Phylogenetic analysis

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Dr. Jaspreet Kaur Dhanjal
Assistant Professor, Center for Computational Biology
Email ID: jaspreet@iiitd.ac.in

September 08, 2025

Character based methods

Character based methods:

- treat the sequences from a vertical perspective.
- they search for each column of the alignment, the simplest explanation for how the characters evolved.

Taxa	Characters
Species A	ATGGCTATTCTTATAGTACG
Species B	ATCGCTAGTCTTATATTACA
Species C	TTCACTAGACCTGTGGTCCA
Species D	TTGACCAGACCTGTGGTCCG
Species E	TTGACCAGTTCTCTAGTTCG

Maximum Parsimony

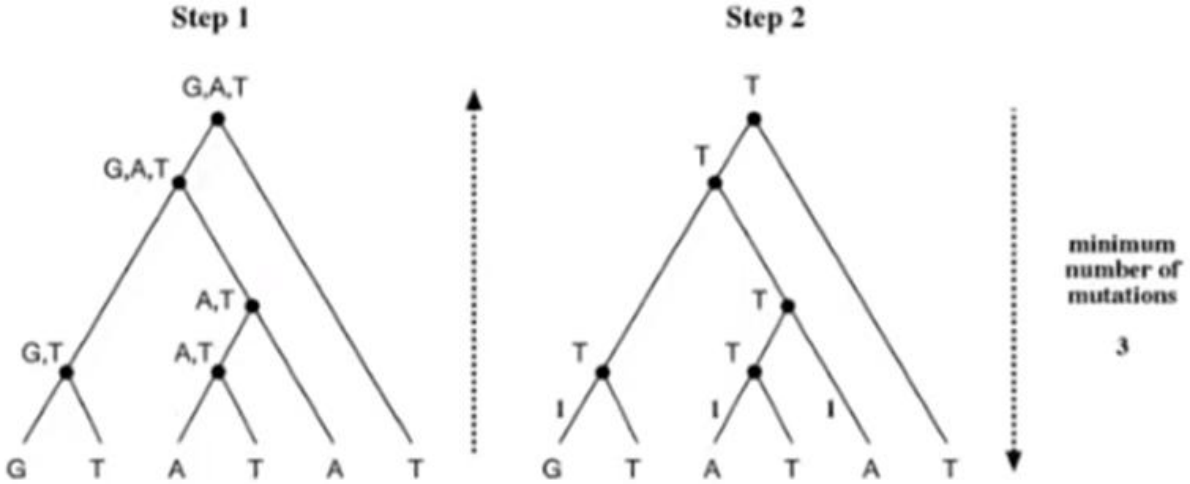
*The principle of **parsimony** dictates that a theory should provide the simplest possible (viable) explanation for a phenomenon.*

1. Student did not prepare well for the exam.
2. Course instructor did not like the student and awarded an *F*.
3. Mean textbook publisher intentionally wrote wrong tree interpretation in the textbook in hope of sabotaging the career of future bioinformaticians.

Maximum Parsimony

In phylogenetics

The best tree is the one that has fewest evolutionary changes



Maximum Parsimony

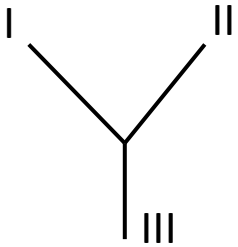
Principle

Select that tree that minimizes the total tree length and the number of nucleic acid substitutions or amino acid replacements required to explain a given set of data.

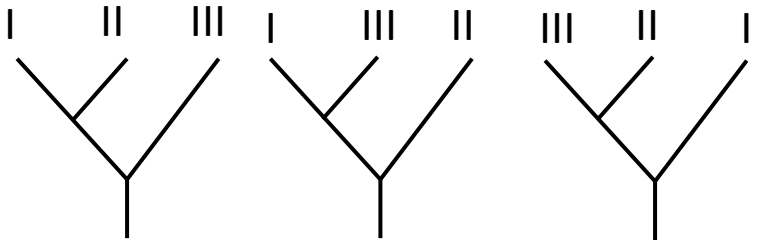
Method

- **A particular topology is considered.**
- **For this topology, the ancestral sequences at each branching point are reconstructed.**
- **The minimum number of events to explain the sequence differences over the whole tree is computed: the minimum number of substitutions is computed for each nucleotide (or amino acid) site, and the numbers for all sites are added.**
- **Another tree topology is chosen.**

Rooted and Unrooted trees



Unrooted tree



Rooted trees

# of rooted trees = $\frac{(2n-3)!}{2^{n-2}(n-2)!}$

# of unrooted trees = $\frac{(2n-5)!}{2^{n-3}(n-3)!}$

Number of possible trees

#no. of species /UTO (n)	#rooted trees	#unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
10	3.44×10^{7}	2.03×10^{6}
15	2.13×10^{14}	7.91×10^{12}
20	8.20×10^{21}	2.21×10^{20}

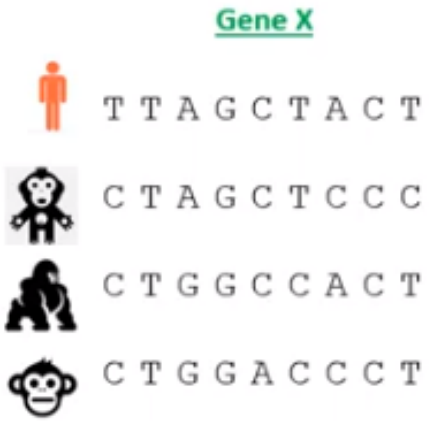
Maximum Parsimony

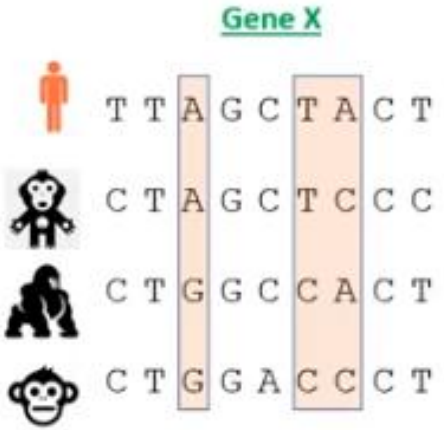
Informative sites

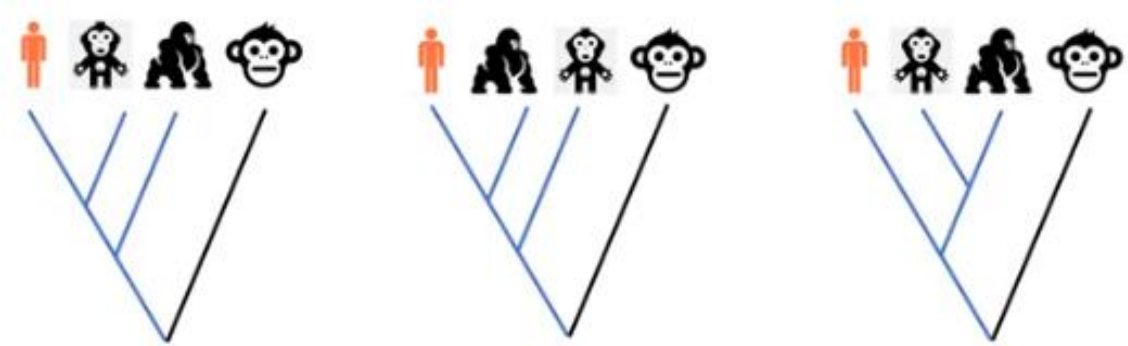
taxa \ sites	1	2	3	4	5	6	7	8
I	A	A	T	T	A	G	C	T
II	G	G	T	C	G	T	A	G
III	A	A	T	G	C	G	C	T
IV	A	G	T	A	A	G	C	A
V	A	C	T	T	C	G	C	G
VI	A	C	A	T	G	G	C	A

A site is informative only when there are at least two different kinds of nucleotides at the site, each of which is represented in at least two of the sequences under study

Maximum Parsimony

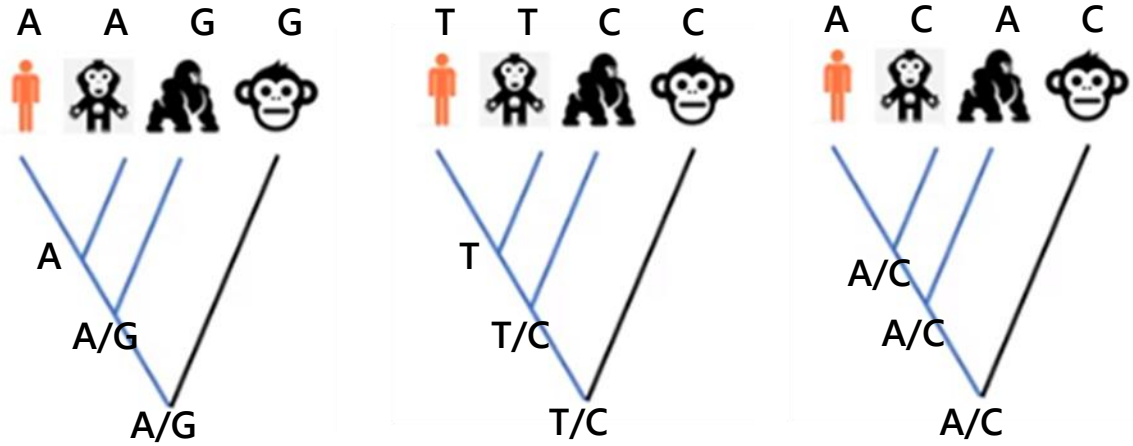


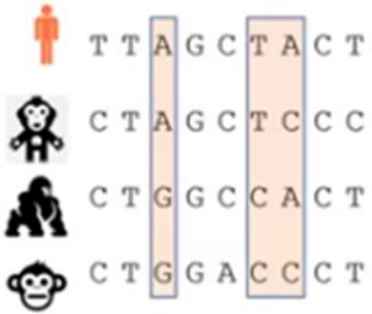




(15 rooted tree topologies are possible with four taxa)

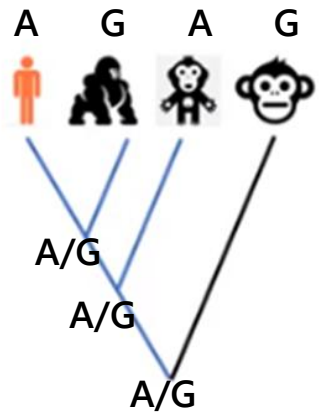
Maximum Parsimony

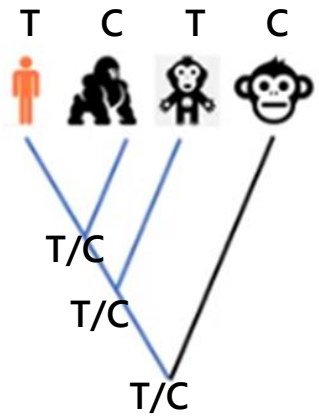


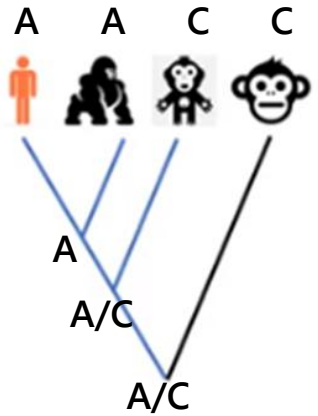


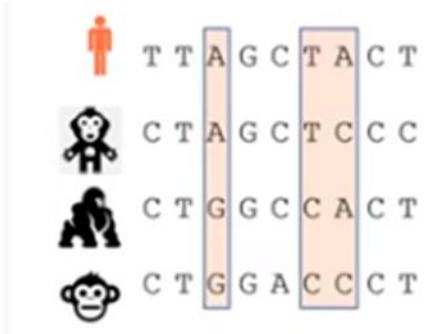
	Position 3	Position 6	Position 7	Total changes
Tree I	1	1	2	4
Tree II				
Tree III				

Maximum Parsimony



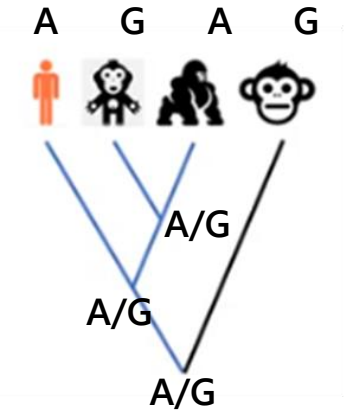


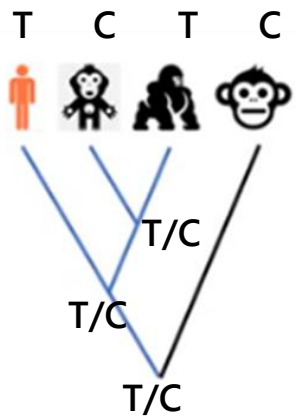


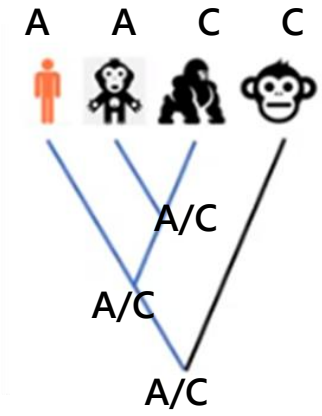


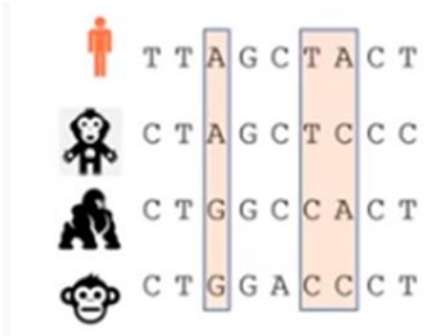
	Position 3	Position 6	Position 7	Total changes
Tree I	1	1	2	4
Tree II	2	2	1	5
Tree III				

Maximum Parsimony









	Position 3	Position 6	Position 7	Total changes
Tree I	1	1	2	4
Tree II	2	2	1	5
Tree III	2	2	2	6

Maximum Likelihood Method

Maximum Likelihood Method

- Another character-based method for the inference of phylogeny.
- It evaluates a hypothesis about evolutionary history in terms of the probability that the proposed model and the hypothesized history would give rise to the observed data set.
- The supposition is that a history with a higher probability of reaching the observed state is preferred to a history with a lower probability.
- The method searches for the tree with the highest probability or likelihood.

Maximum Likelihood Method

Advantages:

- They have often lower variance than other methods (i.e. it is frequently the estimation method least affected by sampling error).
- They tend to be robust to many violations of the assumptions in the evolutionary model.
- Even with very short sequences they tend to outperform alternative methods such as parsimony or distance methods.
- The method is statistically well founded.
- They evaluate different tree topologies while using all the sequence information.

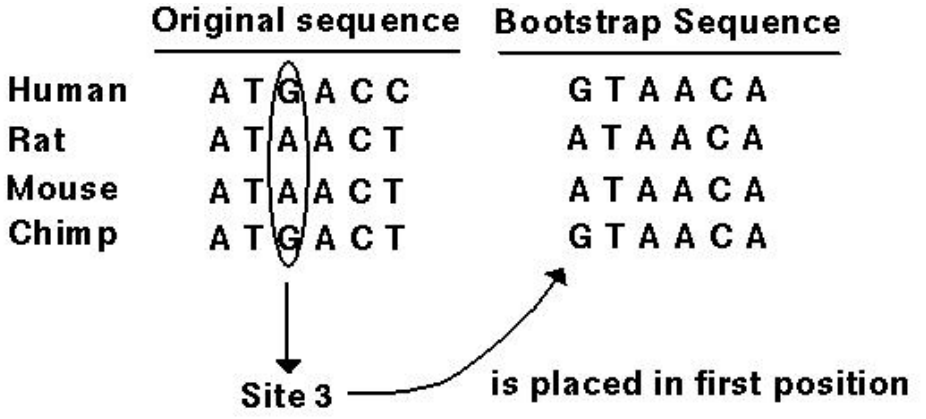
Limitations:

- Maximum likelihood is very CPU intensive and thus extremely slow.
- The result is dependent on the model of evolution used.

Evaluating the reliability of the tree

Bootstrapping

- The bootstrap technique involves generating artificial sequences by randomly sampling sites from the original sequences with replacement.
- This randomly generated data set has the same sequence length but a slightly different composition (i.e. some sites will be oversampled and others not).



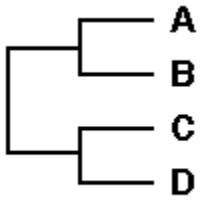
(Then the next five randomly chosen sites: 2, 1, 1, 5, 4, are placed in the next five positions.)

Evaluating the reliability of the tree

Sample 1 0 1 2 0 3 0 1 2 0 1 (<- number of times each site is sampled)

A	A G G C U C C A A A	A	G G G U U U C A A A
B	A G G U U C G A A A	B	G G G U U U G A A A
C	A G C C C C G A A A	C	G C C C C C G A A A
D	A U U U C C G A A C	D	U U U C C C G A A C

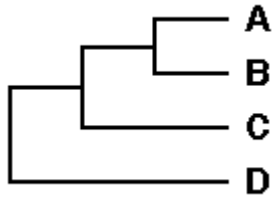
	A	B	C
B	1		
C	6	5	
D	8	7	4



Sample 2 1 0 0 0 2 2 2 0 0 3

A	A G G C U C C A A A	A	A U U C C C C A A A
B	A G G U U C G A A A	B	A U U C C G G A A A
C	A G C C C C G A A A	C	A C C C C G G A A A
D	A U U U C C G A A C	D	A C C C C G G C C C

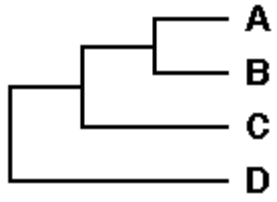
	A	B	C
B	2		
C	4	2	
D	7	5	3



Sample 3 1 0 0 0 2 2 2 0 0 3

A	A G G C U C C A A A	A	A U U C C C C A A A
B	A G G U U C G A A A	B	A U U C C G G A A A
C	A G C C C C G A A A	C	A C C C C G G A A A
D	A U U U C C G A A C	D	A C C C C G G C C C

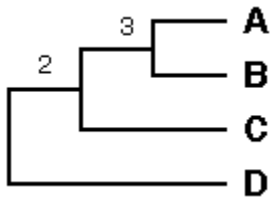
	A	B	C
B	1		
C	3	2	
D	6	3	4

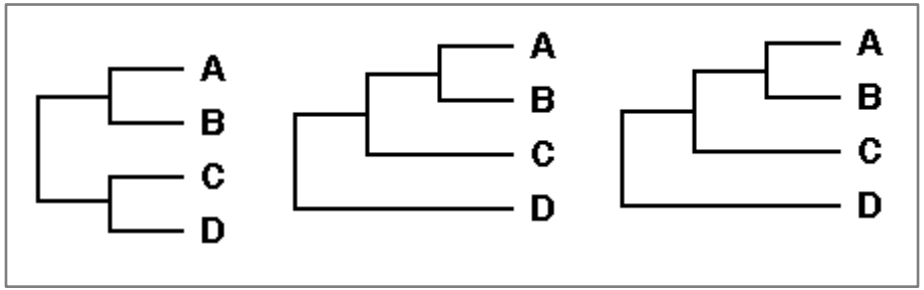


Evaluating the reliability of the tree

A	A G G C U C C A A A
B	A G G U U C G A A A
C	A G C C C C G A A A
D	A U U U C C G A A C

	A	B	C
B	2		
C	3	3	
D	6	4	4





- This whole process is repeated at least 100 times.
- The number of times that a clade is seen among the bootstrap trees is reported.
- The more often a clade is present among the bootstrap trees, the more strongly the data support that clade, because the result is insensitive to which base-pairs happen to be sampled.

Evaluating the reliability of the tree

Jackknifing

- It also used to evaluate the reliability of the generated tree.
- It is a statistical method of numerical resampling based on deleting a portion of the original observations for each pseudo-replicate.
- A 50% jackknife randomly deletes half of the columns from the alignment to create each pseudo-replicate.

Overview of phylogenetic tree construction

1. Selection of sequences for analysis
2. Multiple sequence alignment
3. Tree building
    - *Distance based methods*
        - UPGMA
        - Neighbor joining method
    - *Character based methods*
        - Maximum parsimony
        - Maximum likelihood method
4. Tree evaluation
    - Bootstrapping
    - Jackknifing