BIO 390/CSCI 390/MATH 390
Bioinformatics II
Programming Lecture 11
Phylogenetic Tree

Instructor: Lei Qian
Fisk University

Phylogenetic Tree

Applications

- Study ancestor-descendant relationships
  (Evolutionary biology, adaption, genetic drift, selection, speciation, etc.)
- Paleogenomics: inferring ancestral genomic information from extinct species
  (Comparing Chimpanzee, Neanderthal and Human DNA)
- Origins of epidemics
  (Comparing, at the molecular level, various virus strains)
- Drug design: specially targeting groups of organisms
  (Efficient enumeration of phylogenetically informative substrings)
- Forensic
  (Relationships among HIV strains)
- Linguistics
  (Languages tree divergence times)

Phylogenetic Tree

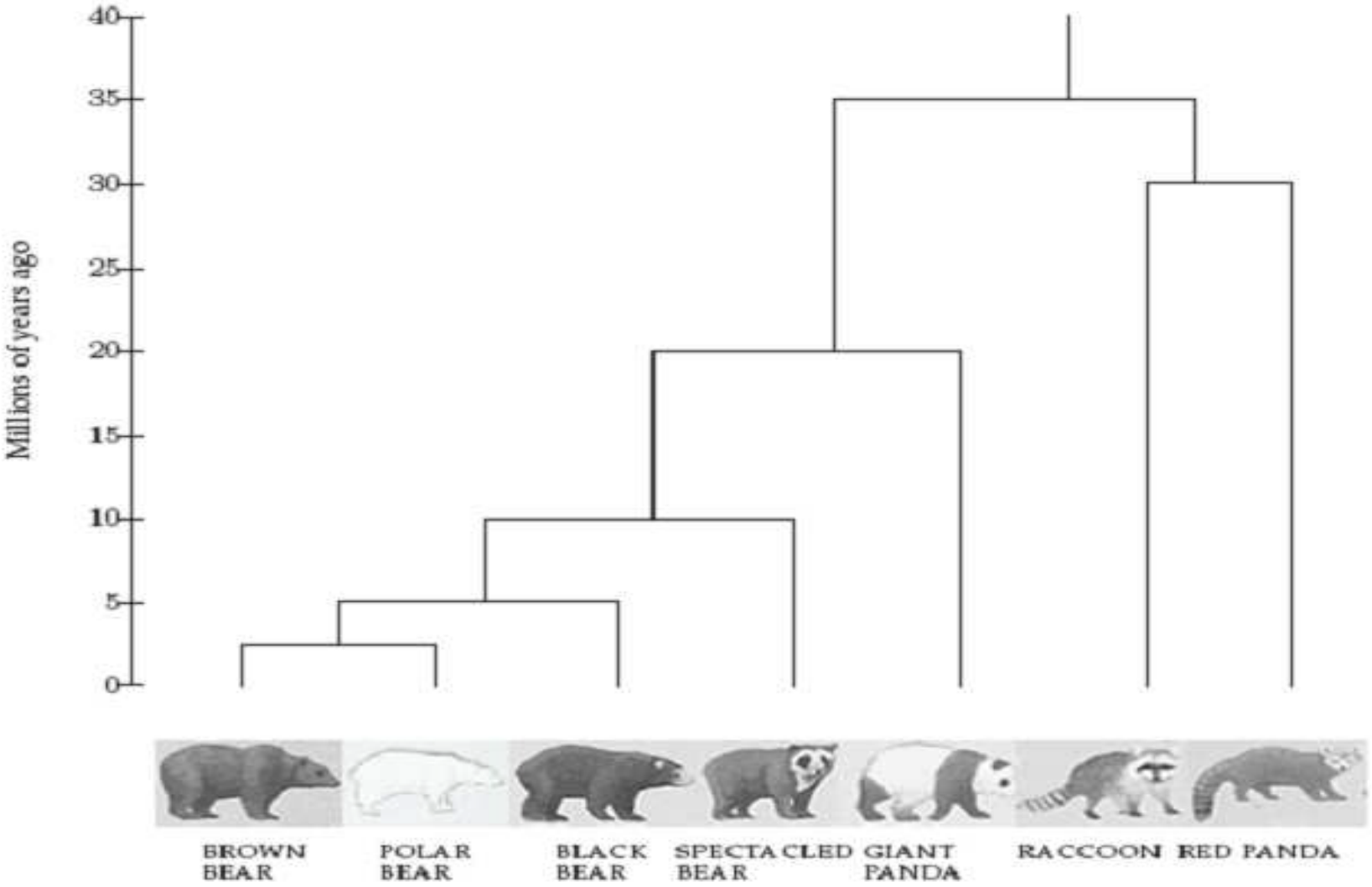
Illustrating success stories in phylogenetics (I)

For roughly 100 years (more exactly, 1870-1985), scientists were unable to figure out which family the giant panda belongs to. Giant pandas look like bears, but have features that are unusual for bears but typical to raccoons: they do not hibernate, they do not roar, their male genitalia are small and backward-pointing.

Anatomical features were the dominant criteria used to derive evolutionary relationships between species since Darwin till early 1960s. The evolutionary relationships derived from these relatively subjective observations were often inconclusive. Some of them were later proved incorrect.

In 1985, Steven O'Brien and colleagues solved the giant panda classification problem using DNA sequences and phylogenetic algorithms.

Phylogenetic Tree



Phylogenetic Tree

Illustrating success stories in phylogenetics (II)

In 1994, a woman from Lafayette, Louisiana (USA), clamed that her ex-lover (who was a physician) injected her with HIV+ blood.

Records showed that the physician had drawn blood from a HIV+ patient that day.

But how to prove that the blood from that HIV+ patient ended up in the woman?

Phylogenetic Tree

HIV has a high mutation rate, which can be used to trace paths of transmission.

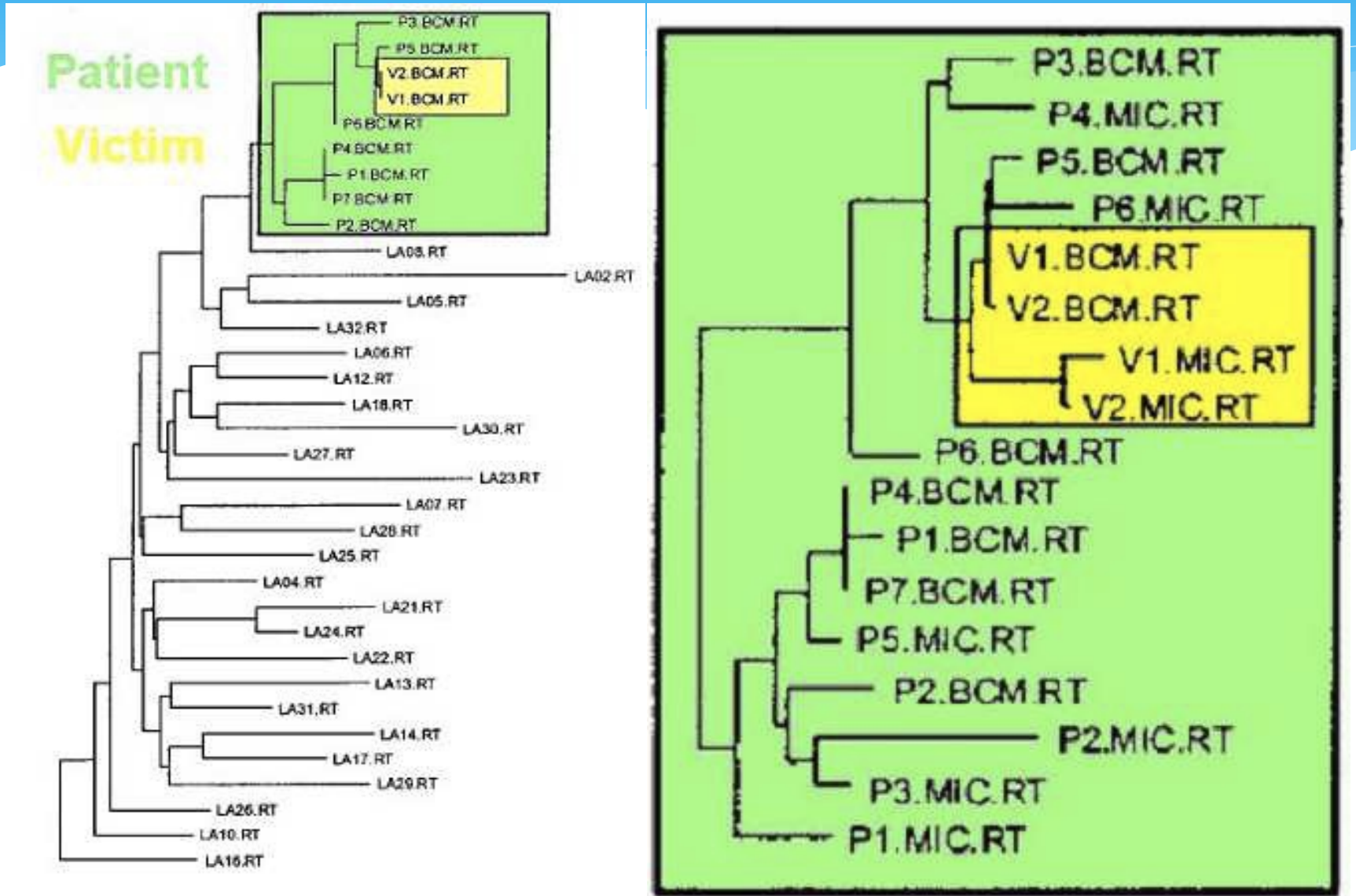
Two people who got the virus from two different people will have very different HIV sequences.

Three different phylogenetic trees (including parsimony-based) were used to track changes in two genes in HIV (gp120 and RT). Multiple samples from the physician's patient, the woman and controls (non-related HIV+ people) were used.

In every reconstruction, the woman's sequences were found to be evolved from the patient's sequences.

This was the first time when phylogenetic analysis was used in court as evidence.

Phylogenetic Tree



Phylogenetic Tree

Deriving Phylogenetic Trees

Aim:

Given a set of data (DNA, protein sequences, protein structure, etc.) that characterize different groups of organisms, try to derive information about the relationships among the organisms in which they were observed.

The distance-based ("phenetic") approach:

Proceed by measuring a set of distances between (data provided for these) species, and generate the tree by a hierarchical clustering procedure.

Note: Hierarchical clustering is perfectly capable of producing a tree even in the absence of evolutionary relationships!

The character-based ("cladistic") approach:

Consider possible pathways of evolution, infer the features of the ancestor at each node, and choose an optimal tree according to some model of evolutionary change (maximum parsimony, maximum likelihood, or based on genealogy or homology).

Phylogenetic Tree

Distance-based Phylogeny

These most intuitive methods of building phylogenetic trees begin with a set of distances d_{ij} between each pair (x_i, x_j) of sequences in the given dataset.

There are many ways of defining a distance.

- Hamming Distance
- Levenshtein Distance
- Scores based on substitution matrices such as PAM and BLOSUM
- Jukes-Cantor distance: $d_{ij} = -\frac{3}{4}\log(1 - f \times \frac{4}{3})$, where f is the fraction of differences between sequences x_i and x_j.

Character-based Phylogeny

A character is a measurable feature having well defined mutually exclusive states. Discrete measurement.

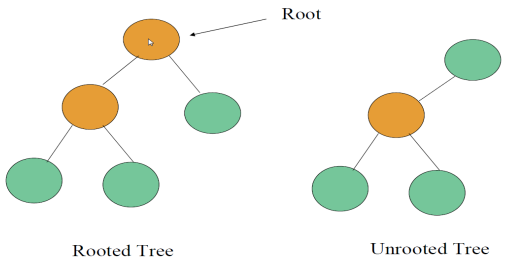
Phylogenetic Tree

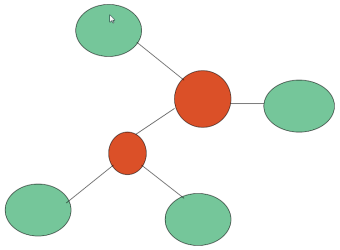
Rooted and Unrooted Trees in Phylogenetic

- **Graph:** (V, E) where V is a set of vertices (nodes) and E is a set of edges between vertices.
- **Trees** Connected graph without cycles.
- **Degree of a node:** The number of edges that touch this node.
- **Leaf:** A node with degree 1.
- **Inner node:** A node with degree > 1.
- **Rooted tree:** A tree has a unique node that has no parent. All other nodes must have a parent. It is directional (toward to or away from the root)
- **Binary rooted tree:** Every inner node has exactly two children.

A rooted tree has three types of nodes: leaf – degree 1, root – degree 2 (unique), other inner nodes – degree 3.

- **Binary unrooted tree:** Every vertex in the tree is of degree 1 (leaf) or degree 3 (inner node). It is unidirectional.





Phylogenetic Tree

Rooted vs. Unrooted Tree

- Root - ancestor of all taxa considered
- Unrooted tree - typical result, unknown common ancestor
- Rooted tree - known common ancestor
- Specify root by means of outgroup
- Outgroup is distant from all other taxa
    - example: mammals and a salamander
    - ancestor of outgroup is presumed root

Phylogenetic Tree

Biologists generally prefer rooted trees.

- Under the molecular clock assumption, the root of the tree would be located at equal distance from all the leaves (contemporary organisms);

- The outgroup method consists of including into the analysis an organism that is known to have branched o earlier than the taxa under study (for which paleontological evidences exist, for instance), the root will be placed along the edge connecting the outgroup to the ancestor of the ingroup (taxa under study).

Phylogenetic Tree

Molecular clock theory

- Proposed by Emile Zuckerkandl and Linus Pauling, 1962.
- Accepted mutations occur at a constant rate.
- The number of accepted mutations is proportional to the length of the time interval.
- Once the "clock" has been calibrated (using fossil evidences, for instance) the unknown length of some time interval can be deduced from the number of accepted mutations.
- Note: different proteins have different clocks (hemoglobin ticks faster than cytochrome c).

Phylogenetic Tree

Maximum Parsimony Method

- Predict the evolutionary tree that minimizes the number of steps required to generate the observed variation in the sequences.
- Find a tree that explains data with a minimal number of changes.
- Appropriate for very similar sequences and a small number of sequences
- Time Consuming (try to examine all possible trees). Need exponential time.
- PHYLIP and PAUP offer maximum parsimony method

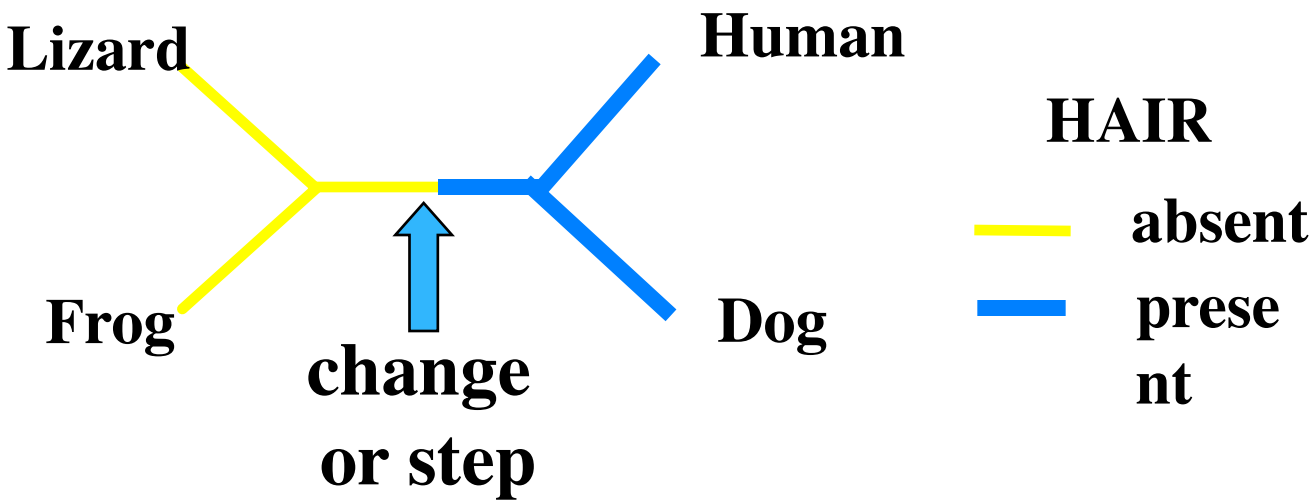
Phylogenetic Tree

Total number of unrooted trees:

	Unrooted	rooted
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425

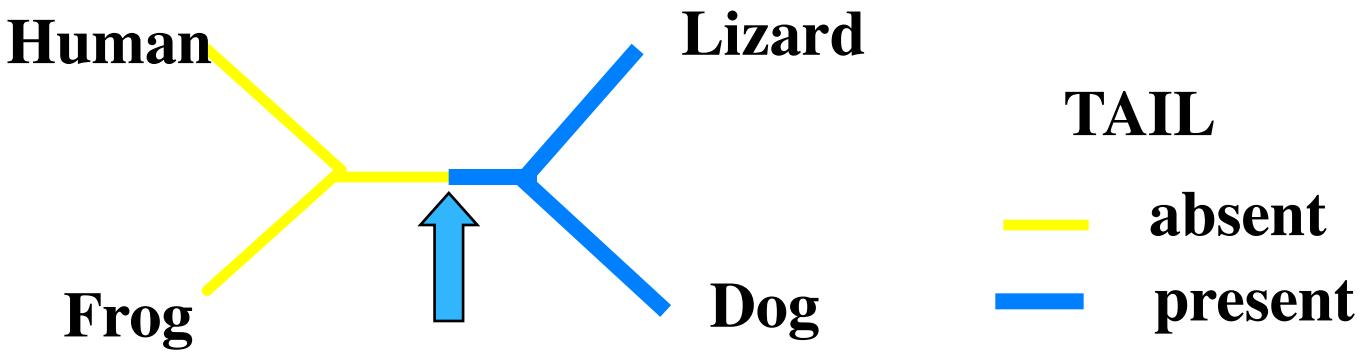
Unique and unreversed characters (apomorphy)

* Because hair evolved only once and is unreversed (not subsequently lost) it is *homologous* and provides unambiguous evidence of relationships



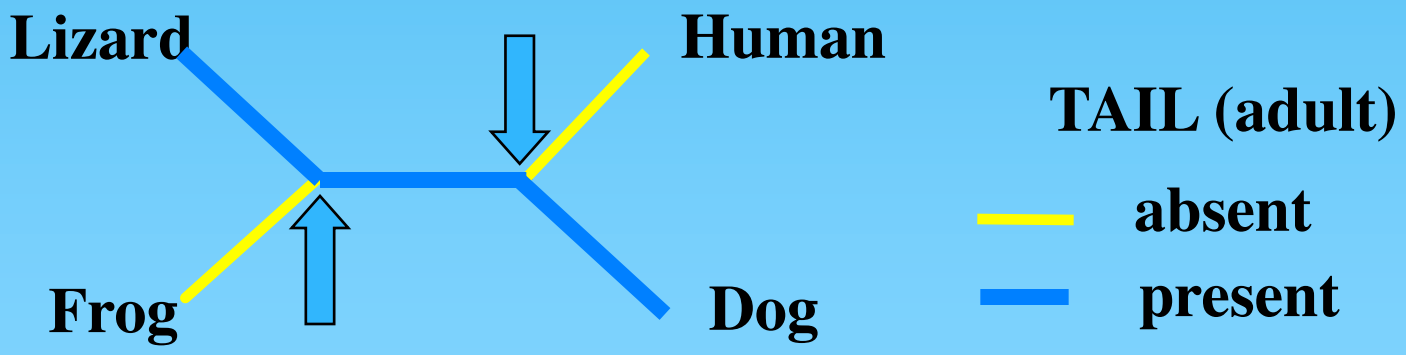
Homoplasy - misleading evidence of phylogeny

* If misinterpreted as homology, the absence of tails would be evidence for a wrong tree: grouping humans with frogs and lizards with dogs

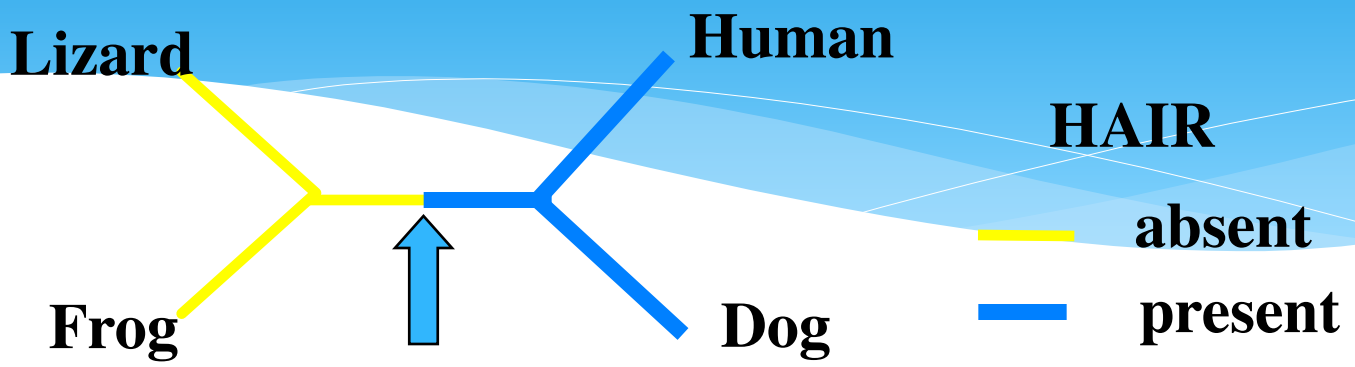


Homoplasy - independent evolution

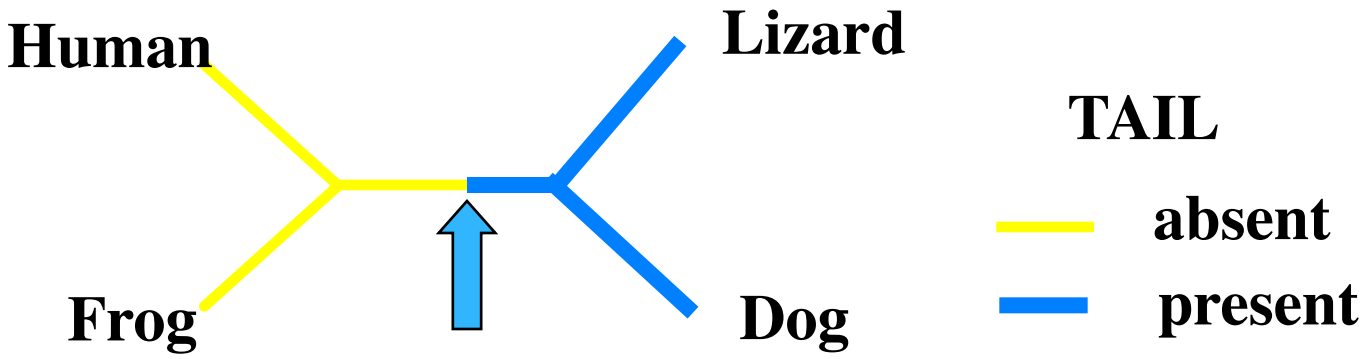
- **Loss of tails evolved independently in humans and frogs - there are two steps on the true tree**



Incongruence or Incompatibility

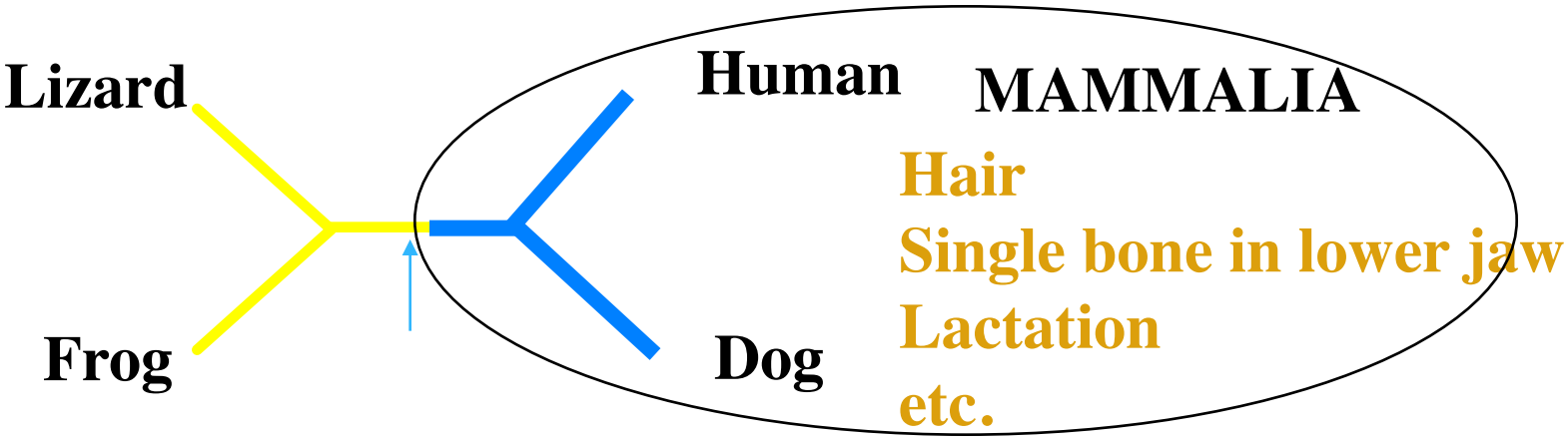


* These trees and characters are incongruent - both trees cannot be correct, at least one is wrong and at least one character must be homoplastic



Congruence

* We prefer the ‘true’ tree because it is supported by multiple congruent characters

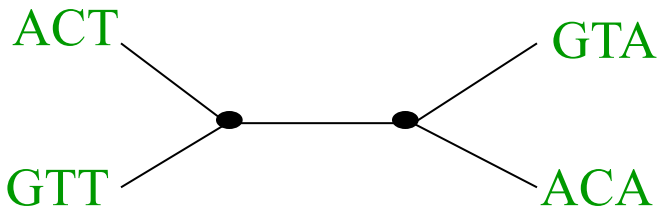


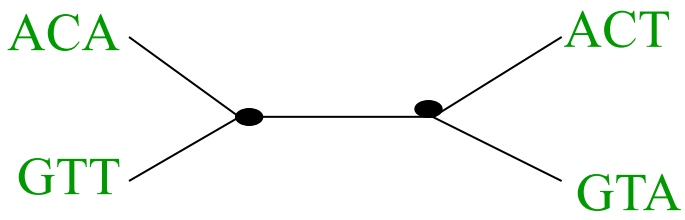
Maximum parsimony (example)

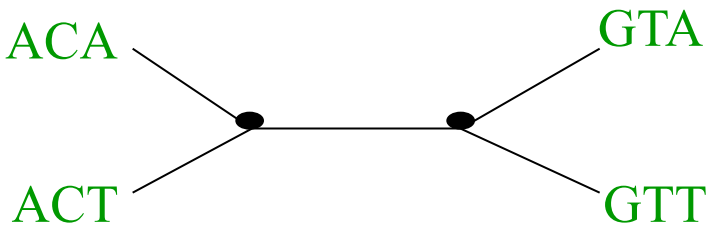
* **Input**: Four sequences
  * ACT
  * ACA
  * GTT
  * GTA
* There are 3 possible unrooted trees with 4 nodes.
* **Question**: which of the three trees has the best scores?

Maximum Parsimony

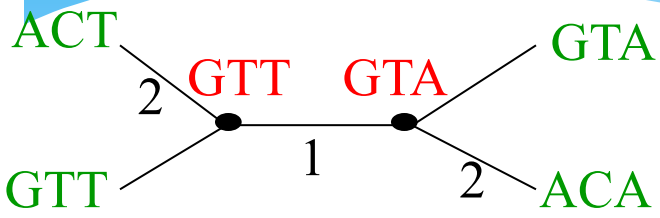
There are three possible unrooted trees



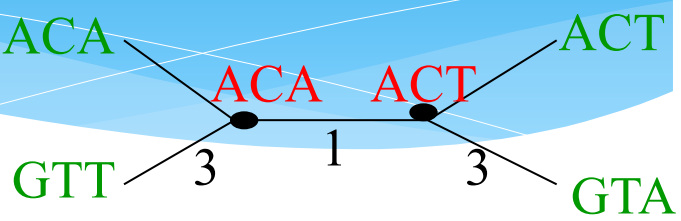




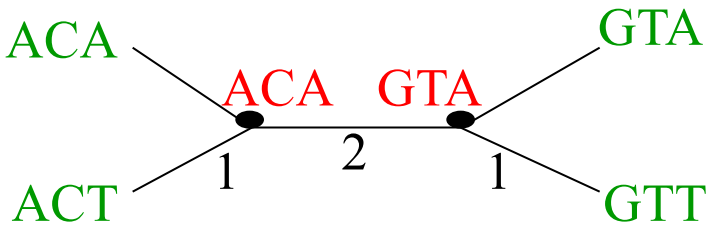
Maximum Parsimony



score = 5



score = 7

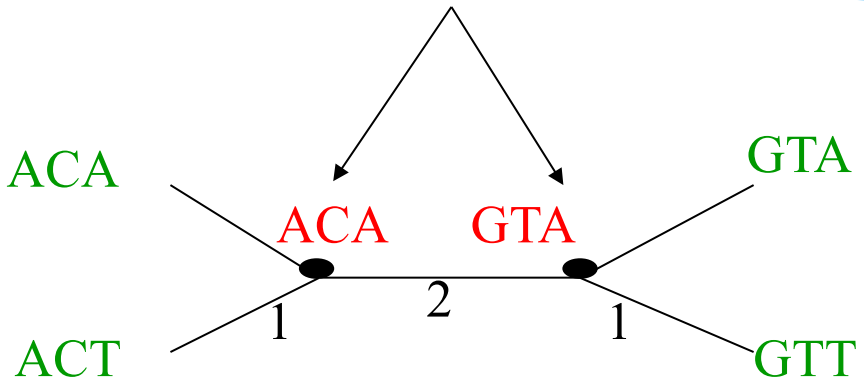


score = 4

Optimal MP tree

Maximum Parsimony: computational complexity

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



score = 4

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

Maximum Parsimony Algorithm

Step 1: Find informative sites for Parsimony Analysis.

A nucleotide site is informative only if it favors a subset of trees over the other possible trees.

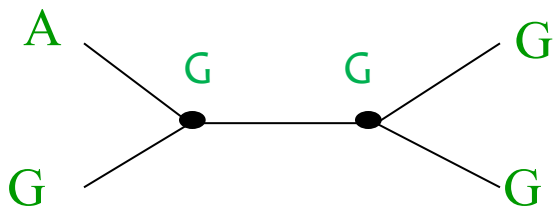
- An example, four OTUs (operational taxonomic units), nine sites

	1	2	3	4	5	6	7	8	9
OTU a	A	A	G	A	G	T	T	C	A
OTU b	A	G	C	C	G	T	T	C	T
OTU c	A	G	A	T	A	T	C	C	A
OTU d	A	G	A	G	A	T	C	C	T

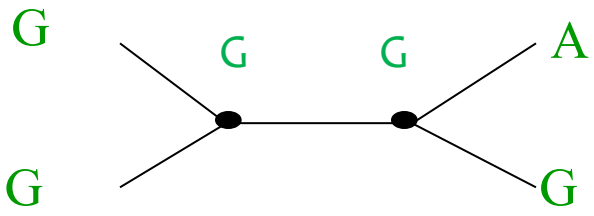
What sites are informative?

Maximum Parsimony Algorithm

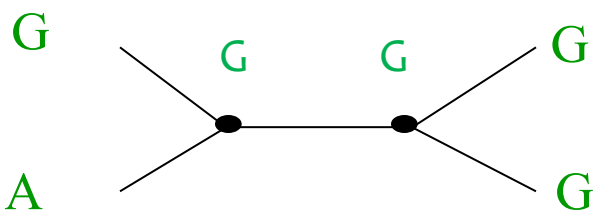
Site 2: A G G G



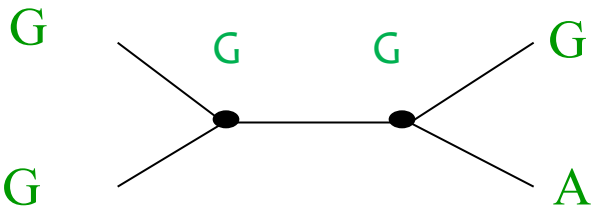
score = 1



score = 1



score = 1

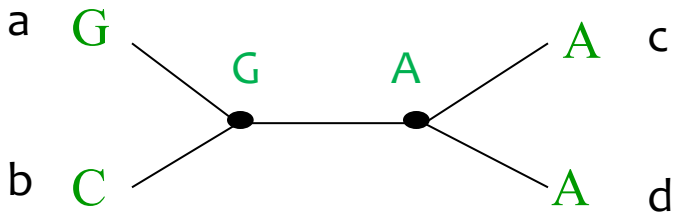


score = 1

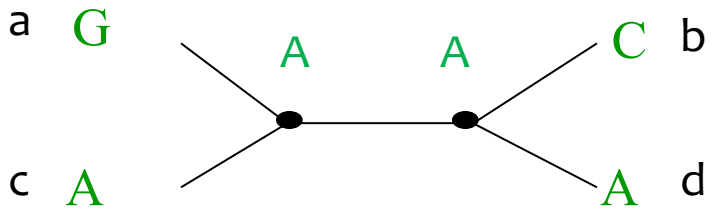
Not Informative!

Maximum Parsimony Algorithm

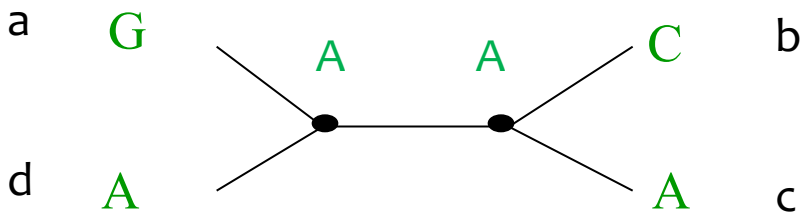
Site 3: G C A A



score = 2



score = 2

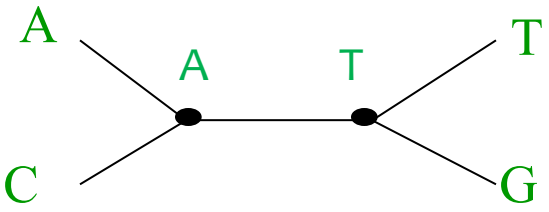


score = 2

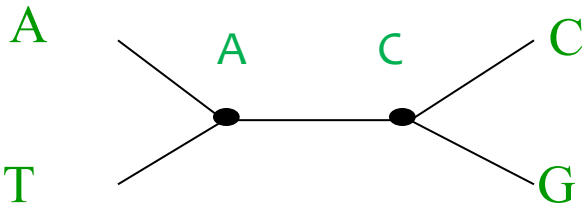
Not Informative!

Maximum Parsimony Algorithm

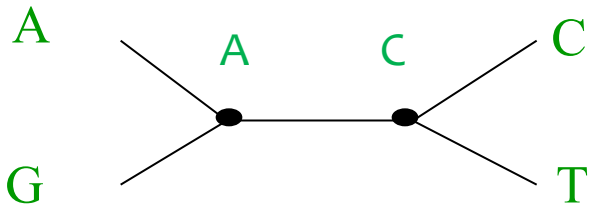
Site 4: A C T G



score = 3



score = 3



score = 3

Not Informative!

Maximum Parsimony Algorithm

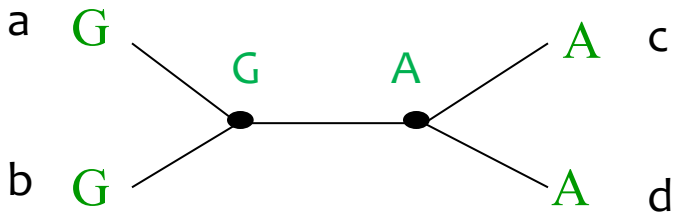
Sites 1, 2, 3, 4, 6, 8 are non-informative
Sites 5, 7 and 9 are formative

- An example, four OTUs (operational taxonomic units), nine sites

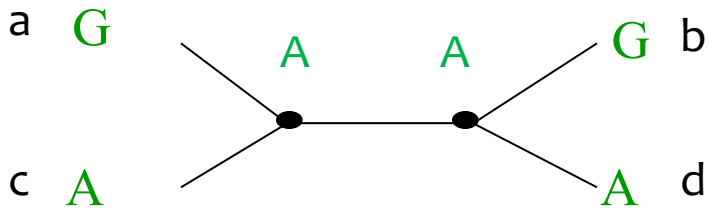
	1	2	3	4	5	6	7	8	9
OTU a	A	A	G	A	G	T	T	C	A
OTU b	A	G	C	C	G	T	T	C	T
OTU c	A	G	A	T	A	T	C	C	A
OTU d	A	G	A	G	A	T	C	C	T

Maximum Parsimony Algorithm

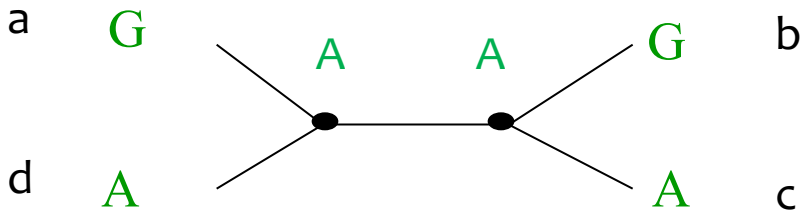
Site 5: G G A A



score = 1



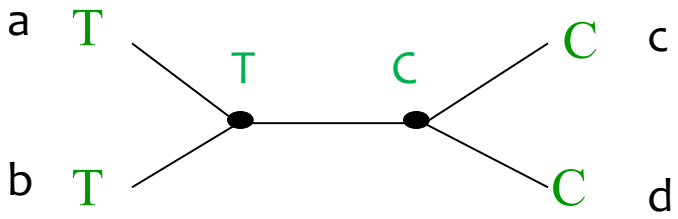
score = 2



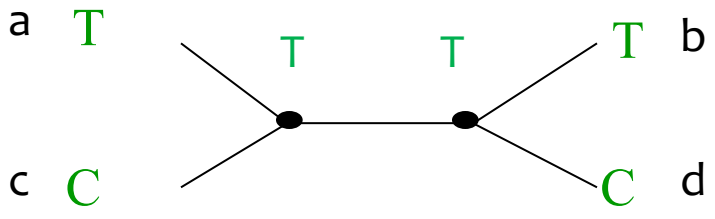
score = 2

Maximum Parsimony Algorithm

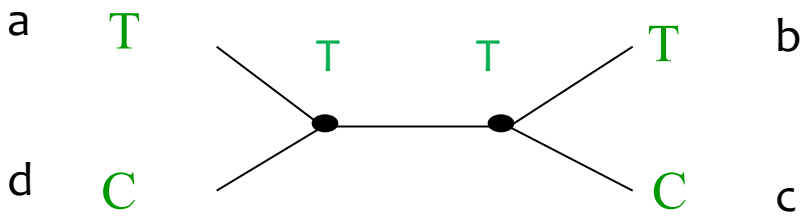
Site 7: T T C C



score = 1



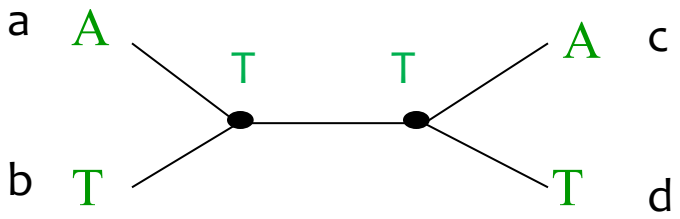
score = 2



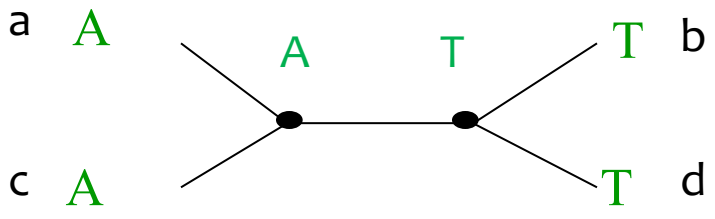
score = 2

Maximum Parsimony Algorithm

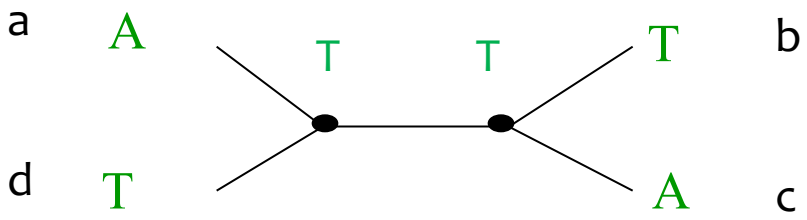
Site 9: A T A T



score = 2



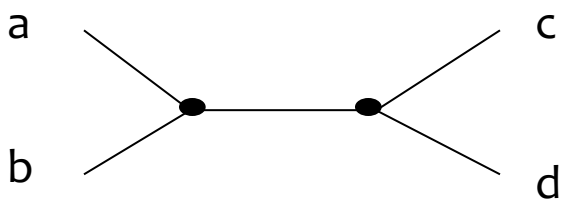
score = 1



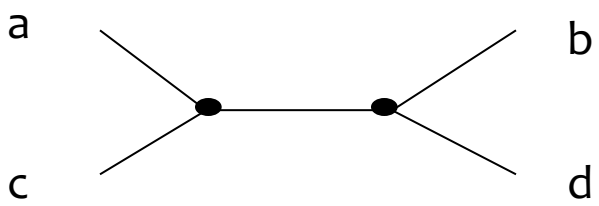
score = 2

Maximum Parsimony Algorithm

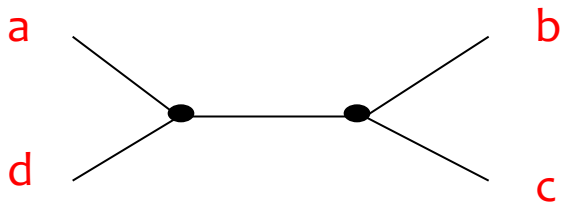
((a, d), (b,c)) is the best tree!



score = 1+1+2 = 4



score = 2+1+2=5



score = 2+2+2=6

Maximum Parsimony Algorithm

UIUC TeachEnG Maximum Parsimony Algorithm Game

http://teacheng.illinois.edu/PhylogeneticTree