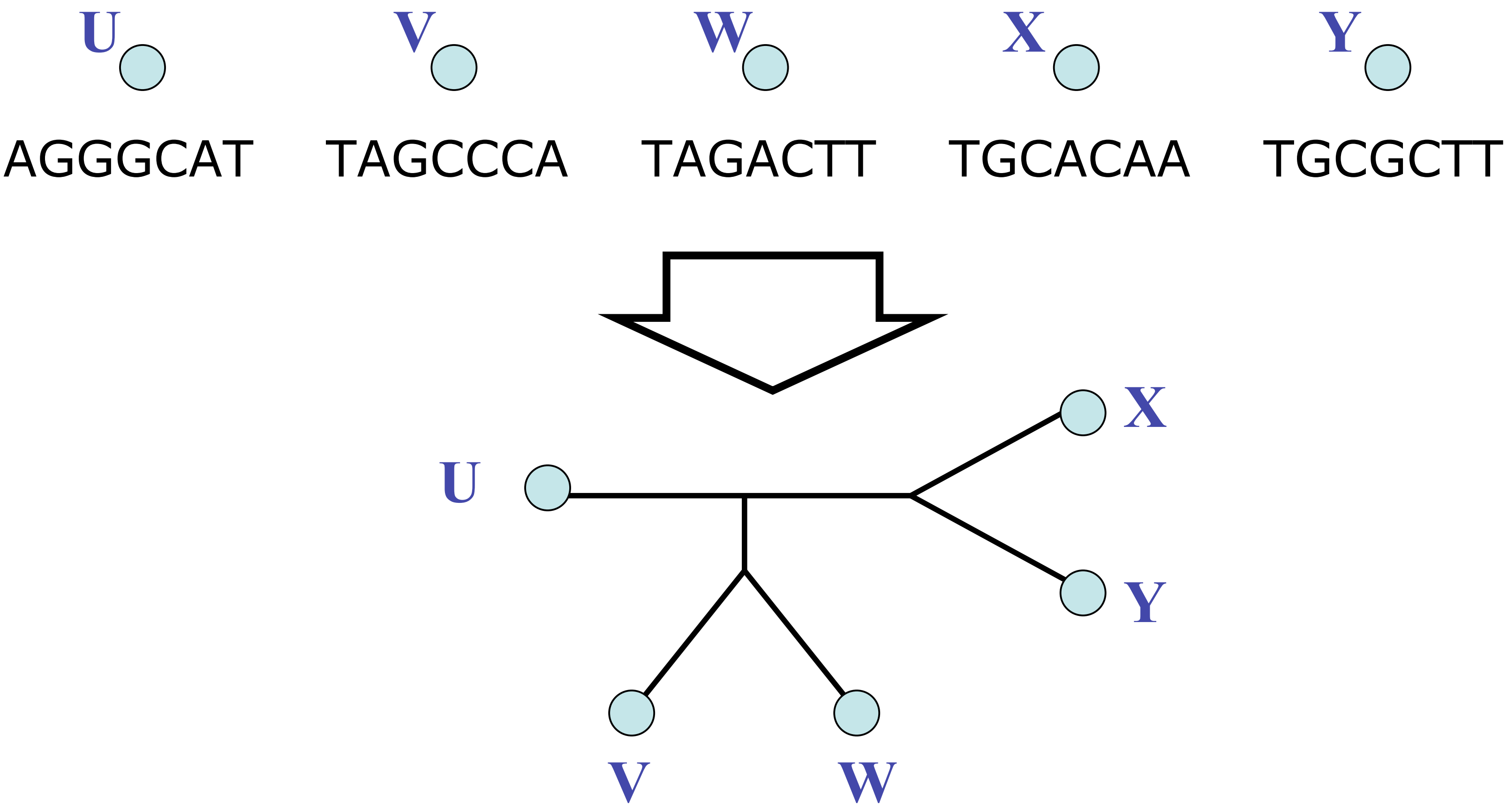
- Biomolecular sequences: DNA, RNA, amino acid, in a multiple alignment
- Molecular markers (e.g., SNPs, RFLPs, etc.)
- Morphology
- Gene order and content

These are “character data”: each character is a function mapping the set of taxa to distinct states (equivalence classes), with evolution modelled as a process that changes the state of a character

DNA Sequence Evolution



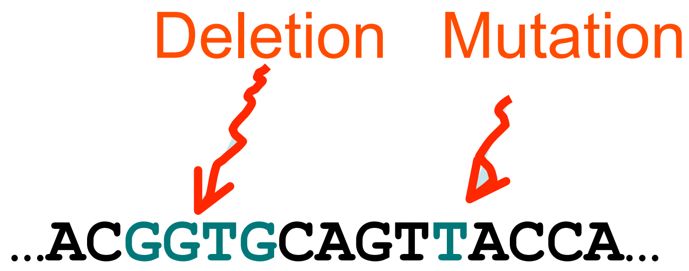
Phylogeny Problem



Indels and substitutions at the DNA level

…**ACGGTGCAGTTACCA**…

Indels and substitutions at the DNA level

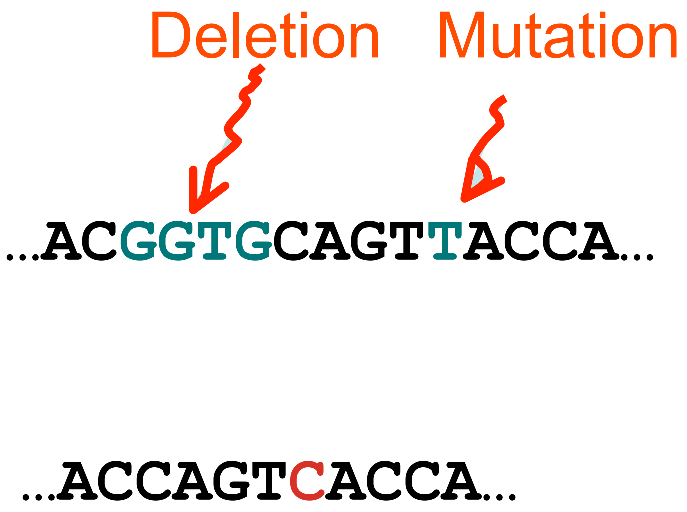


Deletion

Mutation

...ACGGTGCAGTTACCA...

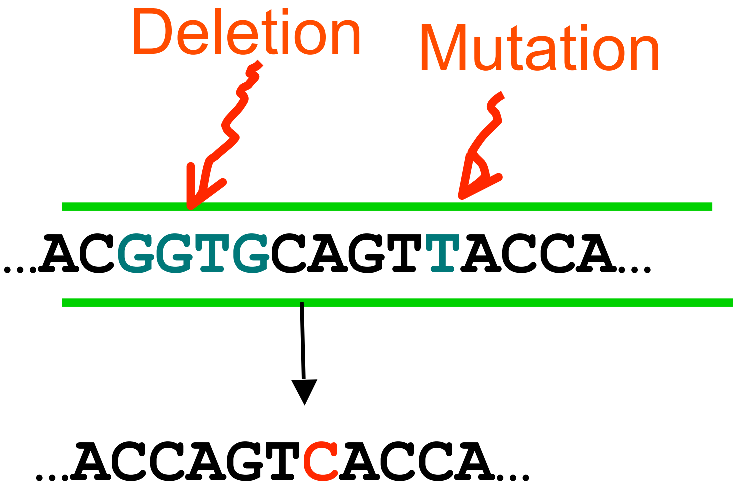
Indels and substitutions at the DNA level



Deletion Mutation

...ACGGTGCAGTTACCA...

...ACCAGTCACCA...



Deletion

Mutation

...ACGGTGCAGTTACCA...

...ACCAGTCACCA...

The **true** 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.

Easy Sequence Alignment

```
B_WEAU160   ATGGAAAACAGATGGCAGGTGATGATTGTGTGGCAAGTAGACAGG 45
A_U455      .............................A.....G......... 45
A_IFA86     ...................................G......... 45
A_92UG037   ...................................G......... 45
A_Q23       ...................C...............G......... 45
B_SF2       ............................................. 45
B_LAI       ............................................. 45
B_F12       ............................................. 45
B_HXB2R     ............................................. 45
B_LW123     ............................................. 45
B_NL43      ............................................. 45
B_NY5       ............................................. 45
B_MN        ............C........................C....... 45
B_JRCSF     ............................................. 45
B_JRFL      ............................................. 45
B_NH52      ........................G.................... 45
B_OYI       ............................................. 45
B_CAM1      ............................................. 45
```

Harder Sequence Alignment

```
B_WEAU160      ATGAGAGTGAAGGGGATCAGGAAGAATTATCAGCACTTG   39
A_U455         ..........T......ACA..G........CTTG....   39
A_SF1703       ..........T......ACA..T...C.G...AA....A   39
A_92RW020.5        ......G......ACA..C..G..GG..AA.....   35
A_92UG031.7        ......G.A....ACA..G.....GG........A   35
A_92UG037.8        ......T......AGA..G........CTTG..G.   35
A_TZ017        ..........G..A...G.A..G............A..A   39
A_UG275A       ....A..C..T.....CACA..T.....G...AA...G.   39
A_UG273A       .................ACA..G.....GG.........   39
A_DJ258A       ..........T......ACA...........CA.T...A   39
A_KENYA        ..........T.....CACA..G.....G.........A   39
A_CARGAN       ..........T......ACA............A......   39
A_CARSAS       ................CACA.........CTCT.C....   39
A_CAR4054      .............A..CACA..G.....GG..CA.....   39
A_CAR286A      ................CACA..G.....GG..AA.....   39
A_CAR4023      .............A.---------..A............   30
A_CAR423A      .............A.---------..A............   30
A_VI191A       .................ACA..T.....GG..A......   39
```

Multiple sequence alignment

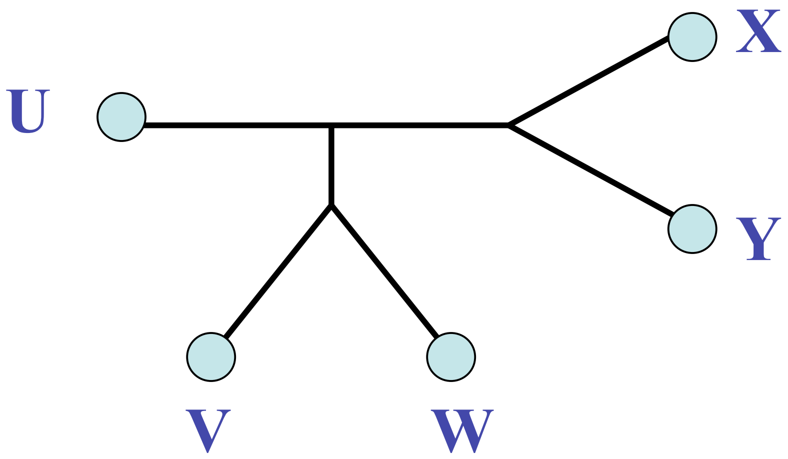
Objective:

Estimate the “true alignment” (defined by the sequence of evolutionary events)

Typical approach:

1. Estimate an initial tree
2. Estimate a multiple alignment by performing a “progressive alignment” up the tree, using Needleman-Wunsch (or a variant) to align alignments

U	AGTGGAT
V	TATGCCCA
W	TATGACTT
X	AGCCCTA
Y	AGCCCGCTT



Input: unaligned sequences

S1 = AGGCTATCACCTGACCTCCA

S2 = TAGCTATCACGACCGC

S3 = TAGCTGACCGC

S4 = TCACGACCGACA

Phase 1: Multiple Sequence Alignment

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

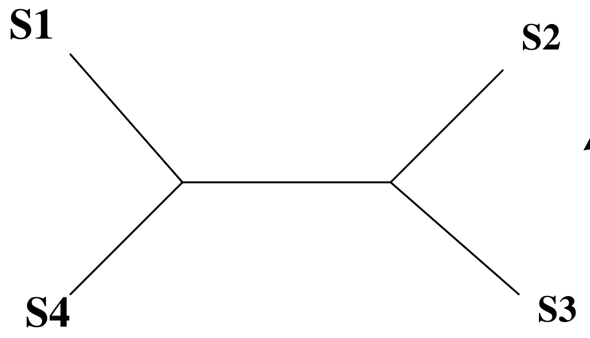
→

```
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
```



So many methods!!!

Alignment method

- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- MAFFT
- Prank
- Muscle
- Di-align
- T-Coffee
- Satchmo
- Etc.

Phylogeny method

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

So many methods!!!

Alignment method

- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- MAFFT
- Prank
- Muscle
- Di-align
- T-Coffee
- Satchmo
- Etc.

Phylogeny method

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

So many methods!!!

Alignment method

- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- MAFFT
- Prank
- Muscle
- Di-align
- T-Coffee
- Satchmo
- Etc.

Blue = used by systematists

Purple = recommended by Edgar and Batzoglou for protein alignments

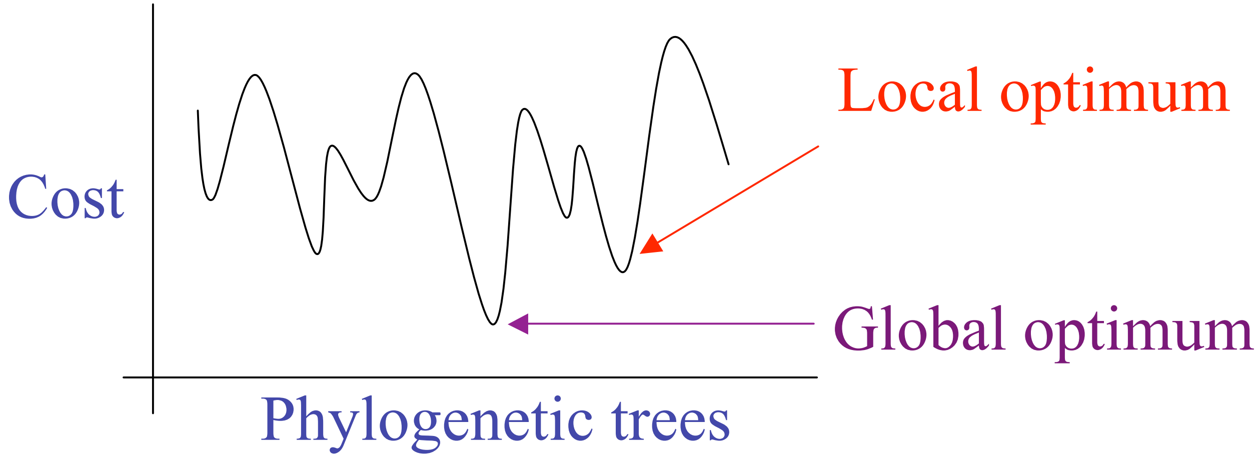
Phylogeny method

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

Phylogenetic reconstruction methods

1. Polynomial time distance-based methods: UPGMA, Neighbor Joining, FastME, Weighbor, etc.

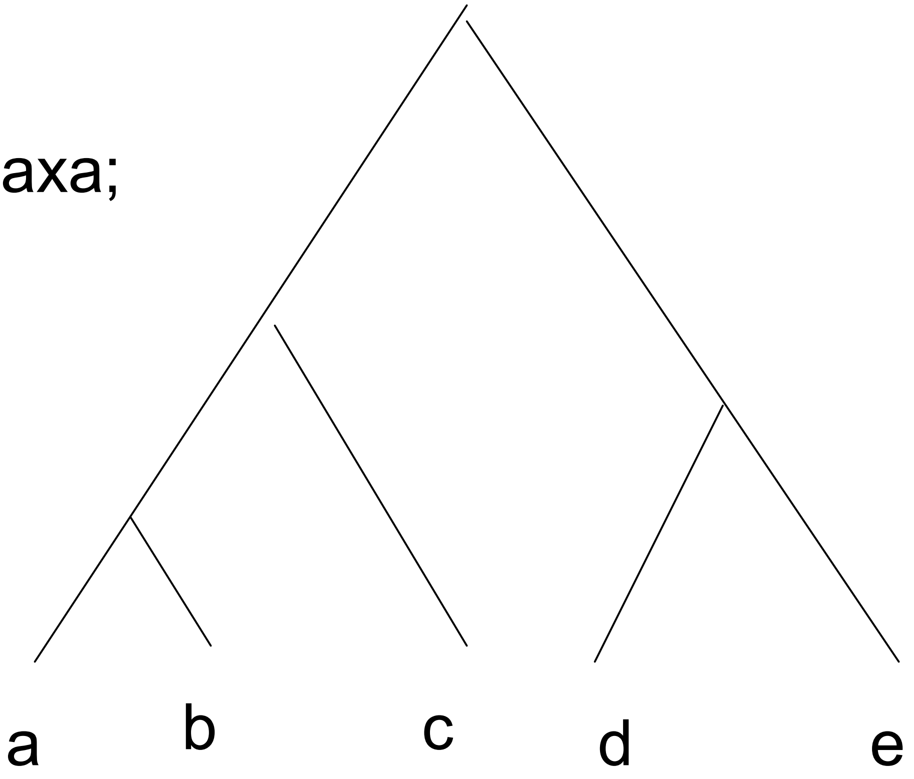
2. Hill-climbing heuristics for NP-hard optimization criteria (Maximum Parsimony and Maximum Likelihood)



3. Bayesian methods

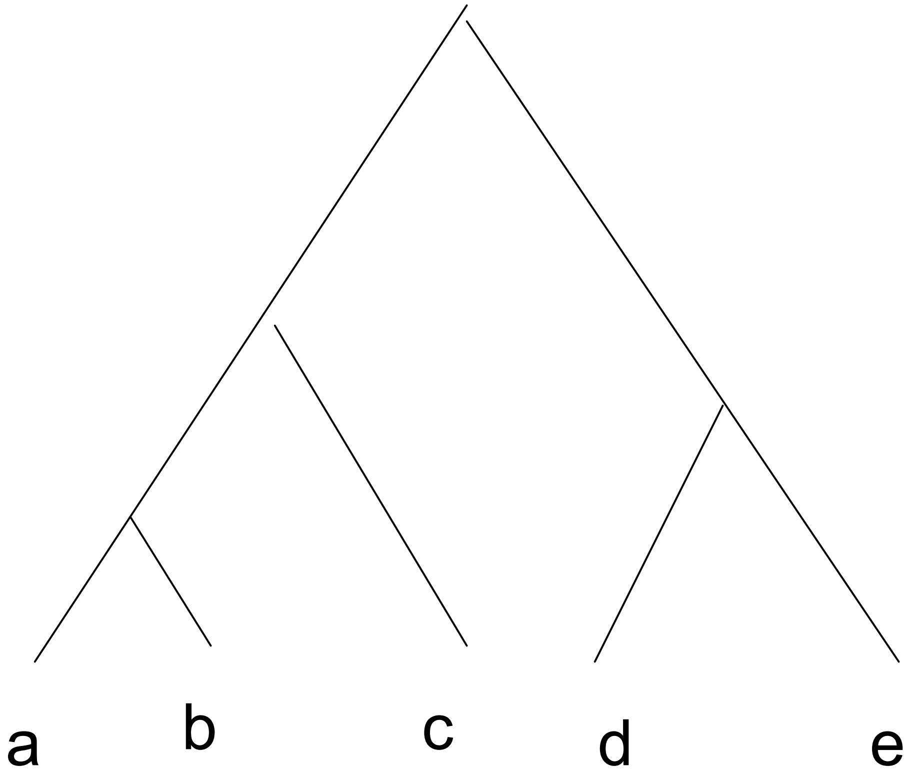
UPGMA

```
While |S|>2:
    find pair x,y of closest taxa;
    delete x
    Recurse on S-{x}
    Insert y as sibling to x
    Return tree
```



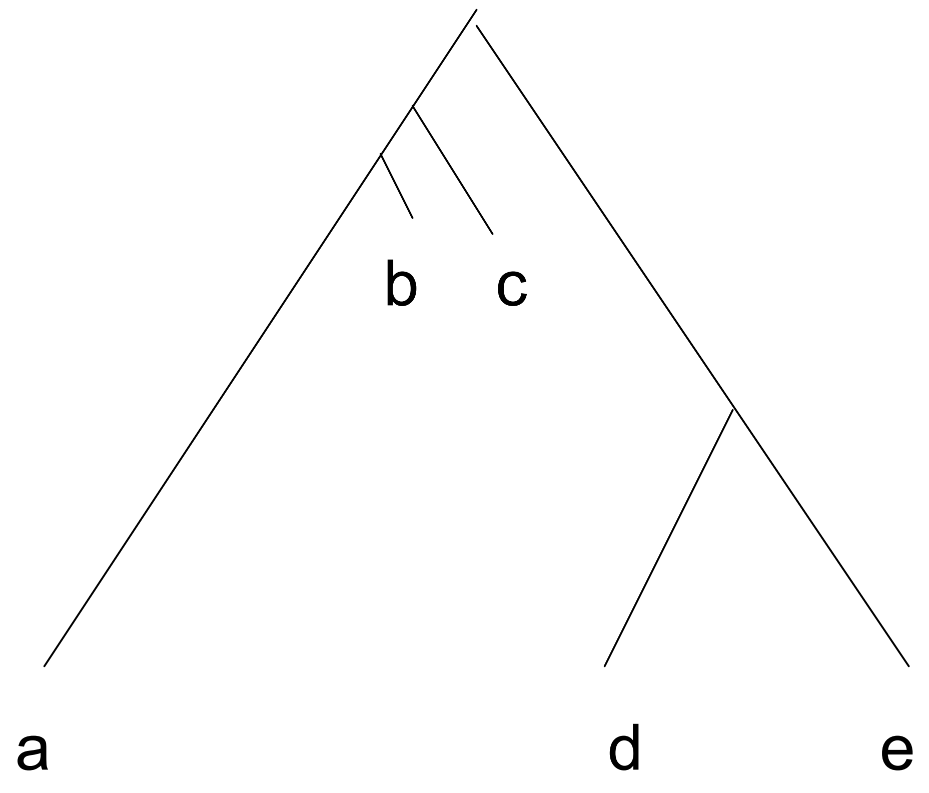
UPGMA

Works when evolution is “clocklike”



UPGMA

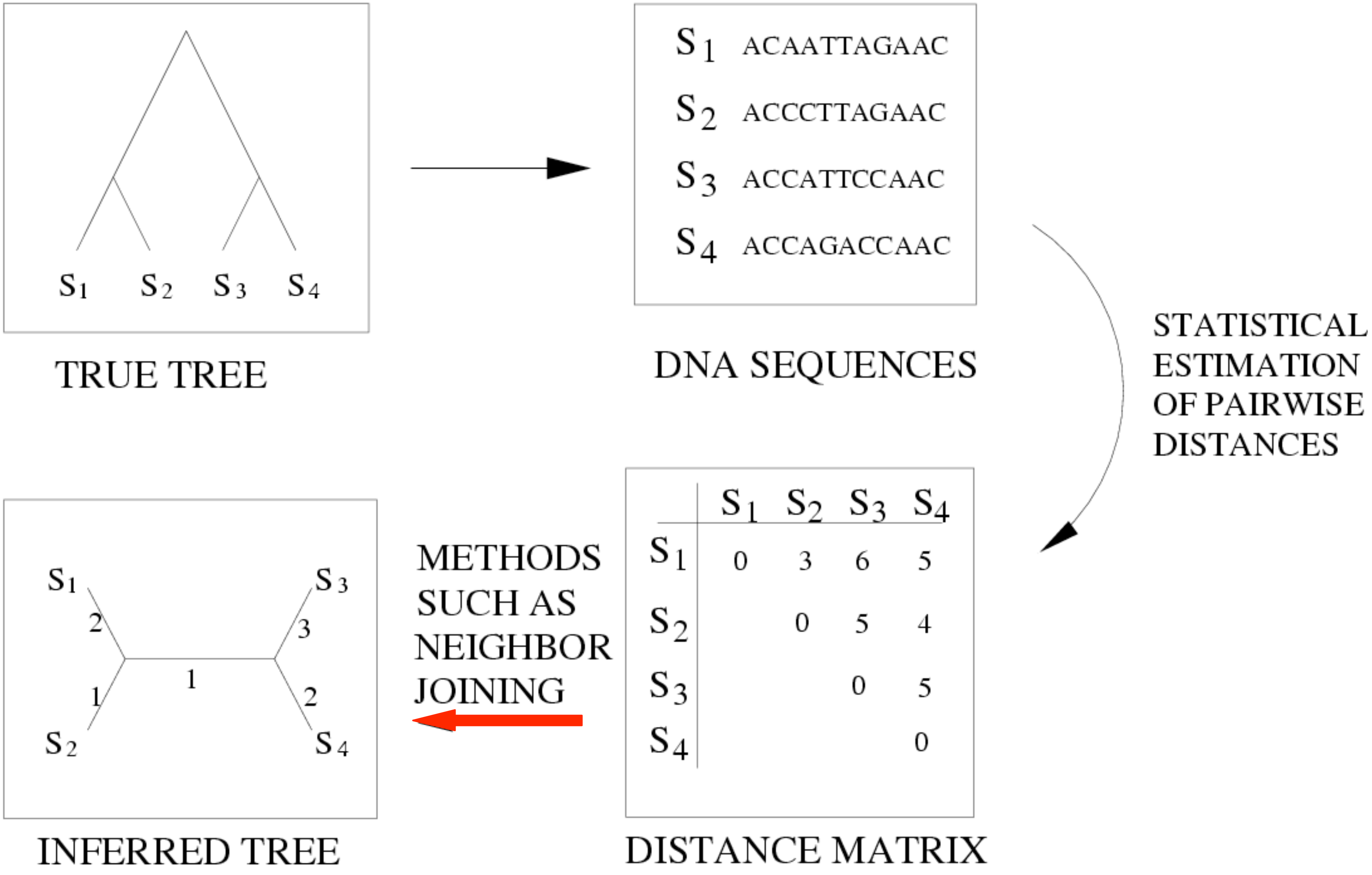
Fails to produce true tree if evolution deviates too much from a clock!



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 mathematical score (e.g. tree length or likelihood score) on real data.

Distance-based Methods



TRUE TREE

S_1 ACAATTAGAAC

S_2 ACCCTTAGAAC

S_3 ACCATTCCAAC

S_4 ACCAGACCAAC

DNA SEQUENCES

STATISTICAL
ESTIMATION
OF PAIRWISE
DISTANCES

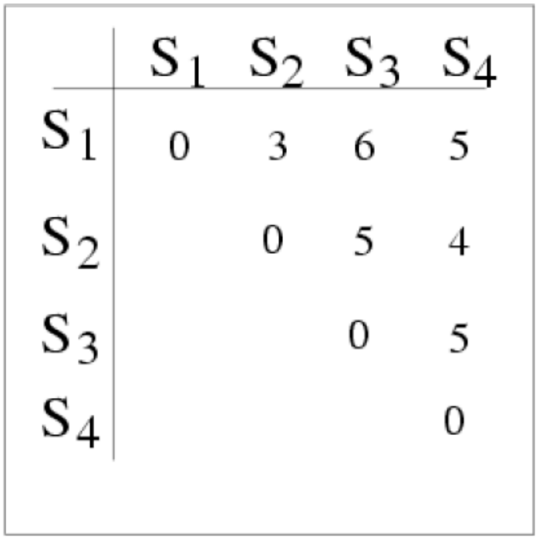
	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

DISTANCE MATRIX

METHODS
SUCH AS
NEIGHBOR
JOINING

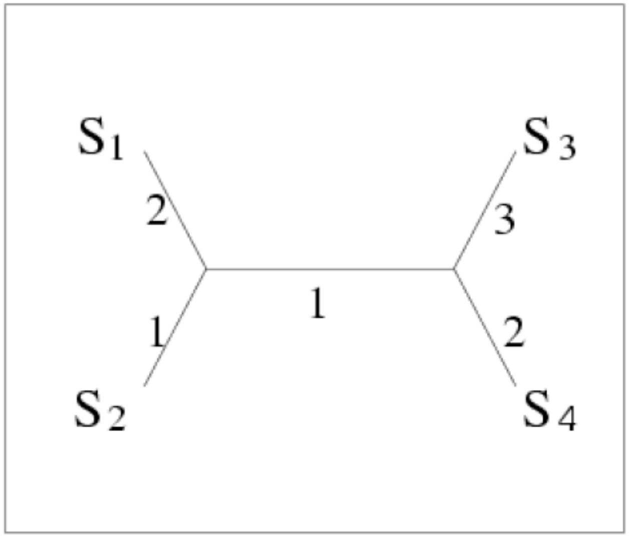
INFERRED TREE

Additive Distance Matrices



	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

POLYTIME
INVERTIBLE



Four-point condition

- A matrix D is additive if and only if for every four indices i,j,k,l, the maximum and median of the three pairwise sums are identical

$$D_{ij}+D_{kl} < D_{ik}+D_{jl} = D_{il}+D_{jk}$$

The Four-Point Method computes trees on quartets using the Four-point condition

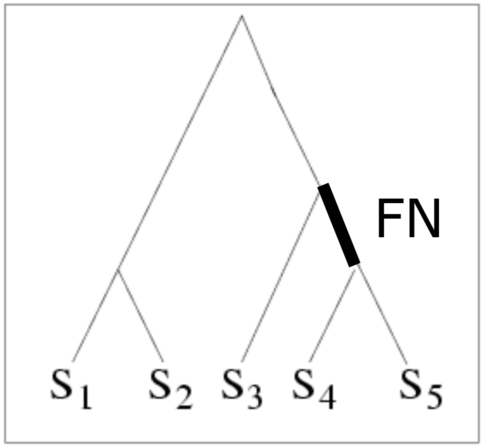
Naïve Quartet Method

- Compute the tree on each quartet using the four-point condition
- Merge them into a tree on the entire set if they are compatible:
  - Find a sibling pair A,B
  - Recurse on S-{A}
  - If S-{A} has a tree T, insert A into T by making A a sibling to B, and return the tree

Better distance-based methods

- Neighbor Joining
- Minimum Evolution
- Weighted Neighbor Joining
- Bio-NJ
- DCM-NJ
- And others

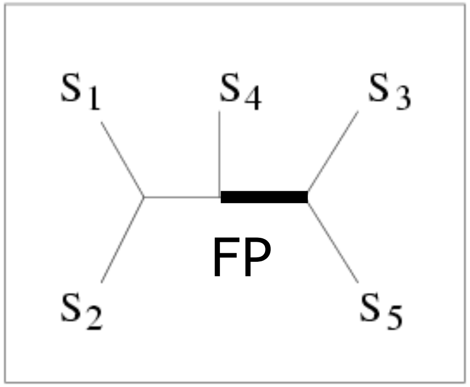
Quantifying Error



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

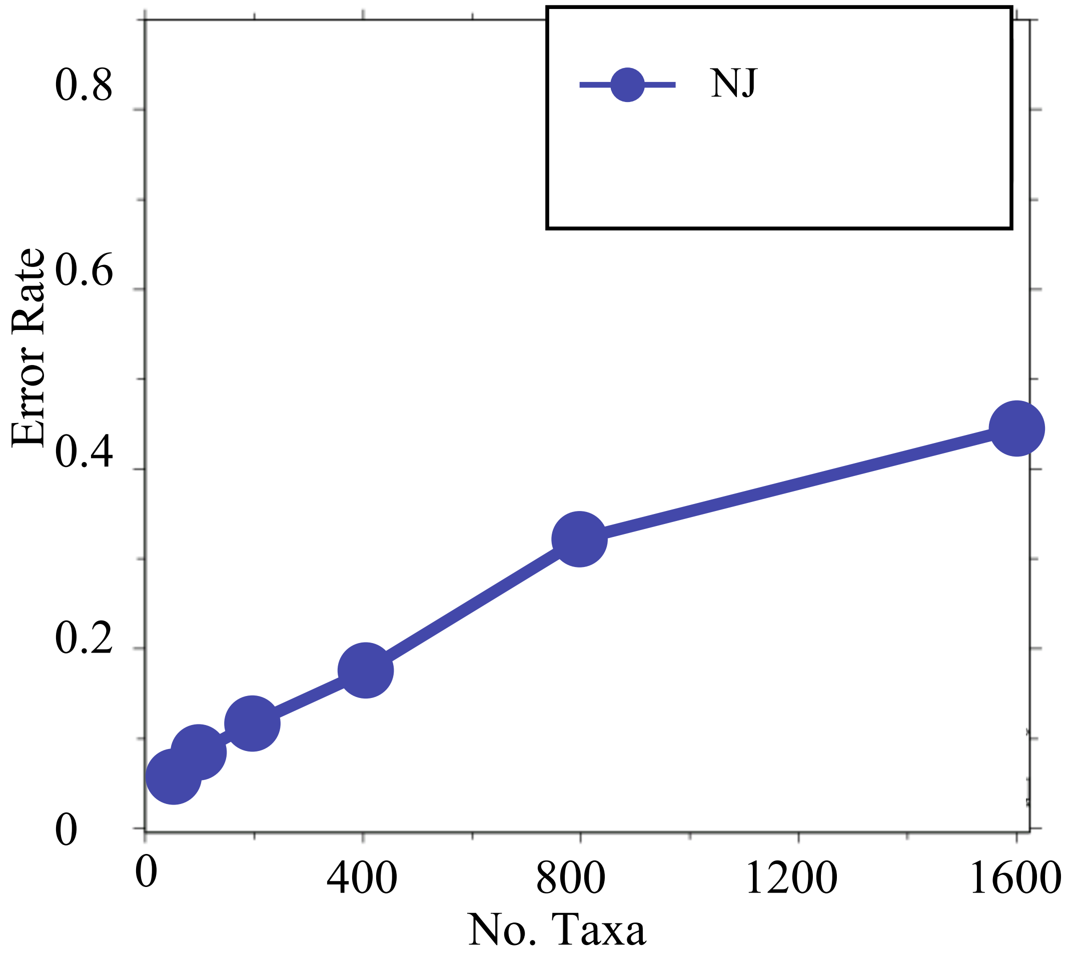


INFERRED TREE

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

50% error rate

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.

“Character-based” methods

- Maximum parsimony
- Maximum Likelihood
- Bayesian MCMC (also likelihood-based)

These are more popular than distance-based methods, and tend to give more accurate trees. However, these are computationally intensive!

Standard problem: Maximum Parsimony (Hamming distance Steiner Tree)

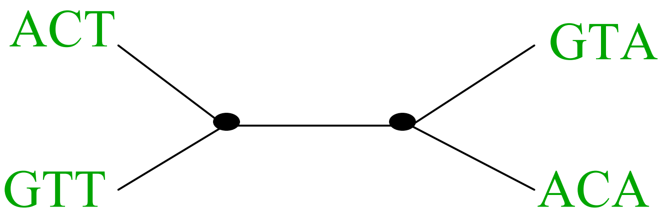
- **Input**: Set S of n aligned sequences of length k
- **Output**: A phylogenetic tree T
  - leaf-labeled by sequences in S
  - additional sequences of length k labeling the internal nodes of T

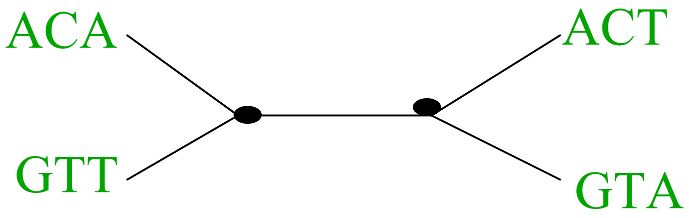
such that $\sum_{(i,j)\in E(T)} H(i,j)$ is minimized.

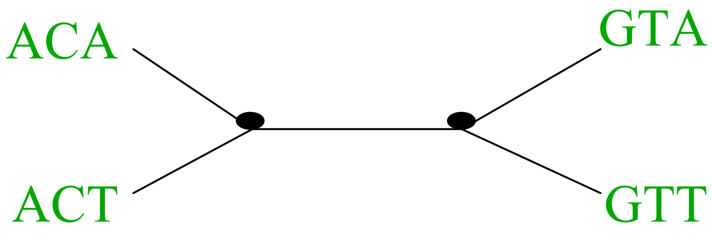
Maximum parsimony (example)

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

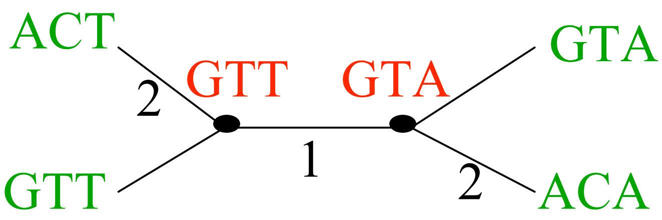
Maximum Parsimony



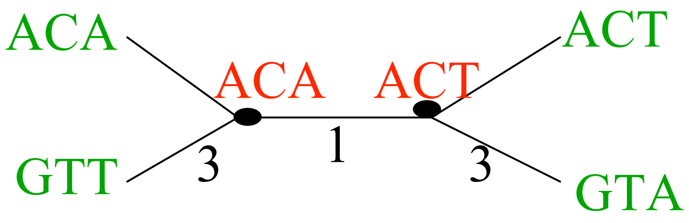




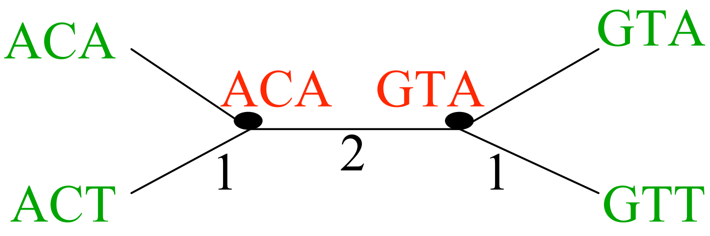
Maximum Parsimony



MP score = 5



MP score = 7

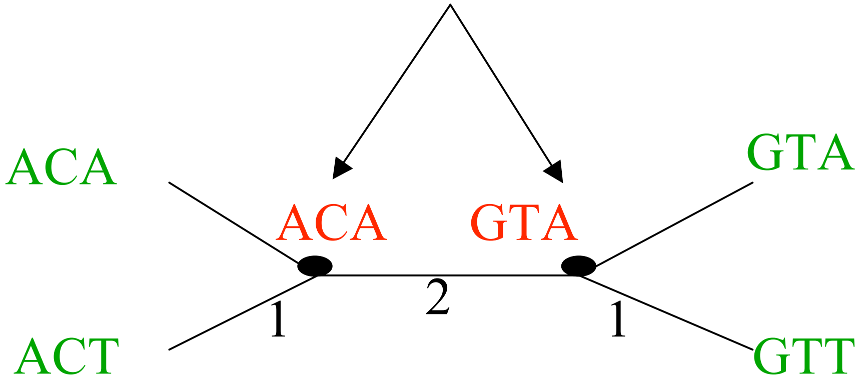


MP score = 4

Optimal MP tree

Maximum Parsimony: computational complexity

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



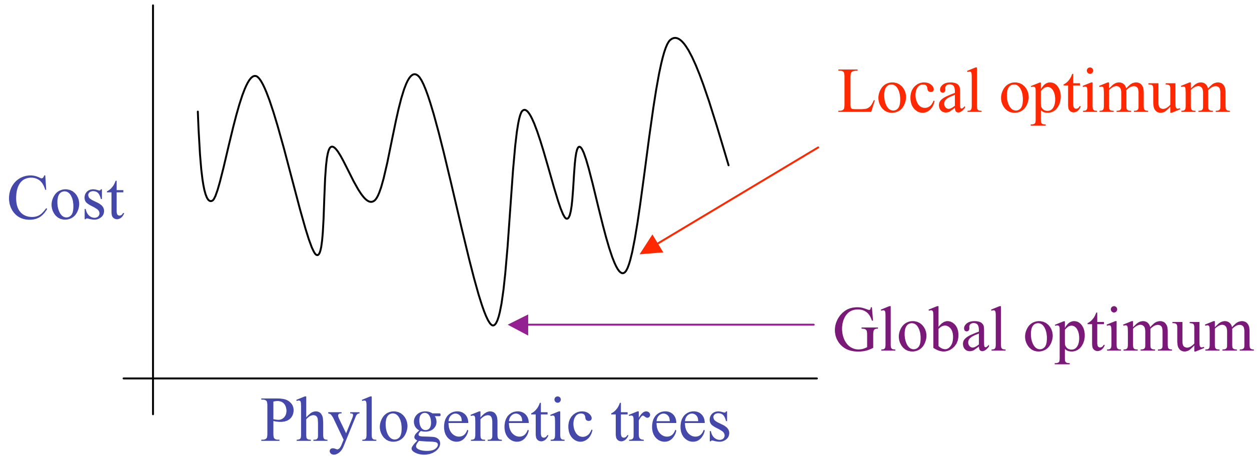
MP score = 4

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

But solving this problem exactly is … unlikely

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

Local search strategies



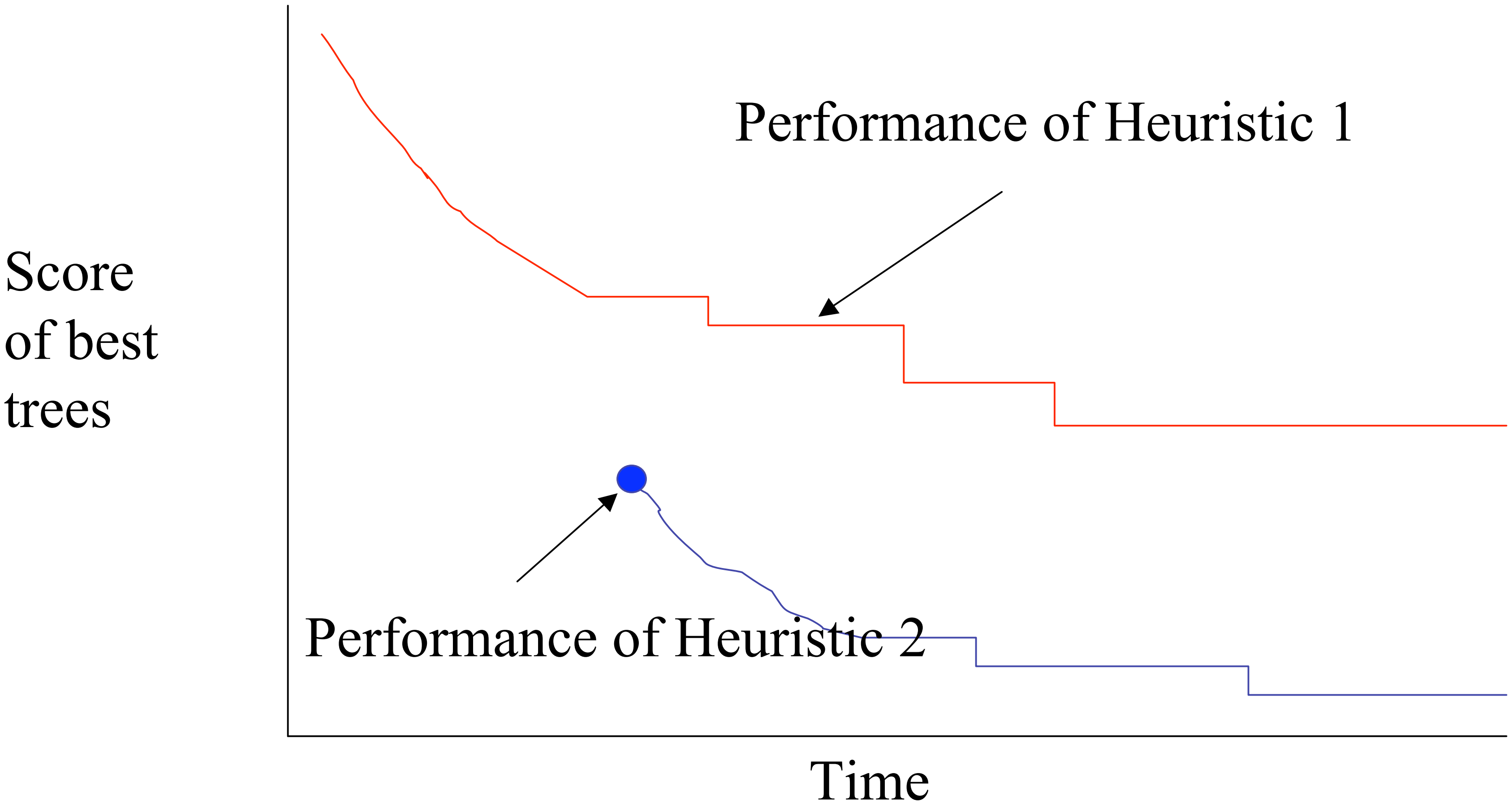
Local search strategies

- Hill-climbing based upon topological changes to the tree

- Incorporating randomness to exit from local optima

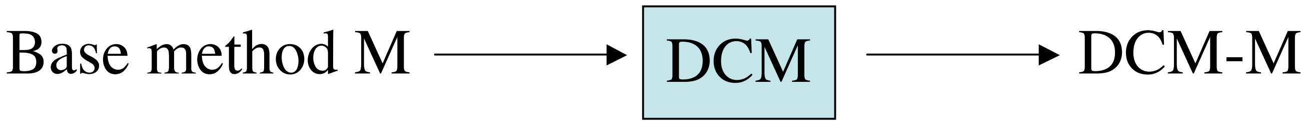
Evaluating heuristics with respect to MP or ML scores

Fake study



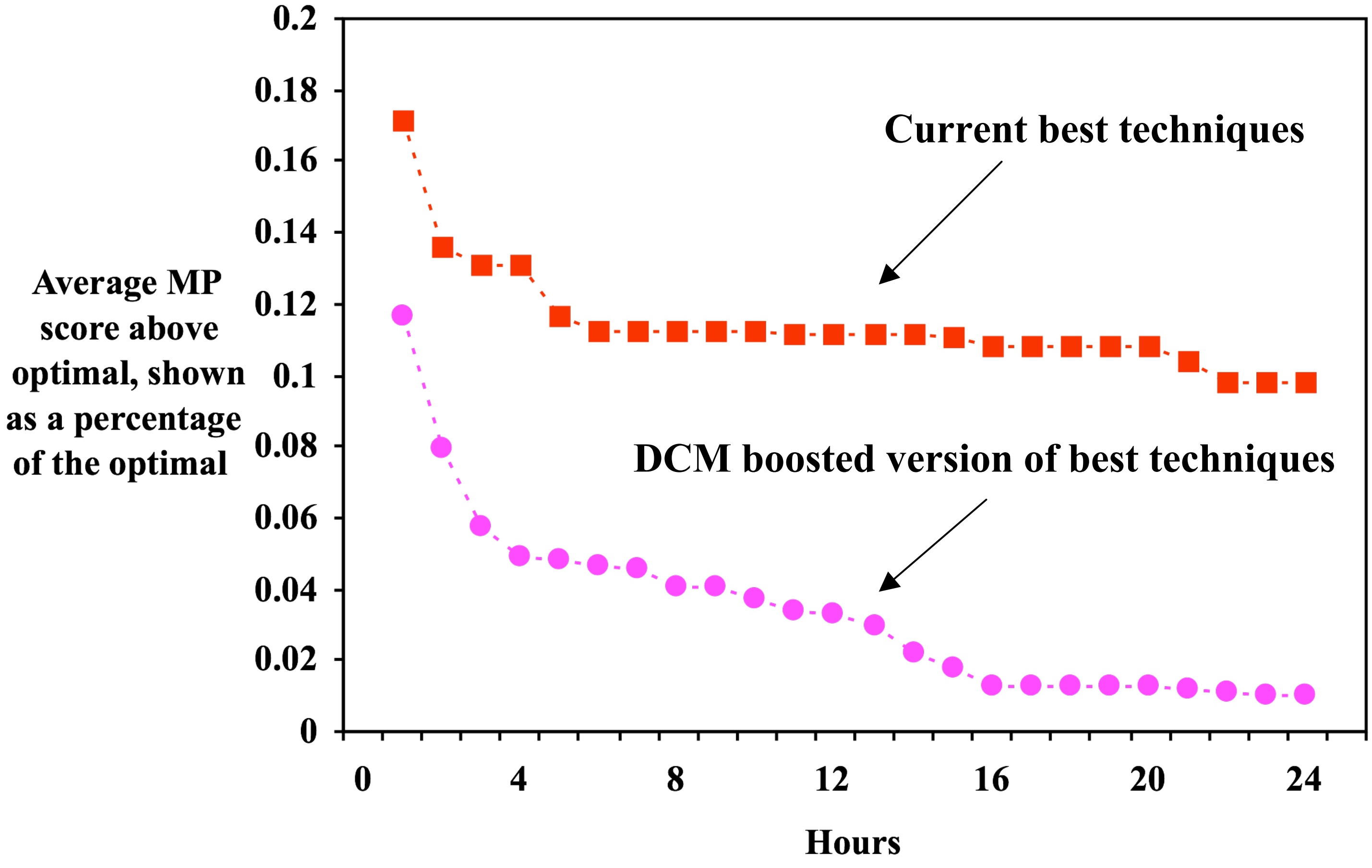
“Boosting” MP heuristics

- We use “Disk-covering methods” (DCMs) to improve heuristic searches for MP and ML



Base method M → DCM → DCM-M

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



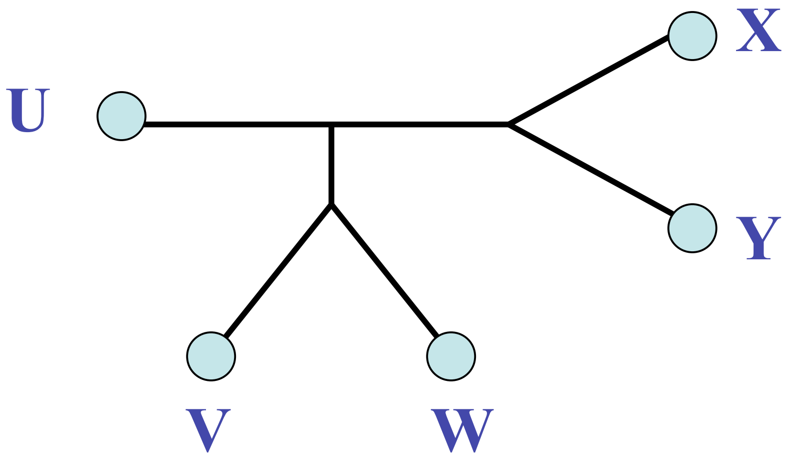
Comparison of TNT to Rec-I-DCM3(TNT) on one large dataset

Current methods

- Maximum Parsimony (MP):
  - TNT
  - PAUP* (with Rec-I-DCM3)
- Maximum Likelihood (ML)
  - RAxML (with Rec-I-DCM3)
  - GARLI
  - PAUP*
- Datasets with up to a few thousand sequences can be analyzed in a few days
- Portal at www.phylo.org

But…

U	AGTGGAT
V	TATGCCCA
W	TATGACTT
X	AGCCCTA
Y	AGCCCGCTT



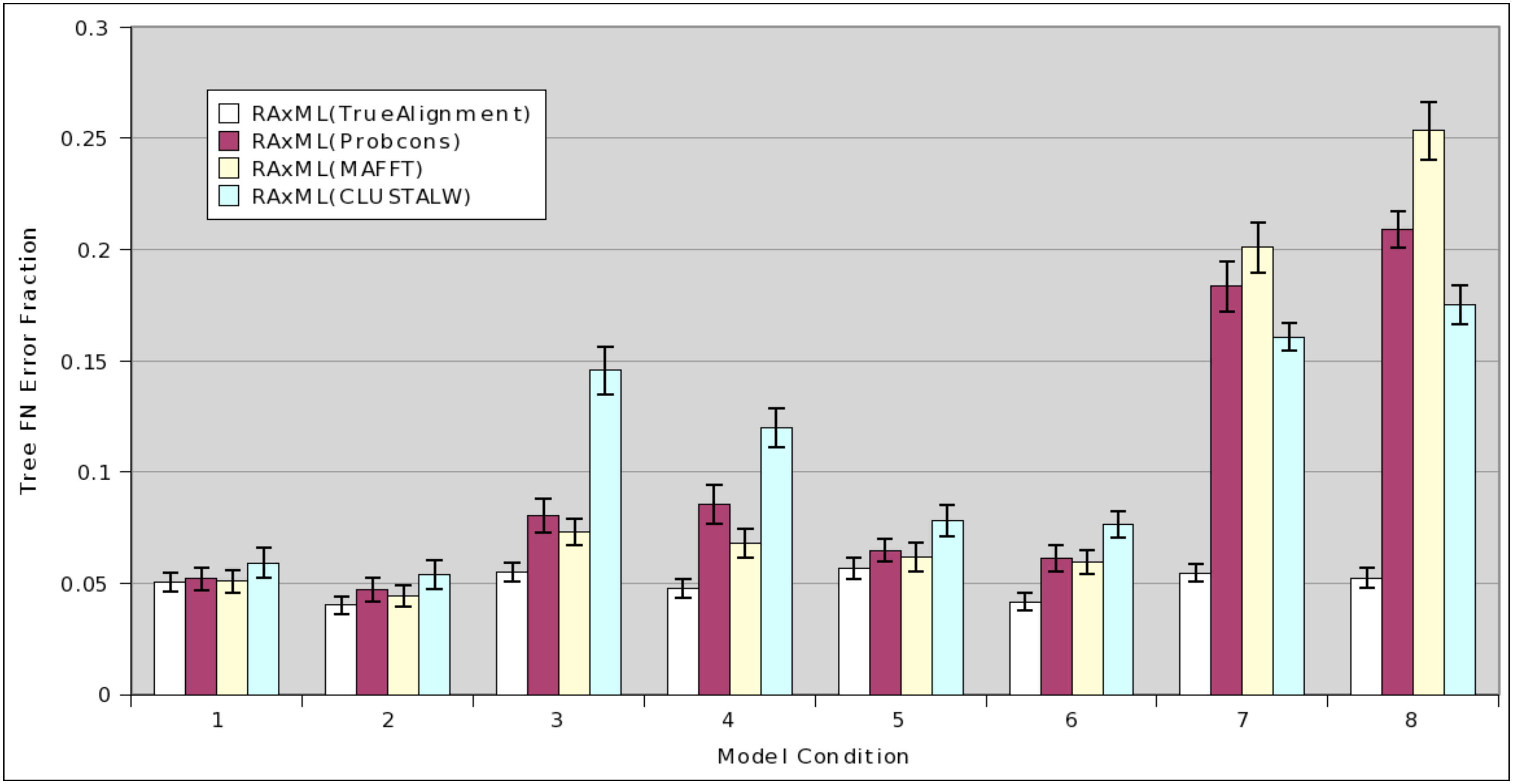
- Phylogenetic reconstruction methods assume the sequences all have the same length.
- Standard models of sequence evolution used in maximum likelihood and Bayesian analyses assume sequences evolve only via substitutions, producing sequences of equal length.
- And yet, almost all nucleotide datasets evolve with insertions and deletions (“indels”), producing datasets that violate these models and methods.

How can we reconstruct phylogenies from sequences of unequal length?

Basic Questions

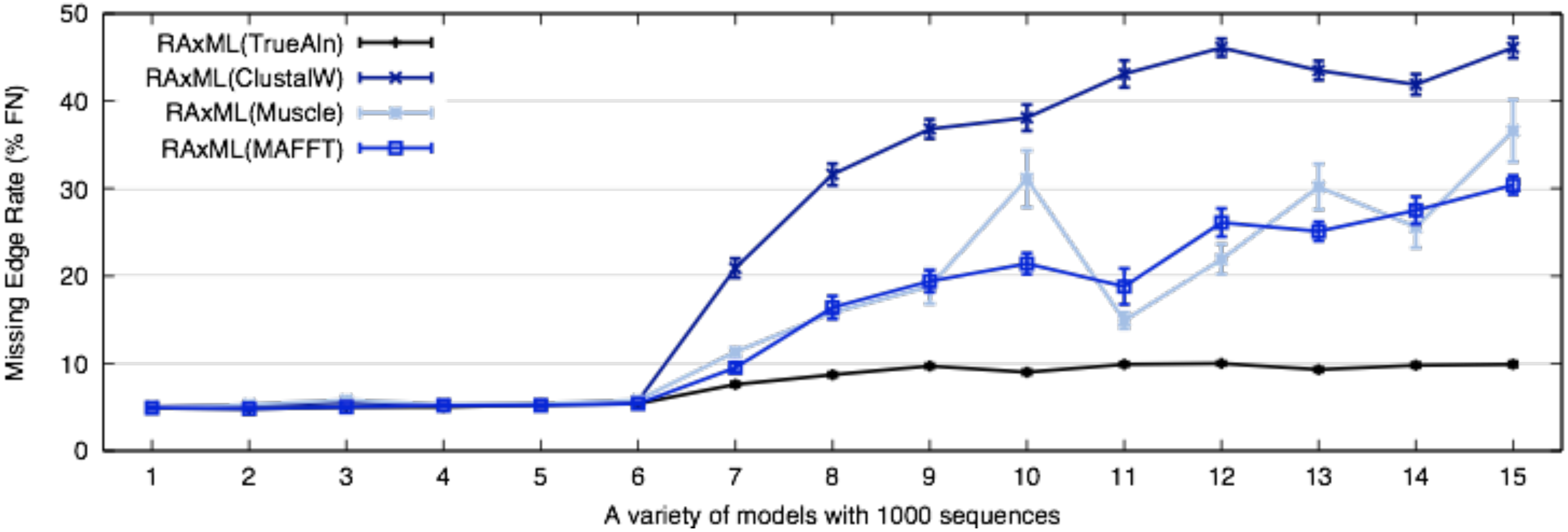
- Does improving the alignment lead to an improved phylogeny?
- Are we getting good enough alignments from MSA methods? (In particular, is ClustalW - the usual method used by systematists - good enough?)
- Are we getting good enough trees from the phylogeny reconstruction methods?
- Can we improve these estimations, perhaps through **simultaneous estimation** of trees and alignments?

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

Results



Model difficulty