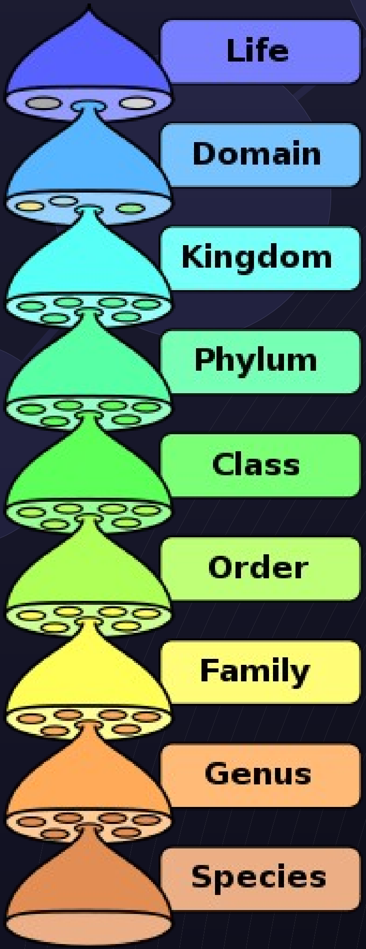
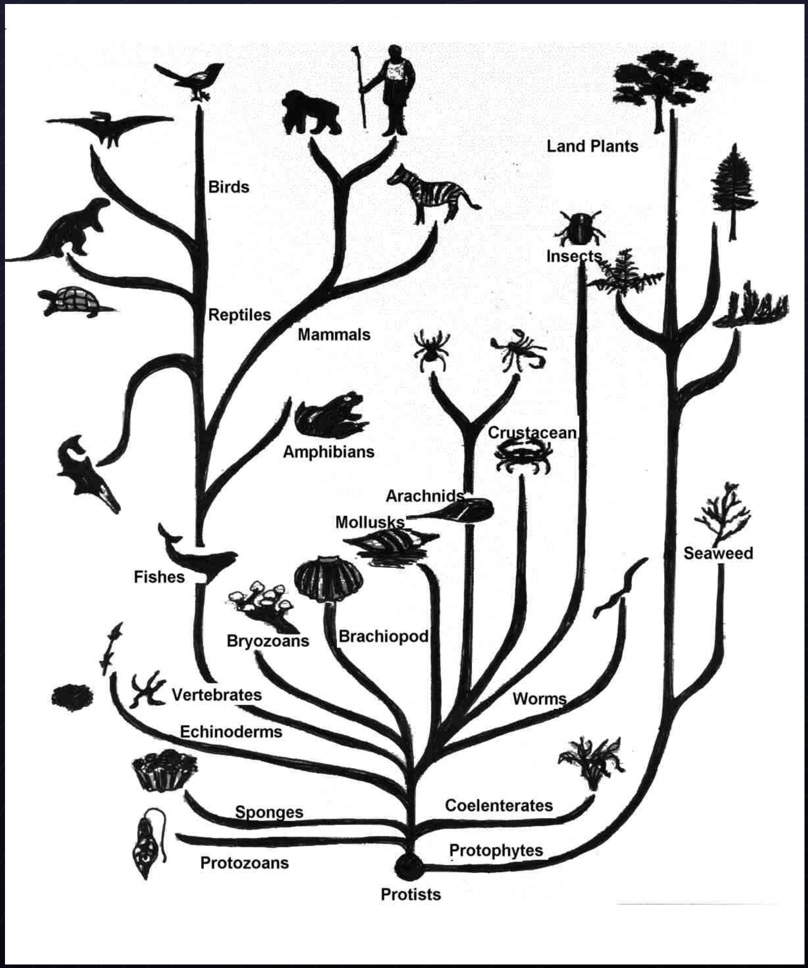
Phylogenetics

Evolution hidden behind your data

Csaba Ortutay, PhD
HiDucator Ltd.
19. October, 2015

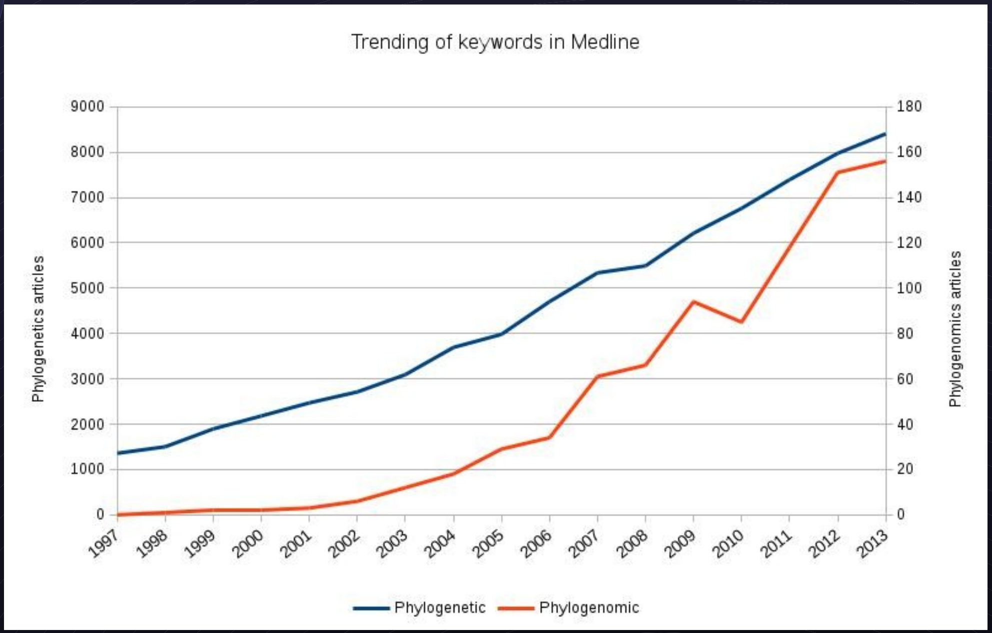
What is phylogenetics?





Why to learn phylogenetics?

- Prediction of gene function
- Genome rearrangements
- Genome assembly from NGS data
- Protein families
- Gene order, synteny
- Phylogenomics

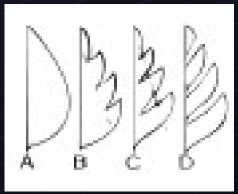


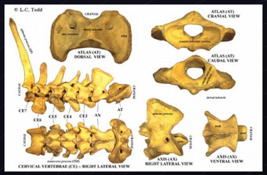
A typical phylogenetic analysis

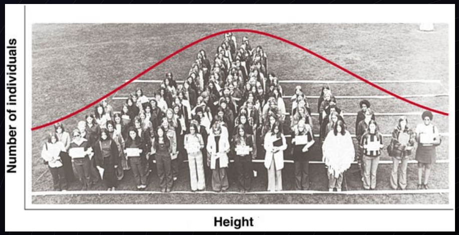
- Data selection
    - Sequences and/or other data?
    - Gene tree or species tree?
- Aligning sequences
    - Use supporting information
    - Good alignment is a necessary but not sufficient precondition
- Generating trees
    - Topology, branch lengths, consensus trees
    - Bootstrapping, tree scoring
- Answer scientific question

Data for pylogenetics

- Categorical
  - Binary
  - Un-ordered
  - Ordinal
- Numerical
  - Discrete
  - Continuous



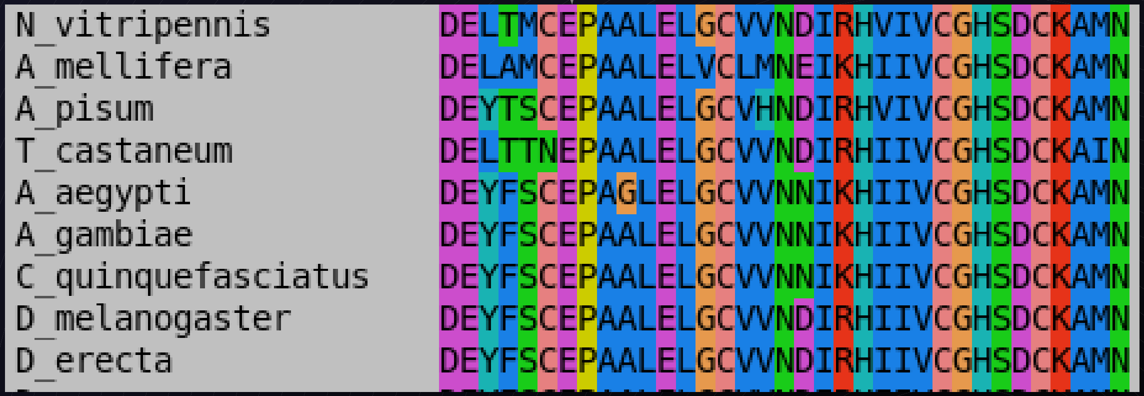




A special un-ordered categorical data type: sequences

Characters

- **Positions**
- **Columns**
- **Sites**



```
N_vitripennis          DELTMCEPAALELGCVVNDIRHVIVCGHSDCKAMN
A_mellifera            DELAMCEPAALELVCLMNEIKHIIVCGHSDCKAMN
A_pisum                DEYTSCEPAALELGCVHNDIRHVIVCGHSDCKAMN
T_castaneum            DELTTNEPAALELGCVVNDIRHIIVCGHSDCKAIN
A_aegypti              DEYFSCEPAGLELGCVVNNIKHIIVCGHSDCKAMN
A_gambiae              DEYFSCEPAALELGCVVNNIKHIIVCGHSDCKAMN
C_quinquefasciatus     DEYFSCEPAALELGCVVNNIKHIIVCGHSDCKAMN
D_melanogaster         DEYFSCEPAALELGCVVNDIRHIIVCGHSDCKAMN
D_erecta               DEYFSCEPAALELGCVVNDIRHIIVCGHSDCKAMN
```

Taxa

- **Species**
- **Sequences**

Protein or nucleotide sequences?

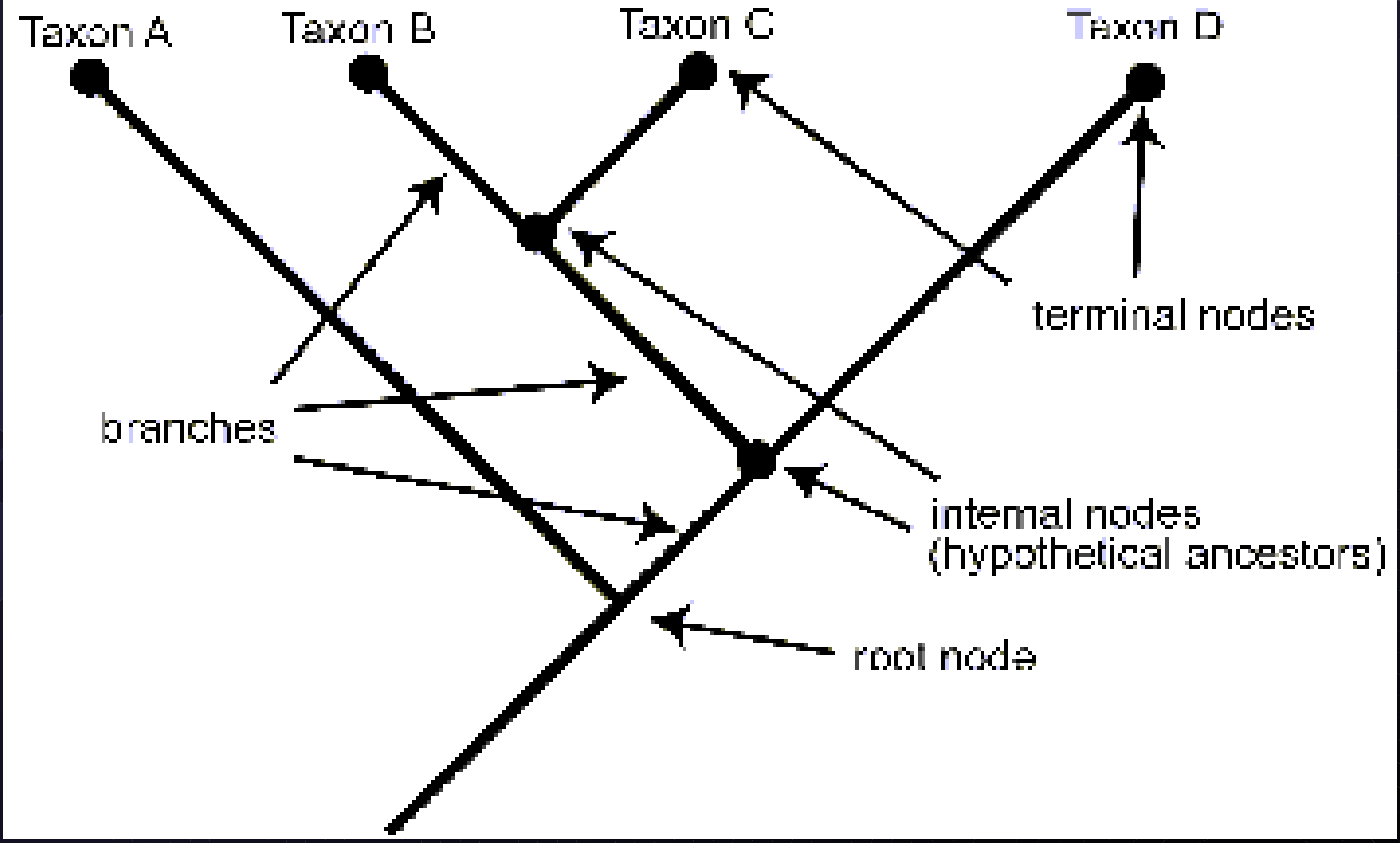
Nucleotide

- Mutations happen on this level
- Some methods have theoretical background only for nucleotide MSA
  - Nucleotide substitution models
- Some methods works only with nucleotide MSA
  - Maximum likelihood (mostly)

Protein

- Easier to align
- Some distance matrices developed for proteins
- Parsimony methods developed for proteins
- In case of frame shifts, homology is not meaningful
- Loosing info of same-sense mutations

Terminology of trees



Algorithms for generating trees

Parsimony

Simple method

Works for any data type

Distance based methods

Clustering method from statistics

- UPGMA
- Neighbor joining

Model based phylogenetics

Nucleotide substitution models

Sophisticated statistics

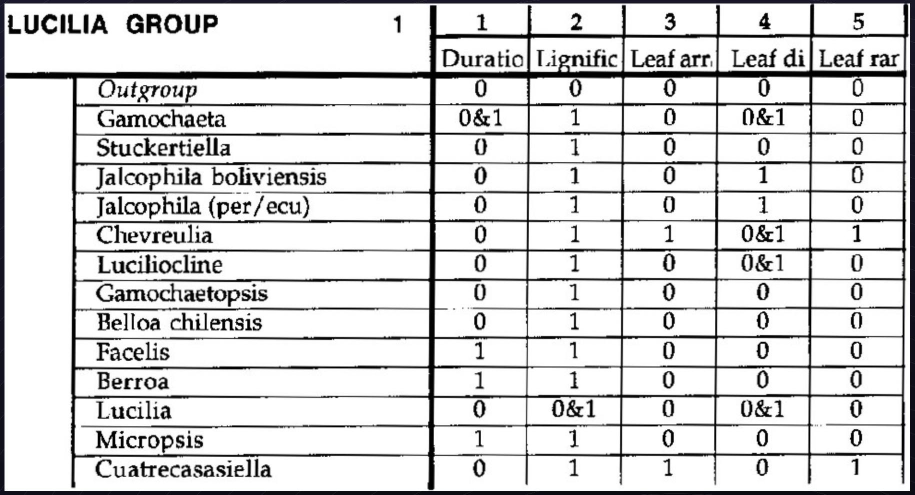
- Maximum likelihood
- Bayesian inference

Parsimony

- Phylogenetic method close to numerical taxonomy
- Directly deals with the characters and their states
- One of the most popular methods
- Easy to understand

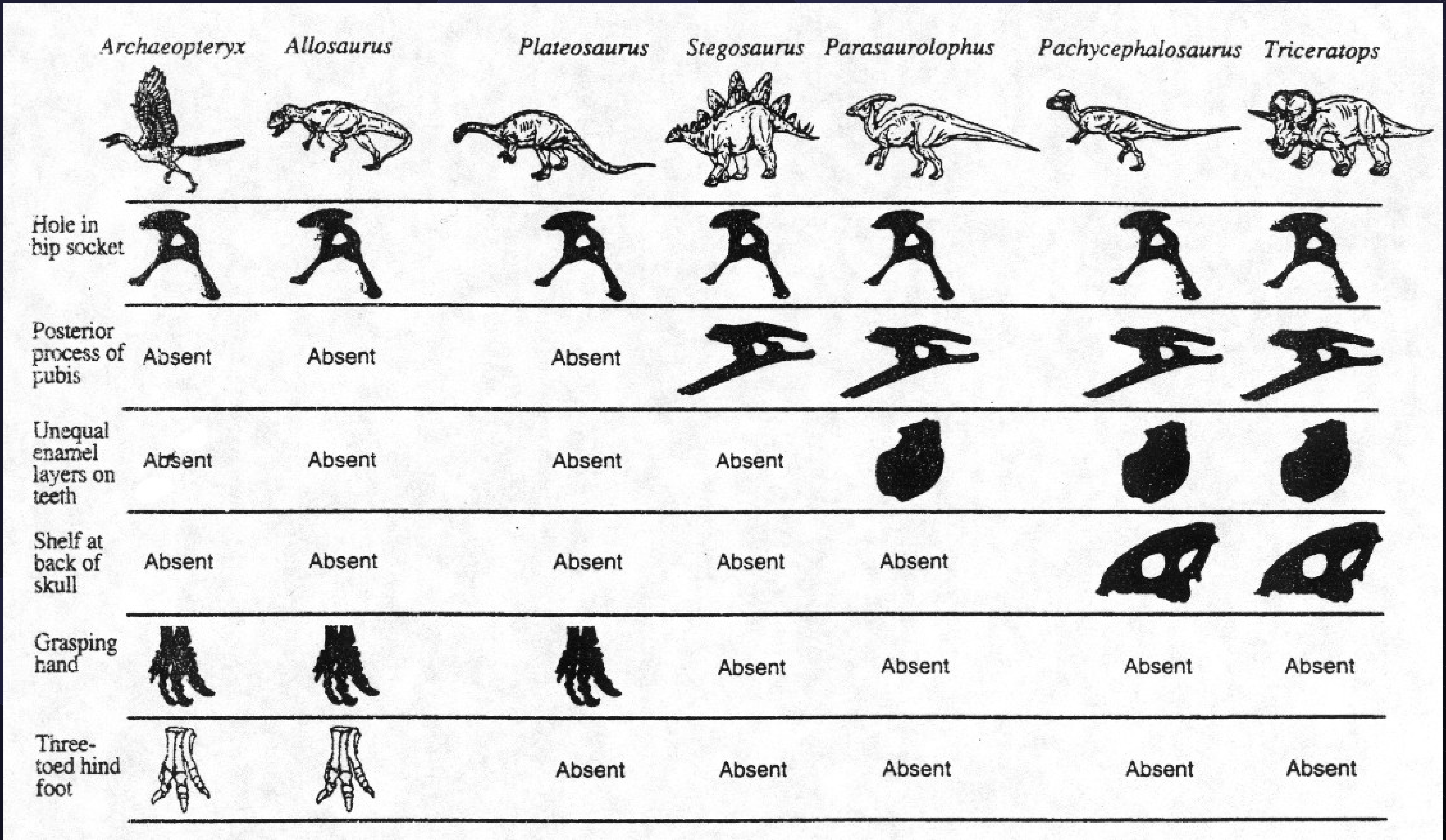
Character states

- Characters: set of homologous features
- Character state: manifestation of feature
  - Coded into table
  - Mostly categorical data



LUCILIA GROUP 1	1	2	3	4	5
	Duratio	Lignific	Leaf arr	Leaf di	Leaf rar
Outgroup	0	0	0	0	0
Gamochaeta	0&1	1	0	0&1	0
Stuckertiella	0	1	0	0	0
Jalcophila boliviensis	0	1	0	1	0
Jalcophila (per/ecu)	0	1	0	1	0
Chevreulia	0	1	1	0&1	1
Luciliocline	0	1	0	0&1	0
Gamochaetopsis	0	1	0	0	0
Belloa chilensis	0	1	0	0	0
Facelis	1	1	0	0	0
Berroa	1	1	0	0	0
Lucilia	0	0&1	0	0&1	0
Micropsis	1	1	0	0	0
Cuatrecasasiella	0	1	1	0	1

Dinosaur characters



	Archaeopteryx	*Allosaurus*	*Plateosaurus*	*Stegosaurus*	*Parasaurolophus*	*Pachycephalosaurus*	*Triceratops*
Hole in hip socket							
Posterior process of pubis	Absent	Absent	Absent				
Unequal enamel layers on teeth	Absent	Absent	Absent	Absent			
Shelf at back of skull	Absent	Absent	Absent	Absent	Absent		
Grasping hand				Absent	Absent	Absent	Absent
Three-toed hind foot			Absent	Absent	Absent	Absent	Absent

Dinosaur characters

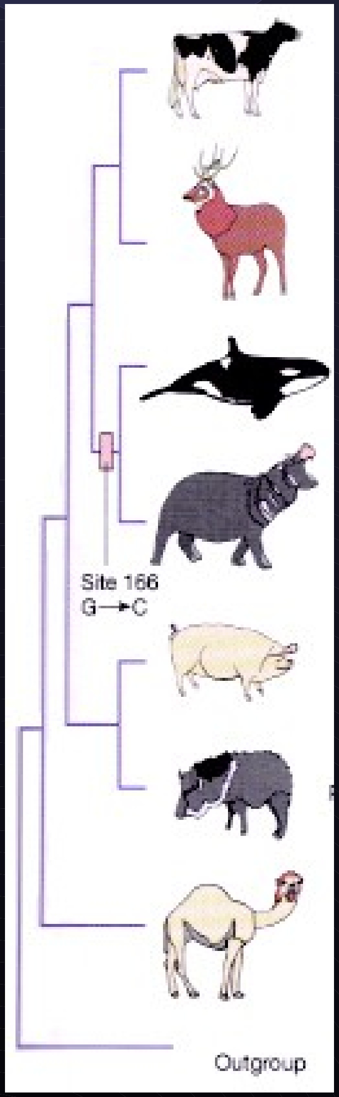
	Archeopteryx	Allosaurus	Plateosaurus	Stegosaurus	Parasaurolophus	Pachy-cephalosaurus	Triceratops
Hip hole	1	1	1	1	1	1	1
Posterior process	0	0	0	1	1	1	1
Unequal teeth layer	0	0	0	0	1	1	1
Skull shelf	0	0	0	0	0	1	1
Grasping hand	1	1	1	0	0	0	0
3 toed foot	1	1	0	0	0	0	0
Body length	0.3 small (1)	8.5 big (3)	9 big (3)	9 big (3)	9 big (3)	4.5 middle (2)	9 big (3)

Unique and unreversed characters

- Best characters
  - Unique – cannot arisen multiple times
  - Unreversed – cannot easily mutate back to the original state
  - E.g.: fur as body cover, grasping hand
- Its presence is a sign of common ancestry
- Its absence means a divergence before the emergence of the character

Homoplasy

- Similarity that is not homologous:
  - Independent evolution (analogy, convergence, parallelism)
  - Reversal (back mutation)
- We cannot deduce the relationship of taxa having homoplasy



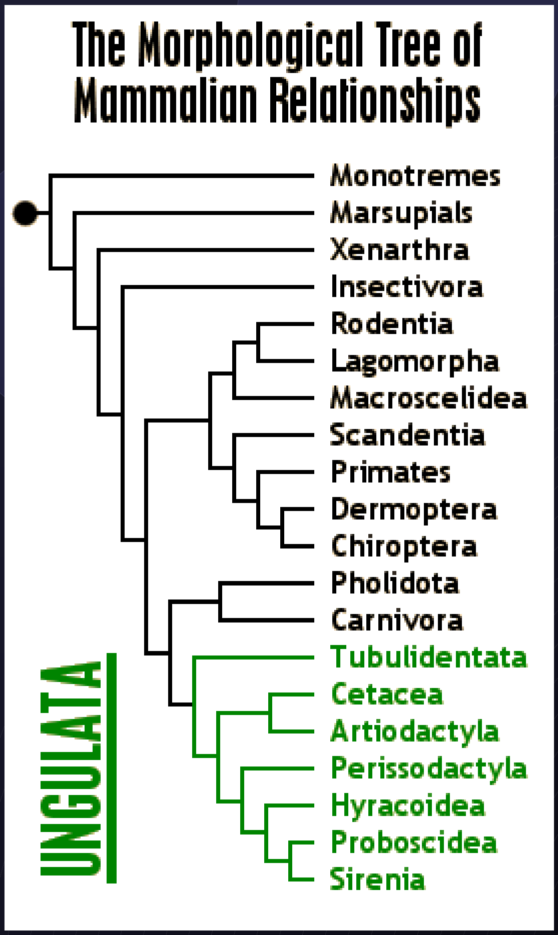
© Brandon Cole / www.brandoncole.com

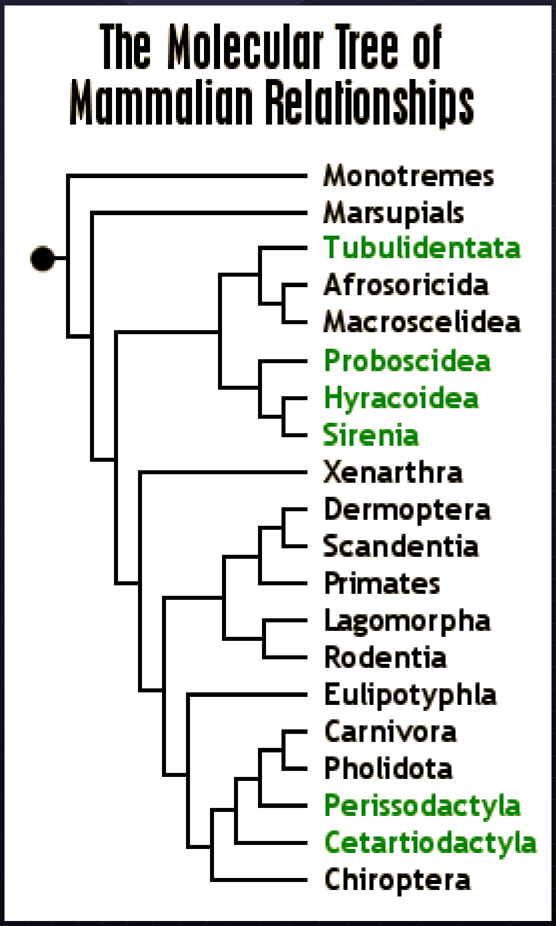
PHOTOGRAPHERS DIRECT.COM

Homoplasy and Incongruence

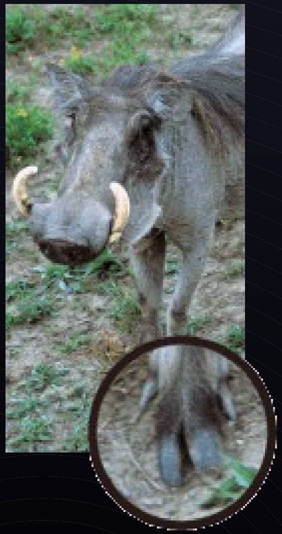
EQUIDAE
Equus grevyi

BOVIDAE
Bison bonasus



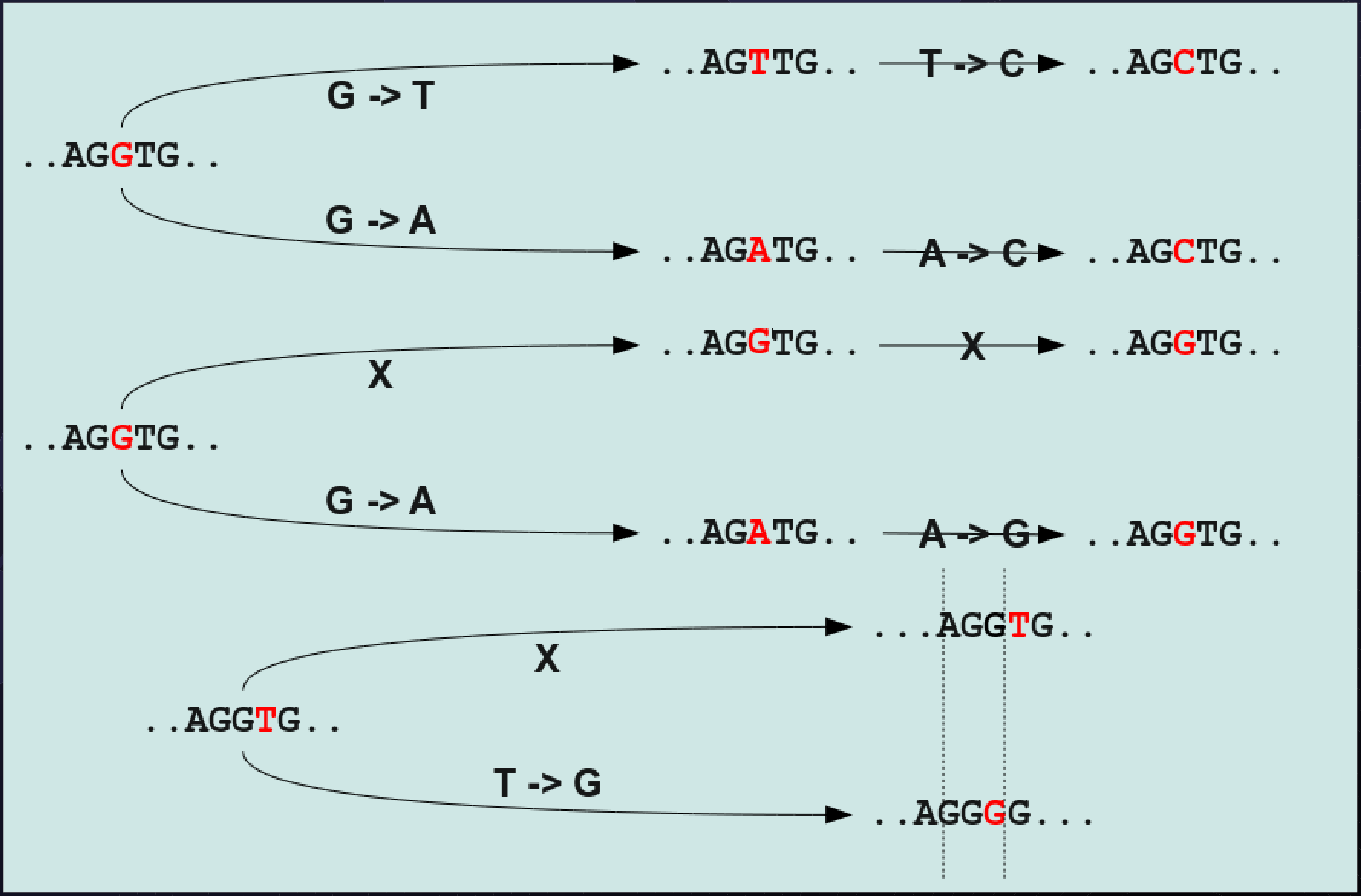


RHINOCEROTIDAE
Ceratotherium simum



SUIDAE
Phacochoerus africanus

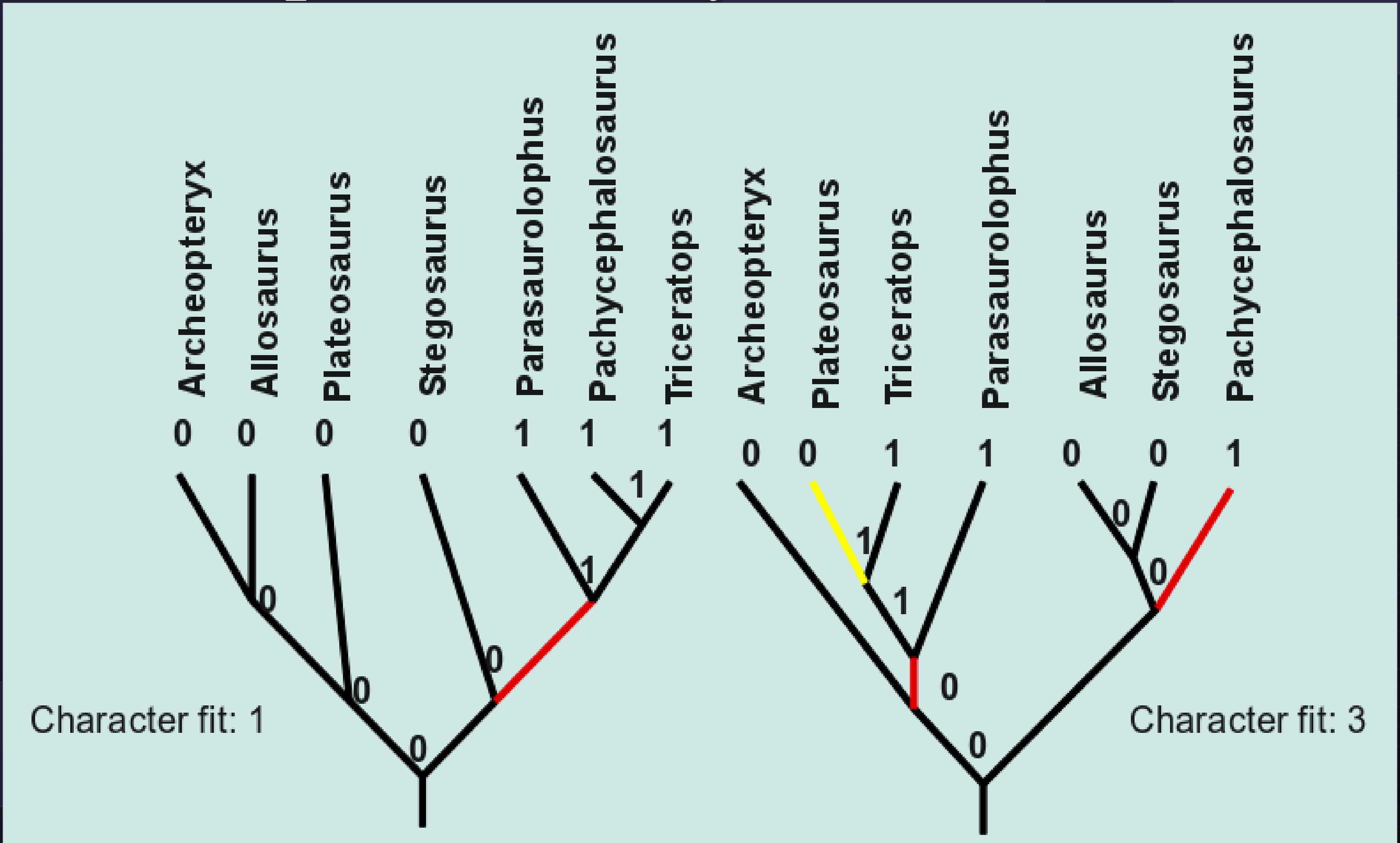
Homoplasy in sequences



Parsimony as maximizing the congruence

- Choosing alternative phylogenetic hypotheses
  - Evaluating alternative trees
  - Parsimony does not create the trees!
- Maximize congruence and minimize homoplasy
- Parsimony helps to find homoplastic characters
- Fit characters to a tree

Unequal teeth layer of dinosaurs

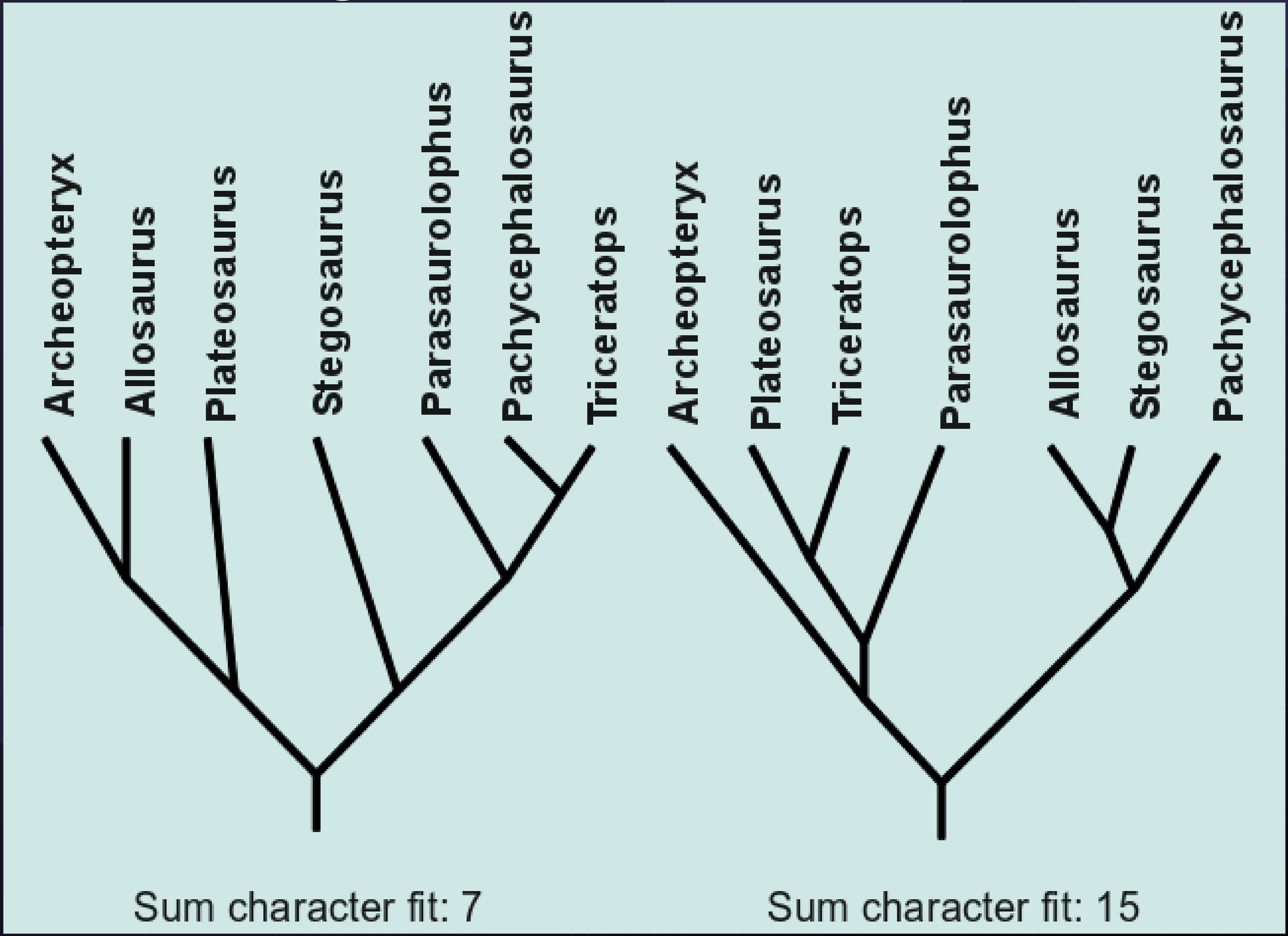


	Archeopteryx	Allosaurus	Plateosaurus	Stegosaurus	Parasaurolophus	Pachycephalosaurus	Triceratops
Unequal teeth layer	0	0	0	0	1	1	1

Most parsimonious trees

- Have the minimum tree length – shortest tree
- If characters not weighted:
  - Most homology, less homoplasy
- If characters weighted
  - A weighted sum of the cost of each character
- You can decide which tree is supported by the dataset
  - No warranty that this is the 'true' tree!

Tree length of dinosaur tree models



Results of parsimony analysis

- We can choose the most parsimonious trees from a large starting set
  - One or more best trees for further analysis
- Hypothesis for the evolution of each characters on a tree
- Branch lengths (calculated from the steps)
- Statistics
  - Character fit
  - Tree length

But what trees to be scored with parsimony?

- Different hypotheses from previous studies
  - Morphology vs sequence based trees
  - Fossils vs taxonomy
- Generated models

Search for trees without preconception

- Exhaustive scoring of all possible trees
  - Possible for max 8-12 taxa
- Heuristic tree generation

Parsimony - advantages

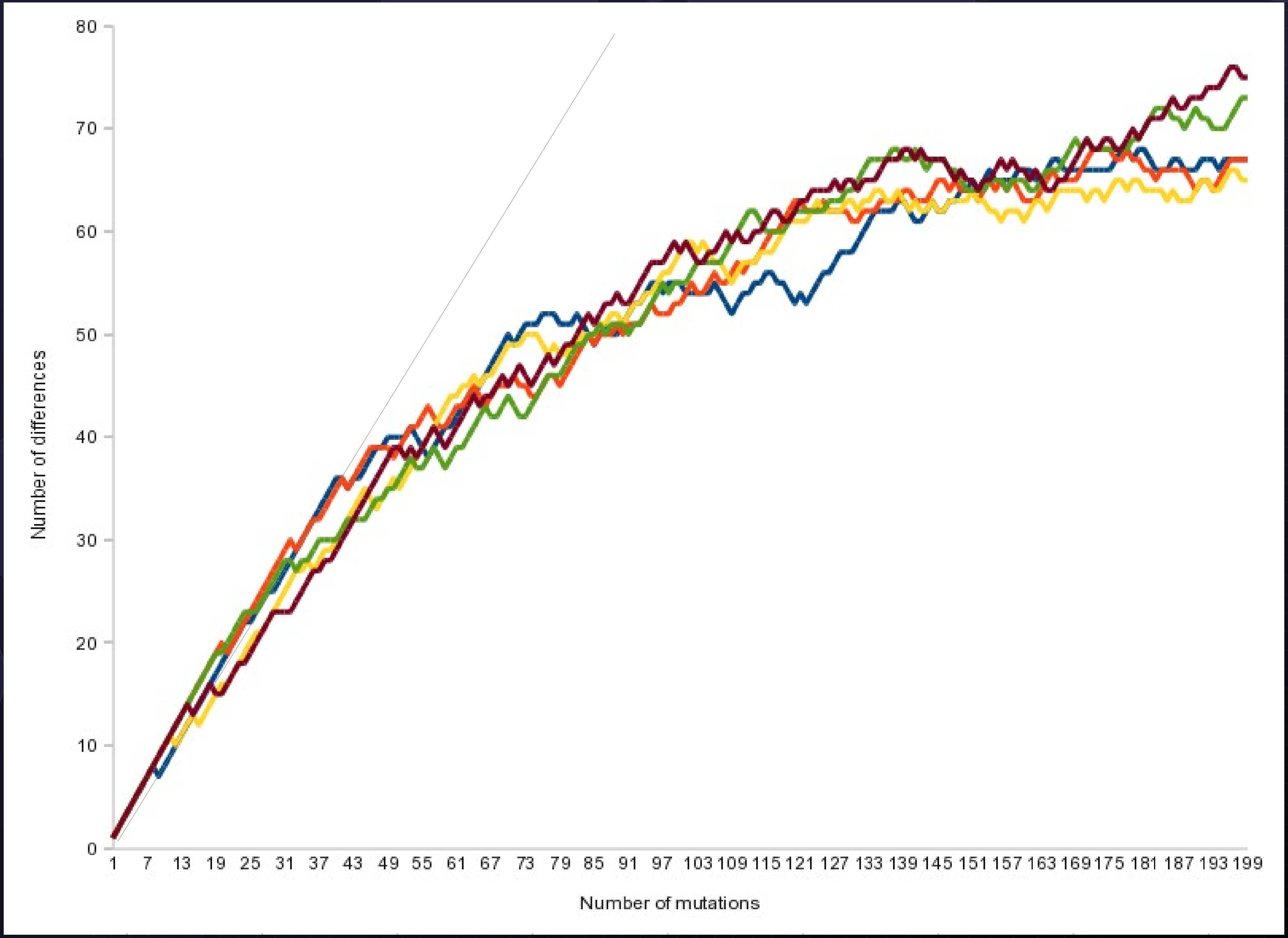
- Simple method – easy to understand
- Does not depend on an explicit model of evolution
- Gives both trees and associated hypotheses of character evolution
- Reliable results if
  - Data is well structured
  - Homoplasy is either rare or widely (randomly) distributed on the tree

Parsimony - disadvantages

- Misleading if homoplasy is common or concentrated in particular parts of the tree
  - thermophilic convergence
  - base composition biases
- Long branch attraction
- Underestimates branch lengths
- Parsimony often justified on purely philosophical grounds
  - Occam's razor

Photo by Ryan Holst

Mutation saturation



Nucleotide substitution models

- Several different models for DNA sequence evolution
- Solid mathematics behind them
- Most important differences between models
  - Nucleotide frequencies
    - 1/4
    - Measured from data
    - Estimated with mathematical models
  - Mutation rates
    - Uniform
    - Transitions/transversions
    - Constant in time/Changing

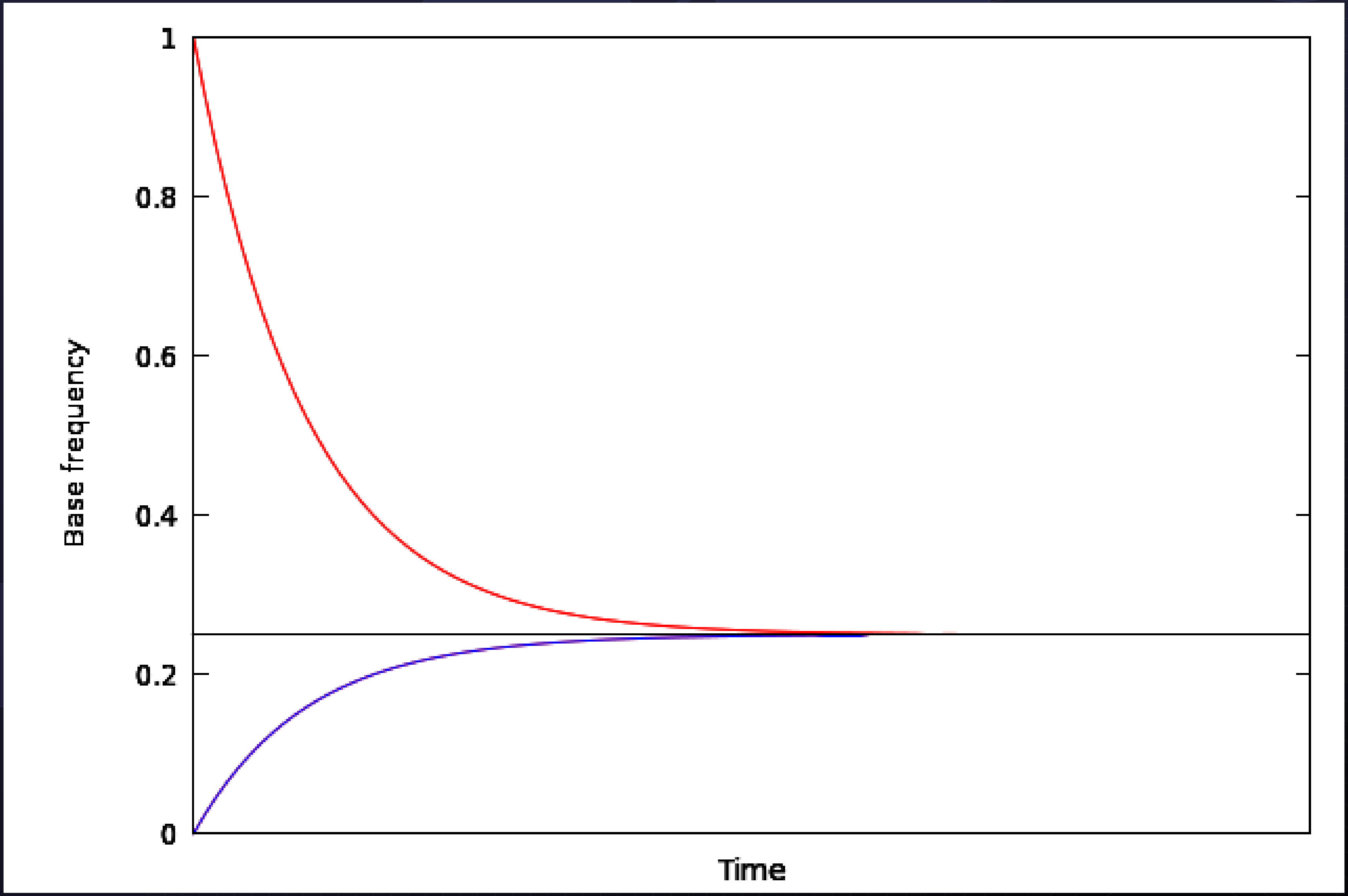
Jukes-Cantor model

- JC69 model (Jukes and Cantor, 1969)
- Equal base frequencies (1/4)
- Single overall mutation rate: μ

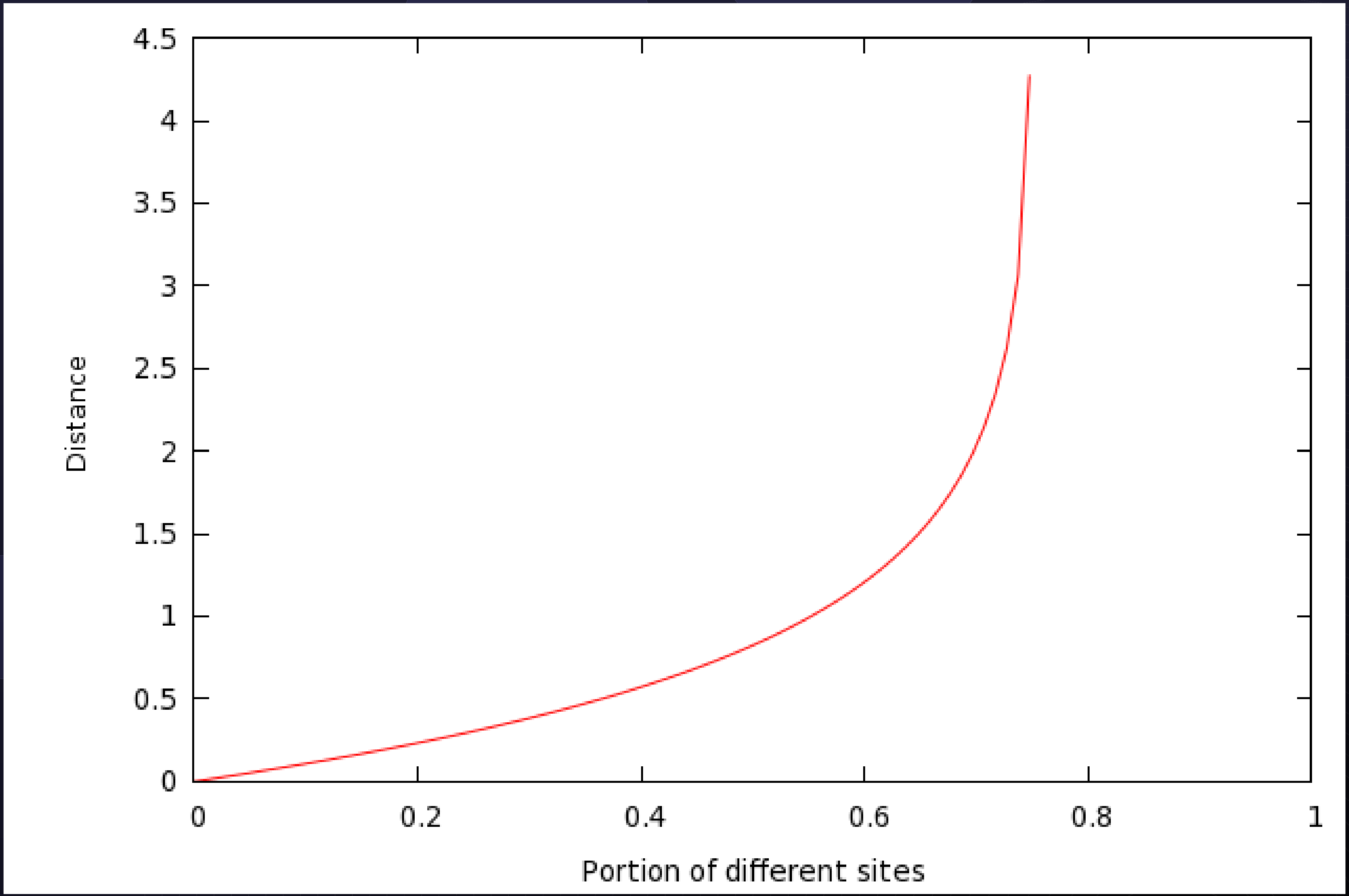
$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

$$d = -\frac{3}{4}\ln(1 - \frac{4}{3}p)$$

Features of JC model



Features of JC model



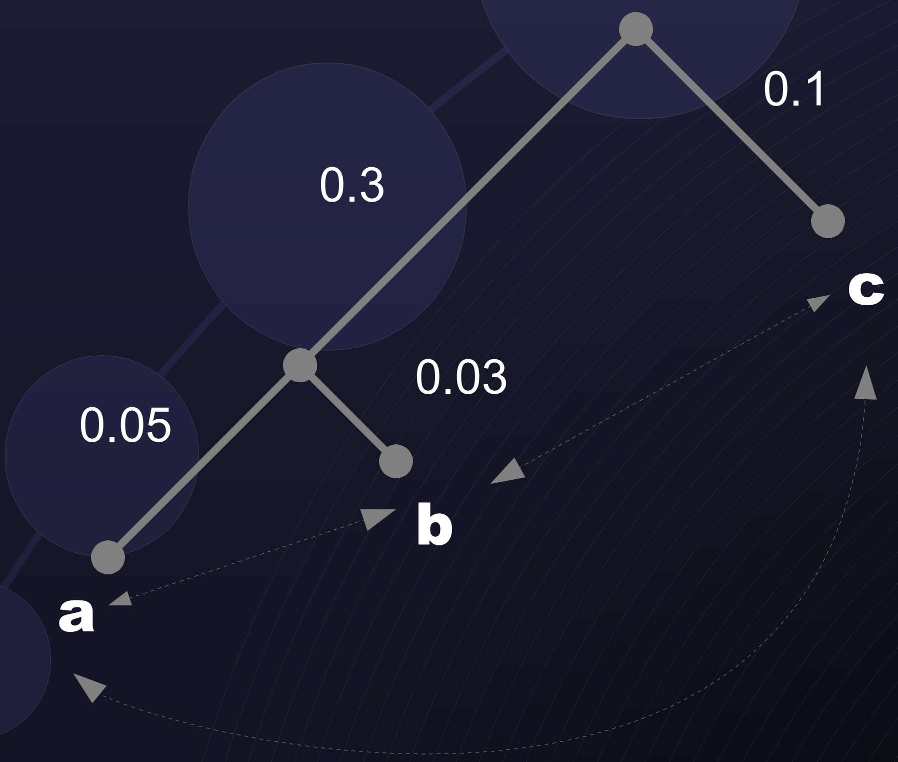
Further models

- K80 model (Kimura, 1980)
  - Different mutation rates for transitions and transversions
- HKY model (Hasegawa, Kishino és Yano 1985)
  - Unique base frequencies
  - Often used for ML
- T92 model (Tamura 1992)
  - GC content
- TN93 model (Tamura és Nei 1993)
  - Unique base frequencies
  - Multiple mutation rates
- Generalized time-reversible model (GTR)
  - Unique base frequencies
  - All mutation rates specified

Use of nucleotide substitution models

- Distance based methods
  - Calculate pairwise distances
  - Model selection: less parameter, less noise: JC69 or K80
- Modeling sequence evolution for ML or Bayesian phylogenetics
  - Model should be fit to the data
  - Less parameters is better

What is pairwise distance?



	a	b	c
a	-	0.08	0.45
b	0.08	-	0.43
c	0.45	0.43	-

Creating a distance matrix

- Mathematical metrics
- Nucleotide substitution models
- Amino-acid substitution matrices
  - PAM vs BLOSUM

	Archeopteryx	Allosaurus	Plateosaurus	Triceratops
A	1	1	1	1
B	0	0	0	1
C	0	0	0	1
D	0	0	0	
E	1	1	1	
F	1	1	0	
G	1	3	3	

	Archeopteryx	Allosaurus	Plateosaurus	Triceratops
Archeopteryx	-	2	3	7
Allosaurus	2	-	1	5
Plateosaurus	2.24	1	-	4
Triceratops	3	2.24	2	-

Amino acid substitution matrices

- Chemical, functional, charge and structural properties of the amino acids
  - Karlin and Ghandour (1985, PNAS 82:8597)
- Weights based on structural similarities and genetic code
  - Dooloittle (Feng et al., 1985 J. Mol. Evol. 21: 112)
- Empirical matrices
  - PAM & BLOSUM

From distance matrix to trees

- Several mathematical possibilities
  - Clustering
- Mostly used methods
  - UPGMA
  - Least squares (LS) method
  - Minimum evolution (ME) method
  - Neighbor Joining (NJ) method
    - Start tree for many other methods

Neighbour-joining

- Greedy algorithm
  - Constructs tree step-wise
- Uses total branch length for evaluating the trees
- Starts with a star shaped tree
- In each step join two taxa and calculates the sum of the branch length, the shortest tree is chosen
- Produces unrooted tree
- Does not assumes equal rates
- Quick and good guess of true phylogeny
- Many very similar implementations

Model based phylogenetics

Maximum likelihood

- What is the probability of seeing the observed data (D) given a model/theory (T)?

Pr(D|T)

Bayesian inference

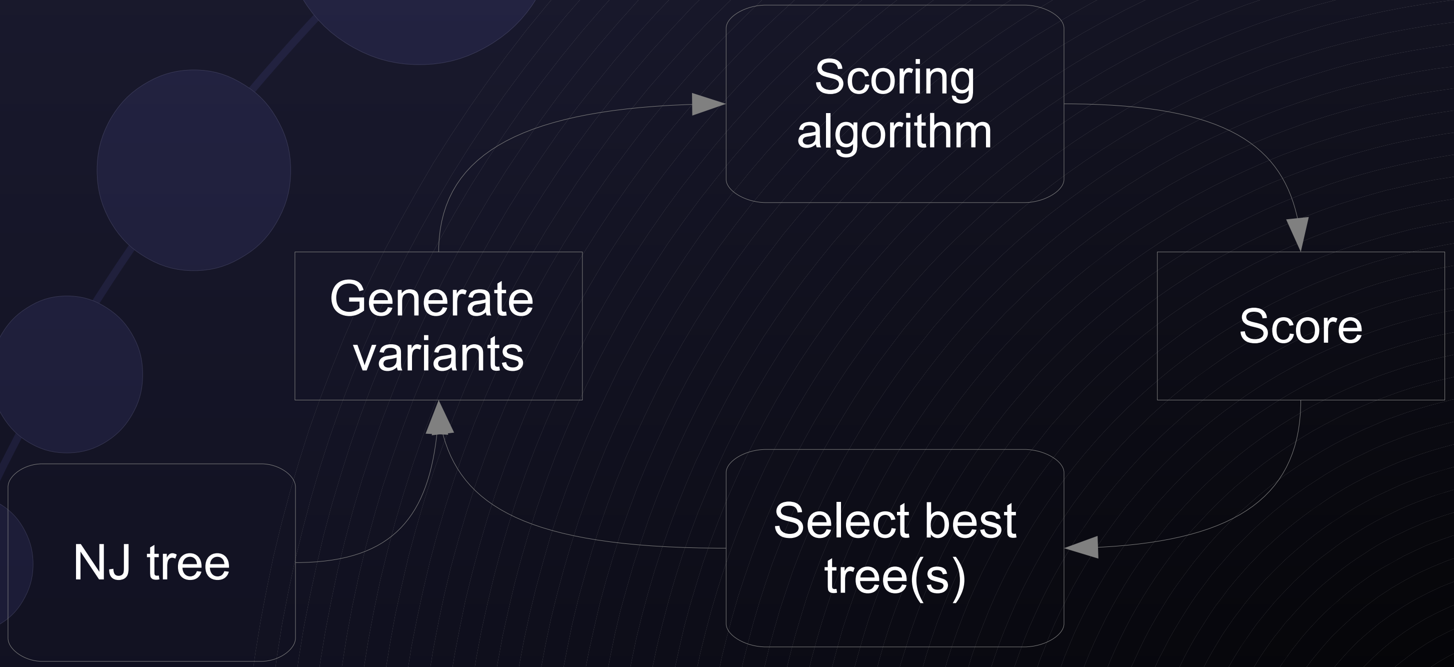
- What is the probability that the model/theory is correct given the observed data?

Pr(T|D)

The basic idea

What is the probability that we would observe an alignment given a tree and a model for the evolution?

Heuristic tree search



Bayesian inference of phylogeny

- Bayesian inference
    - Old marginal statistical method (18th cent.)
    - Suggested for phylogenetics by Felsenstein in 1968
    - Implemented only in 2000
- Efficient numerical solution
    - Quick & dirty
- Applies subjective probability (controversial)
- Very popular method

MrBayes

- You have to specify which model to use
  - Nucleotide substitution model
  - Site rate heterogeneity features
- Model parameters can be provided
  - Base frequency
  - Rate matrix
  - Tree topology
  - Branch lengths
- Model parameters can be estimated

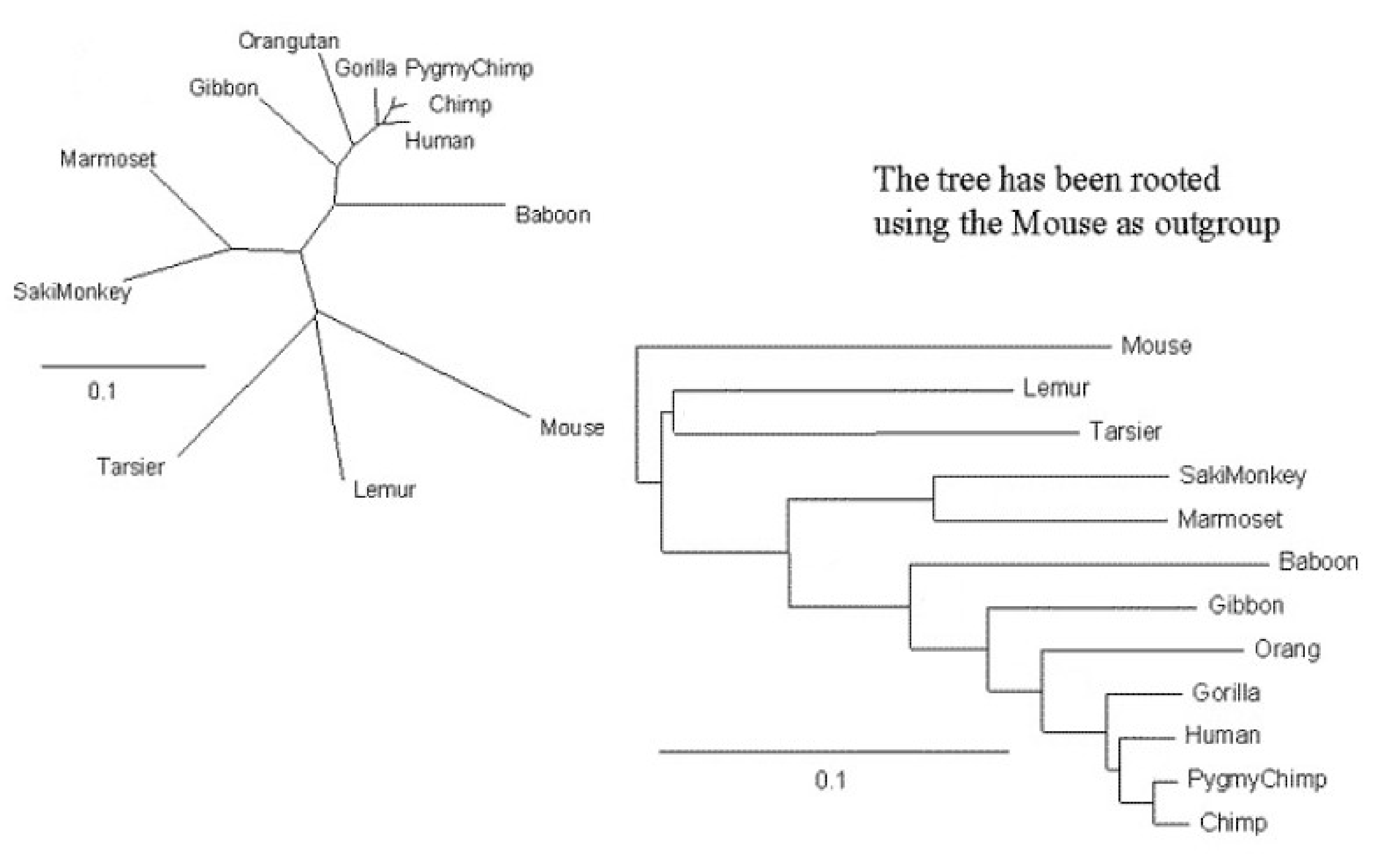
When use model based methods?

- Nucleotide sequences
  - Few: ML
  - Many: Bayesian inference
- You have information on evolution model parameters
- You need information on those parameters
- You have time for the calculations
- You need good statistics for the reliability of the results
- Improving possibilities for othe data types

Auxiliary methodology for phylogenetics

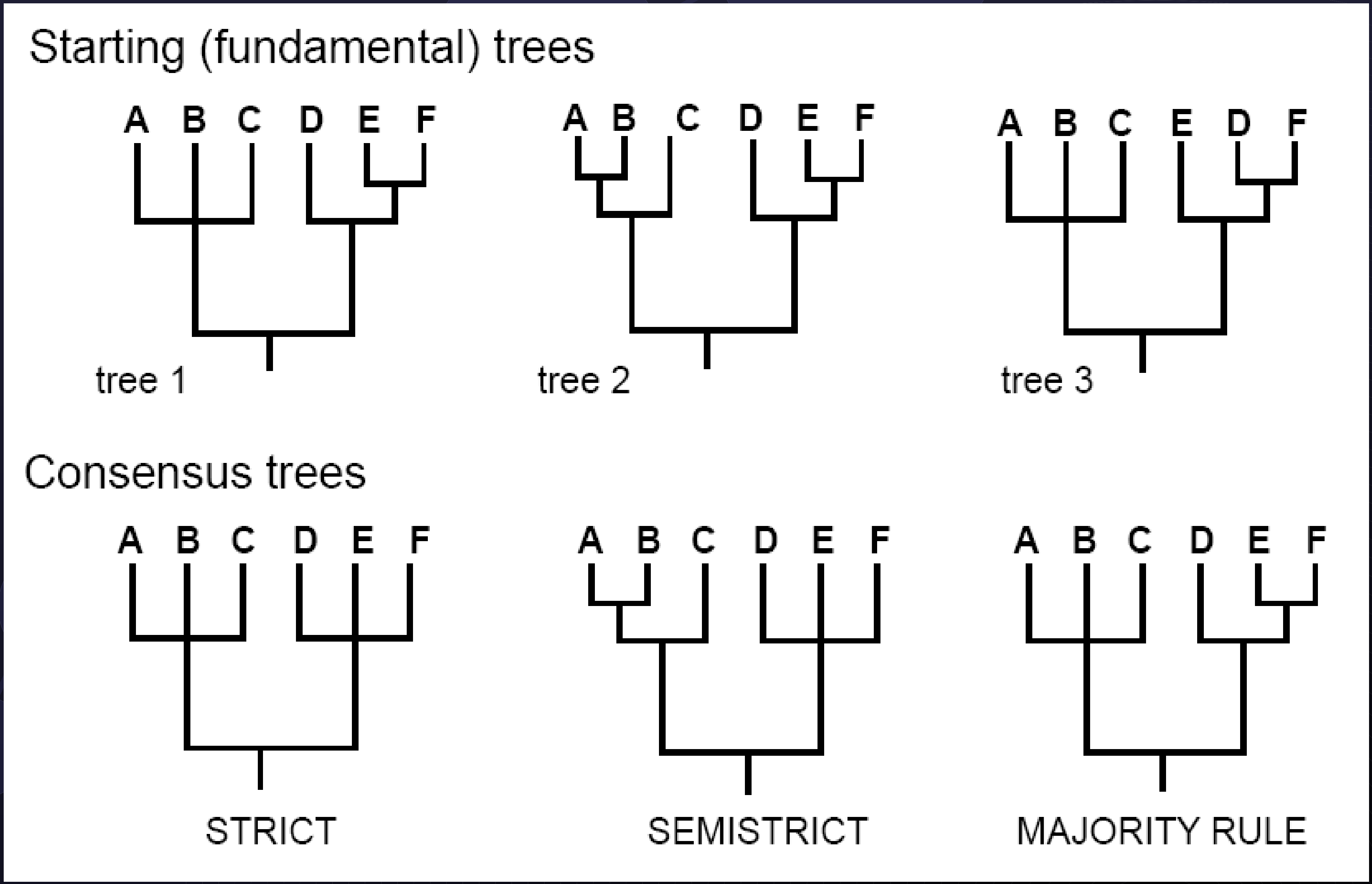
- Rooting trees
- Consensus trees
- Congruence analysis, likelihood ratio test
- Data re-sampling (bootstrapping)
- Tree distances
- Molecular clocks
  - Rate analysis
  - Dating trees

Rooting trees – Outgroup method



The tree has been rooted using the Mouse as outgroup

Consensus methods



Data re-sampling methods

- Generates several sub-samples – replications
  - Non parametric bootstrapping – random site selection with replacement
  - Jackknife – delete half
- Calculate trees for all the sub-samples
  - The method is independent from sampling
- Generate consensus trees
  - 50% majority rule!
  - Only topology is evaluated

A sophisticated phylogenetic analysis

- Sequence selection
  - Representative set
  - Outgroup
- Alignment
- Setting up the dataset
  - Incorporating additional information
    - Stepmatrices
    - Site specific information
    - Topological constrains
    - Non-sequence information

A sophisticated phylogenetic analysis – Choosing a phylogenetic method

- Distance based methods
  - Many sequences
  - Quick tree
- Parsimony
  - Protein sequences
  - Additional non-sequence information

- Maximum likelihood
  - Few (max 20) nucleotide sequences
- Bayesian inference
  - Many/long nucleotide sequences

A sophisticated phylogenetic analysis

- Generating trees
  - Do we have an *a priory* tree?
  - Do we generate trees
    - Exhausting tree search (less than 10 sequences)
    - Heuristics
- Evaluation of the results
  - Bootstrapping
  - Consensus trees

HiDucator

Kérdezz: Csaba.Ortutay@HiDucator.com
www.hiducator.com