MOLECULAR PHYLOGENY

Cont.

Nomenclature of phylogenetic trees

Stages of phylogenetic analysis

[1] Selection of sequences for analysis

[2] Multiple sequence alignment = input data

[3a] Selection of a substitution model – distance-based methods

[3b] Selection of a probabilistic model – character-based methods

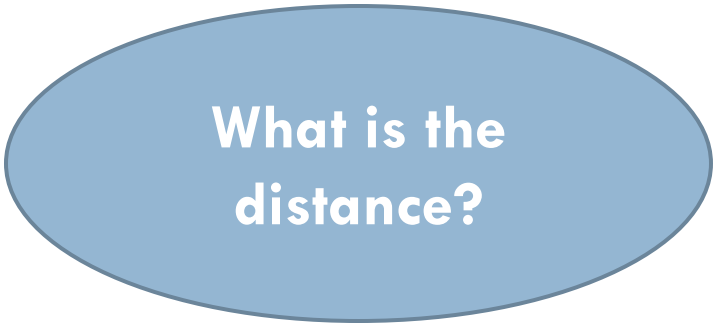
[4] Tree building

[5] Tree evaluation (Bootstraping)

[3] Substitution models in tree building methods

Distance-based

- involve a distance metric, such as the number of amino acid changes between the sequences, or a distance score



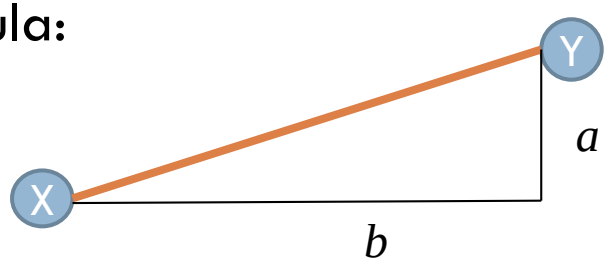
Character-based

- Parsimony analysis involves the search for the tree with the fewest amino acid (or nucleotide) changes that account for the observed differences between taxa
- Maximum likelihood and Bayesian methods are model-based statistical approaches in which best tree is inferred that may account for observed data

Distance function = metric

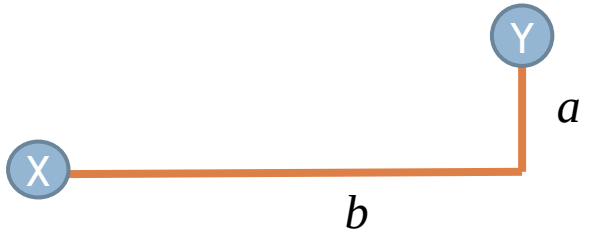
- **Distance** - numerical description of how far apart objects are
- **Metric = distance function** defines distance between elements of a *set* (*metric space*)
- **Euclidean metric (distance) –** the shortest way from X to Y; distance function is given by the Pythagorean formula:

$$D = \sqrt{a^2 + b^2}$$



- **Manhattan distance:** distance between X and Y is the sum of the absolute differences of their coordinates

$$D = a + b$$



Distances in protein sequence alignments (see MSA)

Triangular distance

Manhattan – like metric in protein sequences comparisons:

Dis (A,B) = Dis (A,C) + Dis (B,C)

Applicable only for closely related proteins

Kimura distance

Distance based on a probability that one residue will change into another

allowing multiple changes in one position

…

Distance in trees building – DNA diversity

Distance formula should provide a model describing the probability that one residue (nucleotide) will change into another

Hamming distance

align pairs of sequences, than count the number of differences.

Thus, degree of divergence (distance) D is:

$$D = n / N$$

N – length of an alignment

n – number of differences

Note: *observed differences do not equal genetic distance!*
Genetic distance involves mutations that are not observed directly.

Models of nucleic acids substitution

Jukes and Cantor (1969) proposed another corrective formula for DNA alignments

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

p-proportion of residues that differ

Assumptions:

- each residue is equally likely to change into any other (i.e. the rate of transversions equals the rate of transitions).
- all four nucleotides are present in DNA sequence with the same frequences

[3] Models of nucleic acids substitution

Jukes and Cantor formula: $D = (-\frac{3}{4}) \ln (1 - \frac{4}{3} p)$

Consider an alignment where 3 per 60 aligned residues differ.

The normalized Hamming distance is: $D_H = 3/60 = 0.05$.

The Jukes-Cantor correction is: $D_{JC} = (-\frac{3}{4}) \ln (1 - \frac{4}{3} 0.05) = 0.052$

Consider an alignment where 30/60 aligned residues differ:

$D_H = 0.5$ $\quad$ $D_{JC} = (-\frac{3}{4}) \ln (1 - \frac{4}{3} 0.5) = 0.82$

The Jukes-Cantor correction is more substantial!

[3] Models of nucleotide substitution – mutations frequency in DNA

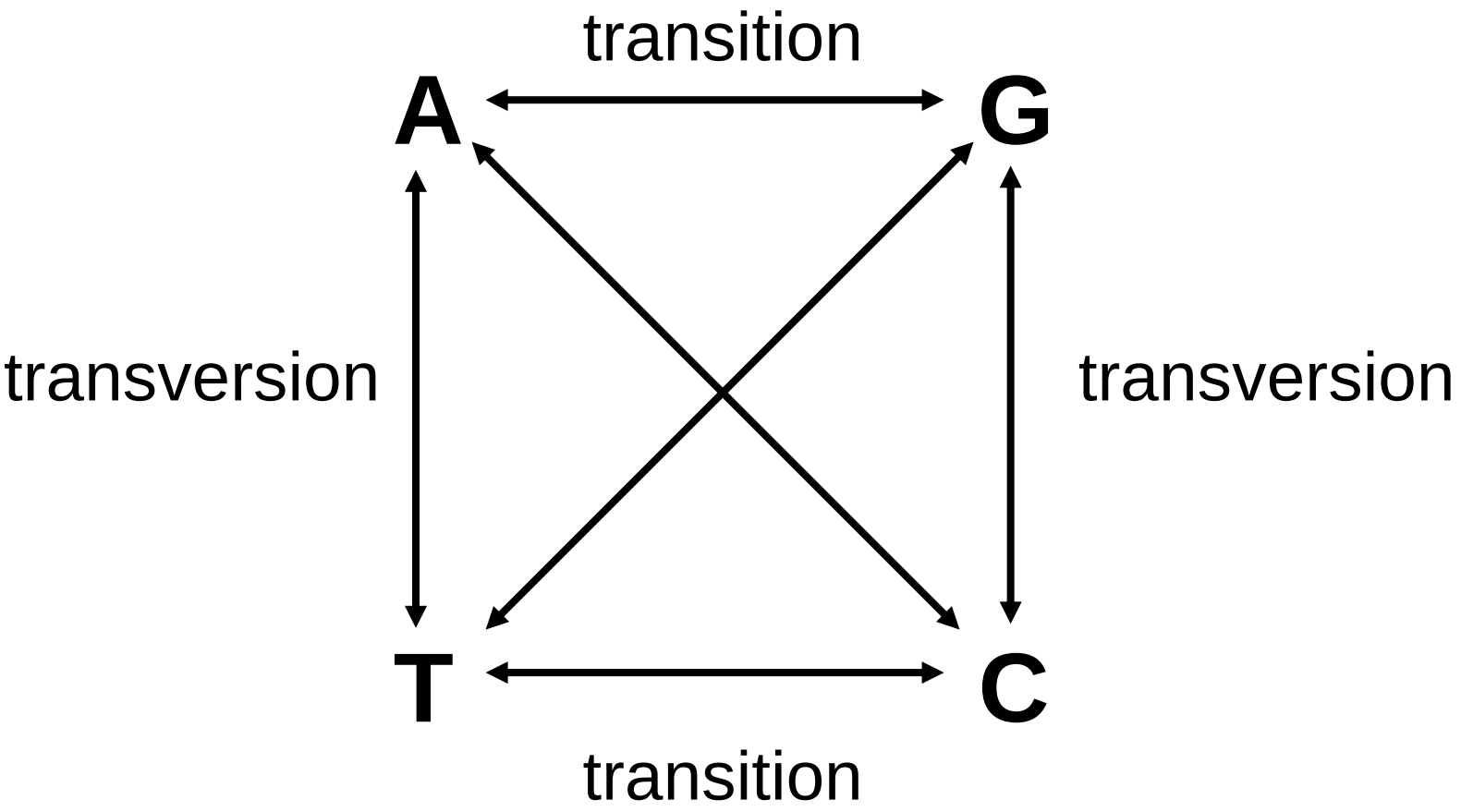
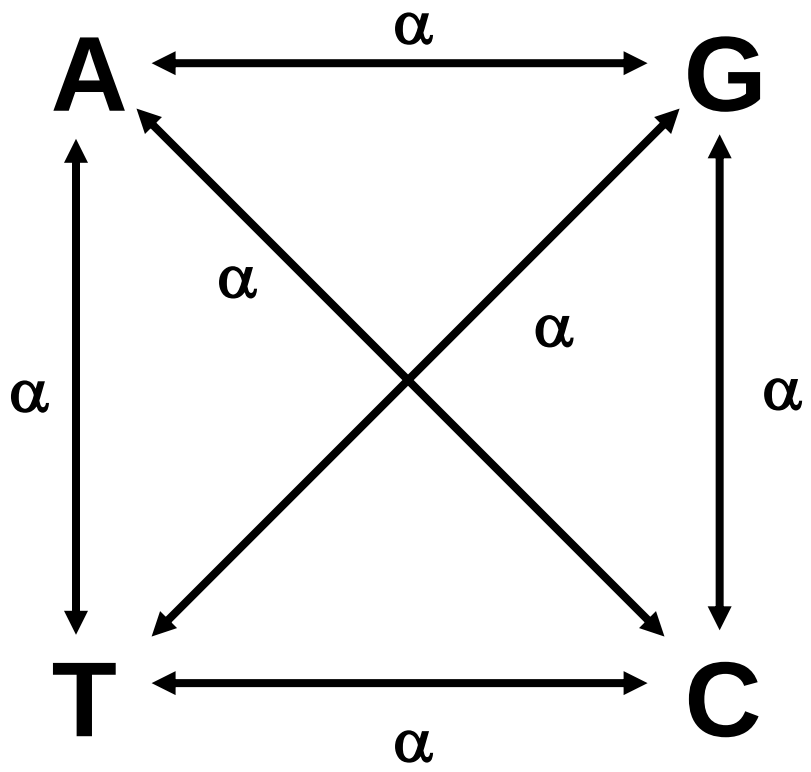


Fig. 7.21
Page 250

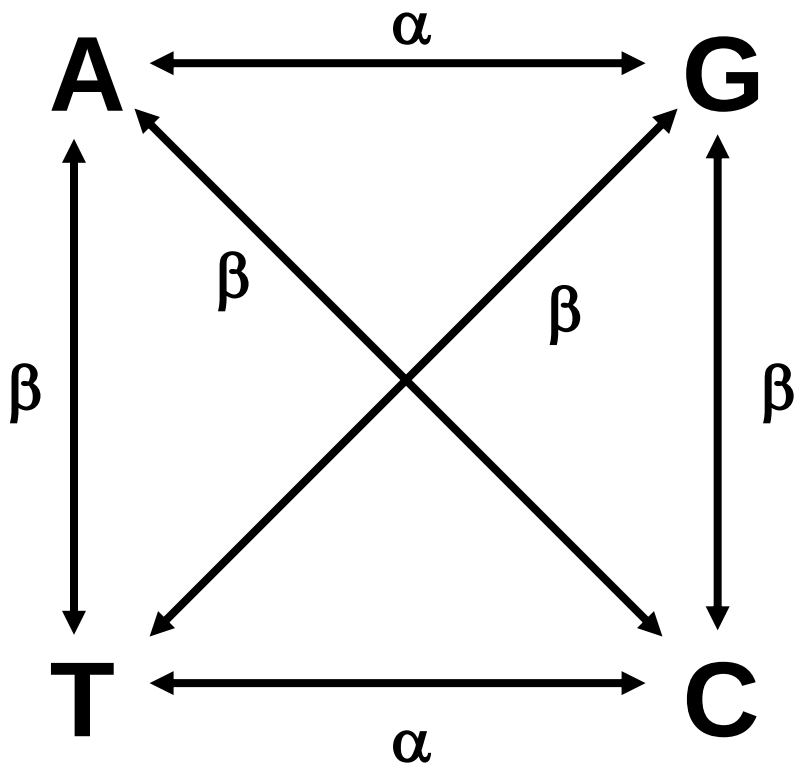
[3] Models of nucleotide substitution



e.g. Jukes and Cantor one-parameter model
assumes equal frequency of trasitions and transvertions

Fig. 7.21
Page 250

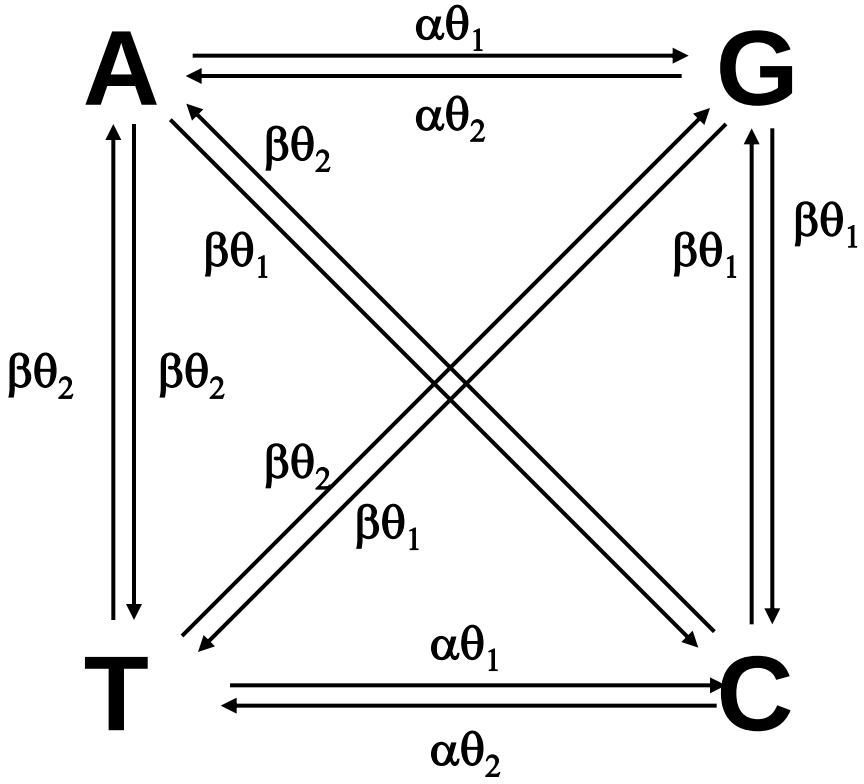
[3] Models of nucleotide substitution



Kimura's model of nucleotide substitution assumes $\alpha \neq \beta$ & $\beta > \alpha$

Fig. 7.21
Page 250

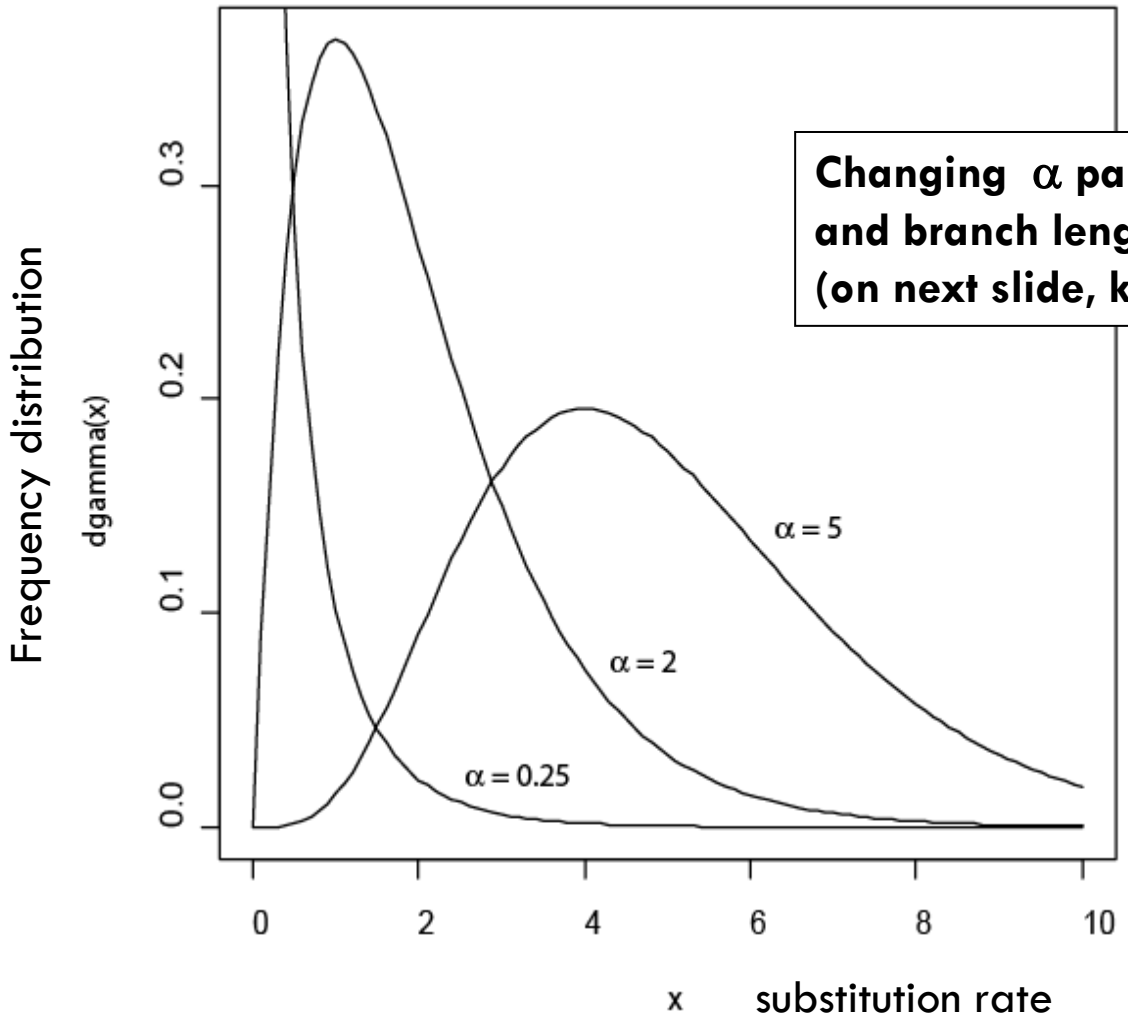
[3] Models of nucleotide substitution



Tamura's model accounts for variations in GC content

Fig. 7.21
Page 250

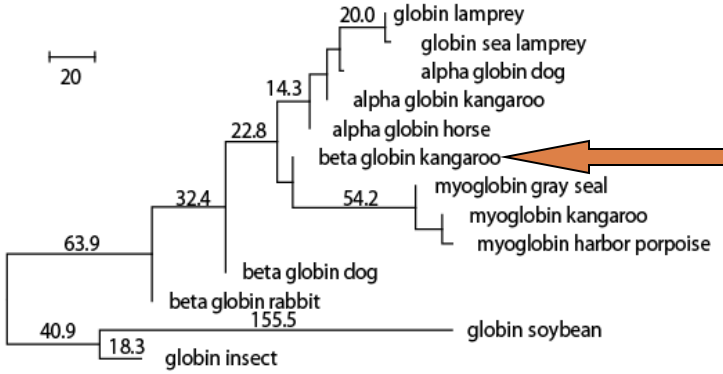
Gamma distribution – based models account for unequal substitution rates across variable sites



Changing α parameter does alter the topology and branch lengths of the tree…
(on next slide, kangaroo globin switches clades)

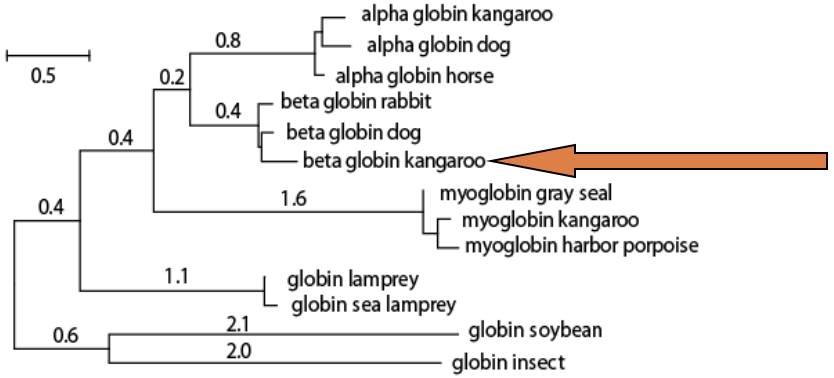
Fig. 7.22
Page 252

(a) Neighbor-joining tree with Poisson correction and gamma distribution shape parameter α=0.25



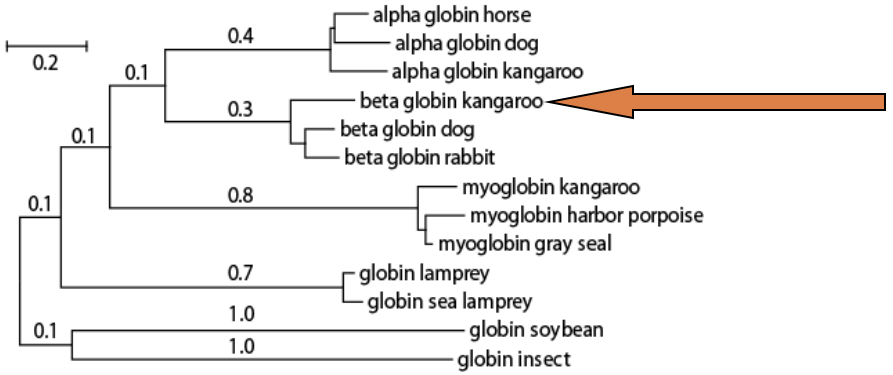
α = 0.25

(b) Neighbor-joining tree with Poisson correction and gamma distribution shape parameter α=1



α = 1

(c) Neighbor-joining tree with Poisson correction and gamma distribution shape parameter α=5



α = 5

Fig. 7.23
Page 253

Distance in trees building - models of aa substitution

Poisson correction to Hamming distance to correct for multiple substitutions at a single site:

$$D = -\ln(1-p)$$

p – proportion of residues that differ

Assumptions:

equal substitution rates across sites

equal amino acids frequencies

example from MSA:

$$D = -\ln S_{eff}$$

S_{eff} = normalized similarity score

$$S_{eff} = (S_{real(ij)} - S_{rand\ (ij)}) / (S_{iden(ij)} - S_{rand(ij)}) \times 100$$

Poisson distribution

Further assumptions:

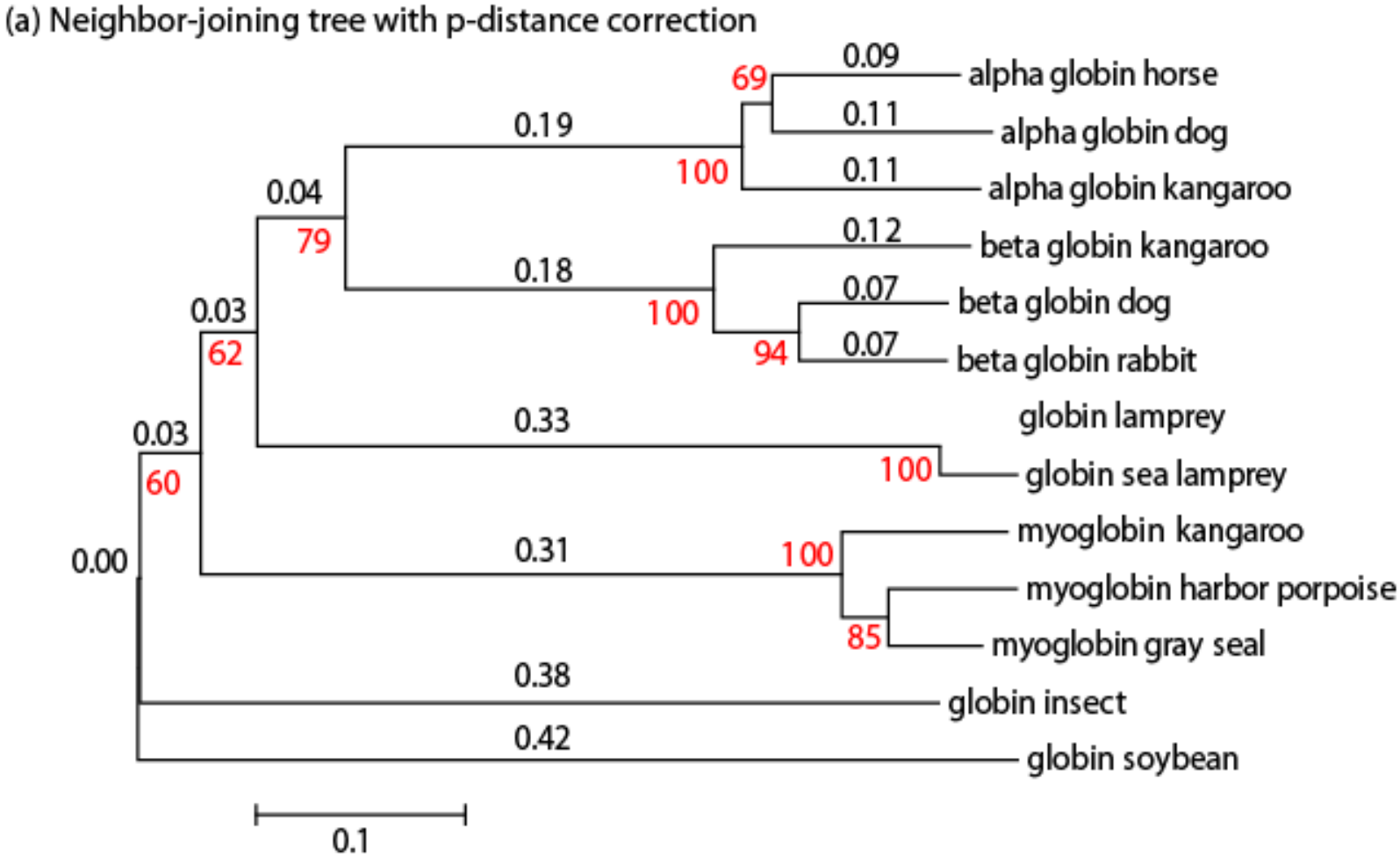
1. Probablilty of observing a change is small and proportional to the lengh of time interval
2. Number of changes is constant in time
3. Changes occur independently

Poisson distribution: $P(X) = e^{-\mu}\mu^{X} / X!$

P(X) – probability of X occurances per unit of time,

μ- population mean number of changes over time

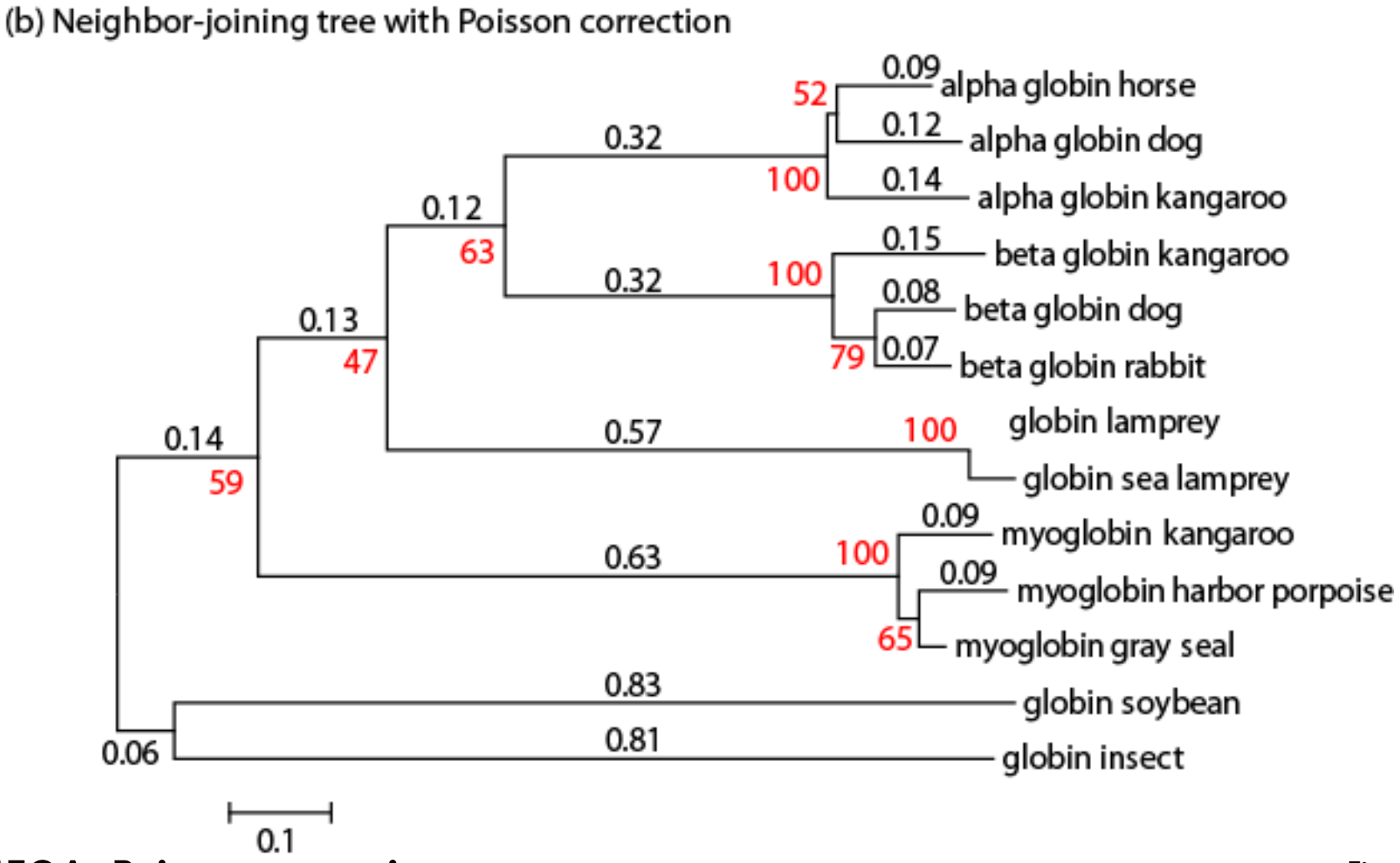
Choice of substitution model influences the length of branches in a tree



MEGA: „p-distance correction" = Hamming distance

Fig. 7.20
Page 249

Choice of substitution model influences the length of branches in a tree



MEGA: Poisson correction

Fig. 7.20
Page 249

Stages of phylogenetic analysis

[1] Selection of sequences for analysis

[2] Multiple sequence alignment = input data

[3a] Selection of a substitution model – distance-based methods

[3b] Selection of a probabilistic model – character-based methods

[4] Tree building

[5] Tree evaluation (Bootstraping)

[4] Tree-building methods

Distance-based

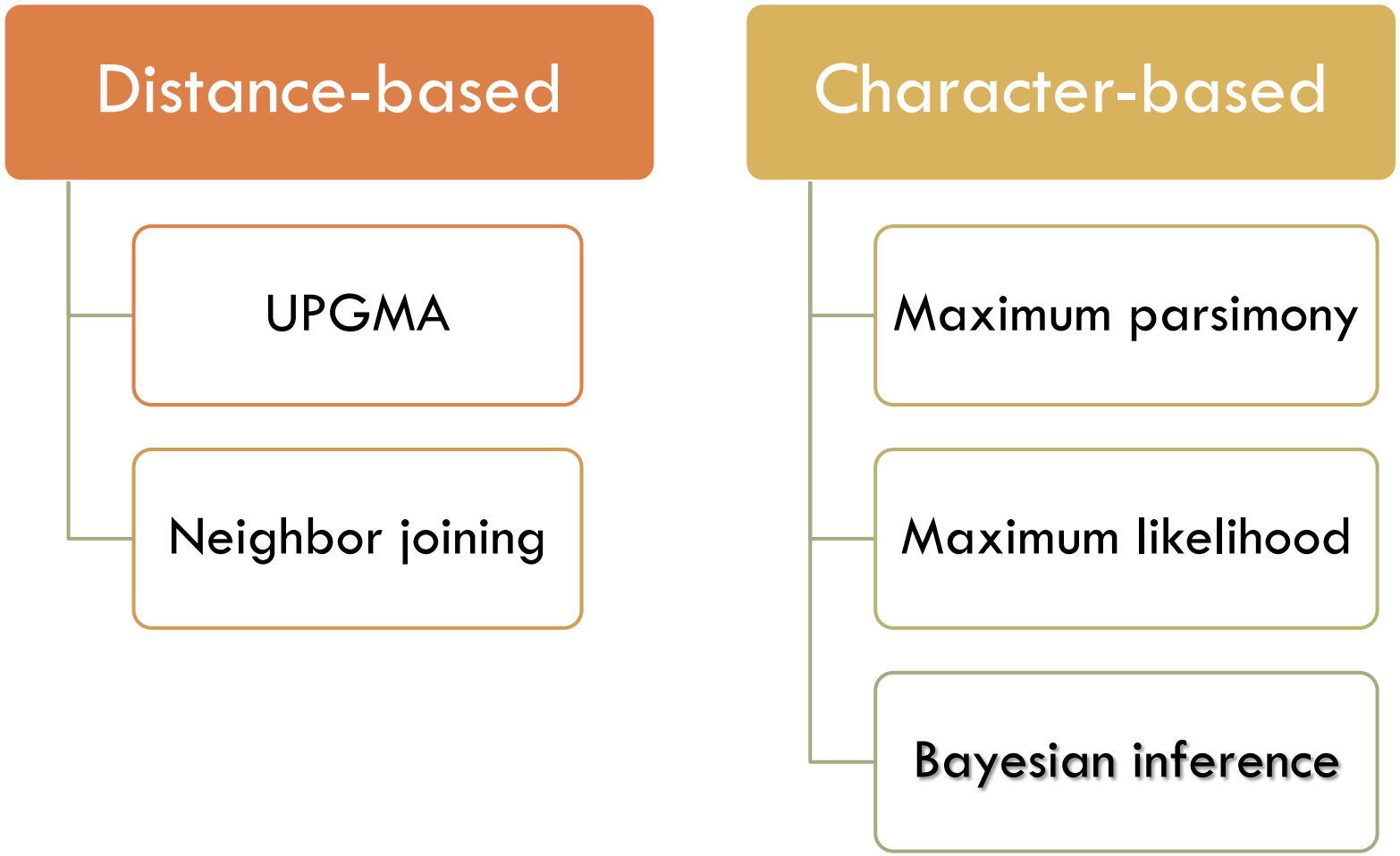
- involve a distance metric, such as the number of amino acid changes between the sequences, or a distance score

Distance formula should provide a model describing the probability that one residue will change into another – e.g. computed on the basis of all possible pairwise alignments in the protein seqs. set; models of nt substitution may assume transitions/transversions rates

Character-based

- identify positions that best describe how residues are derived from common ancestors
- **Parsimony analysis** involves the search for the tree with the fewest amino acid (or nucleotide) changes that account for the observed differences between taxa
- **Maximum likelihood** and **Bayesian** methods are model-based statistical approaches in which best tree is inferred that may account for observed data

[4] Tree-building methods



Character-based methods: maximum parsimony

- Rather than pairwise distances between proteins, evaluate the aligned columns of characters (amino acid residues)

- The goal:

  To find the tree with the shortest branch lengths possible.
  Thus we seek the most **parsimonious ("simple") tree**

[4.3] Tree-building methods: maximum parsimony

[1] Identify informative sites – *constant characters are not usefull.*

[2] Construct *all possible* trees, counting the number of changes required to create each tree.

For 12 taxa or fewer - evaluate all possible trees exhaustively;
For >12 taxa perform a heuristic search.

[3] Select the shortest tree (or trees).

[4.3] Tree-building methods: Maximum parsimony

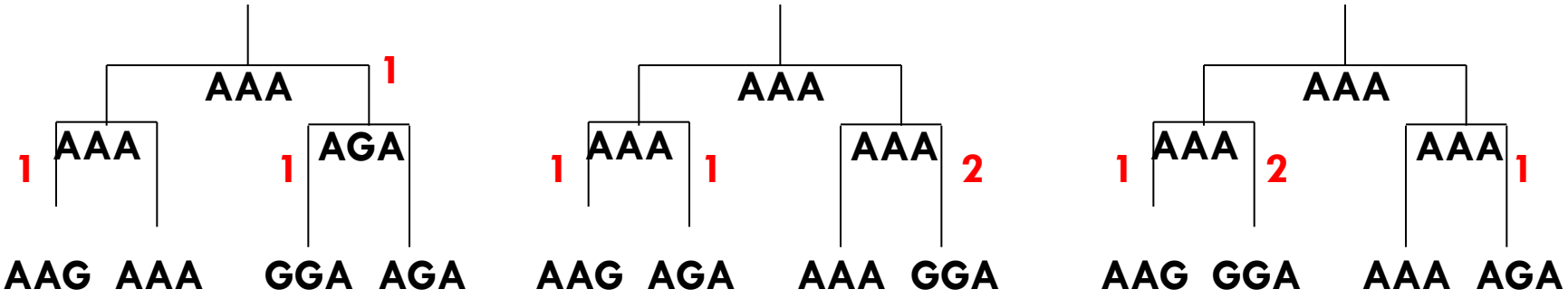
Consider these four taxa (OTU):

AAG
AAA
GGA
AGA

How might they have evolved from a common ancestor such as AAA?

[4.3] Tree-building methods: Maximum parsimony

3 examples of possible trees:



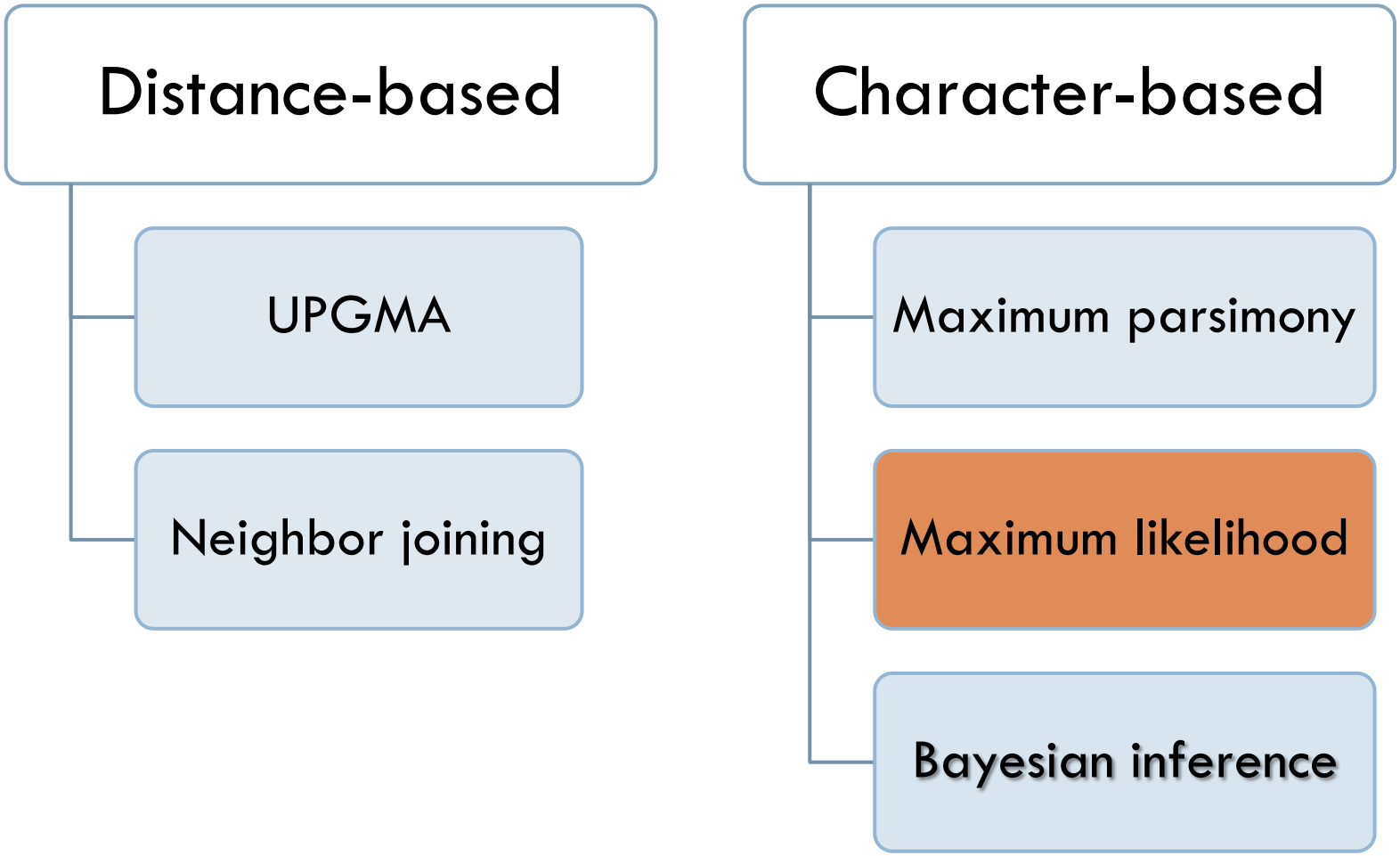
Cost = 3 **Cost = 4** **Cost = 4**

Choose the tree(s) with the lowest cost (lowest number of changes).

In maximum parsimony, there may be more than one tree having the lowest total branch length.

You may compute the consensus best tree

[4] Tree-building methods



Character-based methods: Maximum likelihood

- Maximum likelihood is computationally intensive.
- A likelihood is calculated for the probability of each residue in an alignment, based upon some model of the substitution process.

- ***Goal:***

  What are the tree topology and branch lengths that have the greatest likelihood of producing the observed data set?

ML is implemented in the TREE-PUZZLE program, as well as PAUP and PHYLIP

Maximum likelihood applied in Tree-Puzzle

Quartet puzzling - heuristic algorithm for maximum likelihood trees building method (Strimmer & von Haeseler, 1996)

[1] Reconstruct all possible quartets A, B, C, D from whole set of N input sequences; construct all possible unrooted trees for the quartets:
((A,B), (C,D)); ((A,C),(B,D)) and ((A,D),(B,C))

For 12 myoglobins there are 495 possible quartets.

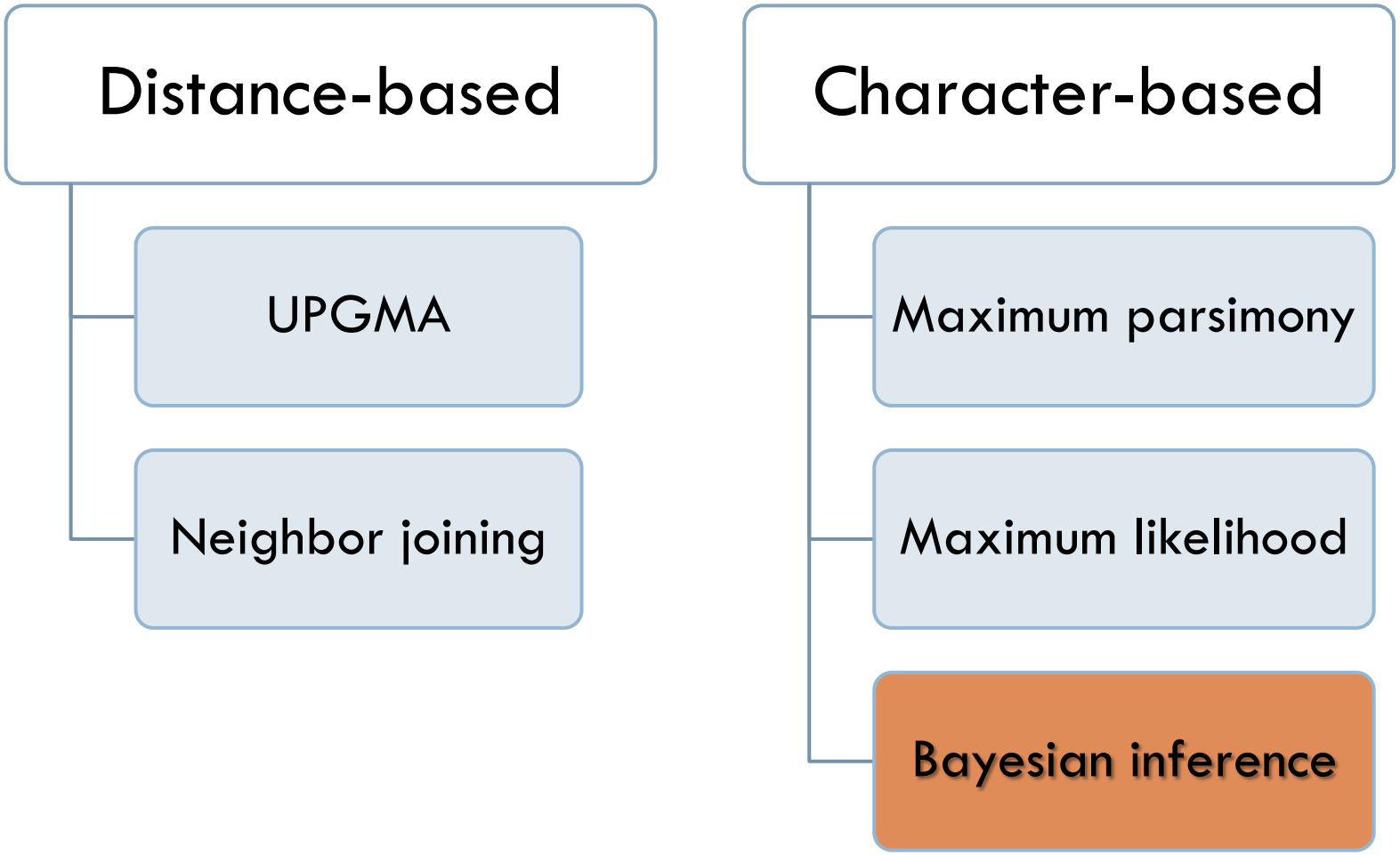
[2] Puzzling step: begin with one quartet tree. N-4 sequences remain on ***random list***. Add them to the branches of quarted tree from [1] systematically, optimising each new branch. **Compute likelihood** of the resulting tree.

[3] Repeat whole procedure for numerous ***puzzled*** random lists of sequences

[4] Report a consensus tree(s) = with the most frequent topology

Higgs, Attwood pp. 258-61

[4] Tree-building methods



Bayesian inference - Bayes' theorem

The probability of an event A *given an event B* depends not only on the relationship between events A and B but on the probability of occurrence of each event

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

P(*A*) – the prior probability of *A (regardless of any other information)*.
P(*A* | *B*) is the conditional probability of *A*, given *B*.
P(*B* | *A*) is the conditional probability of *B* given *A*.
P(*B*) - the prior probability of *B (regardless of any other information)*

Bayes' theorem- evaluation of drug test results

Corporation decides to test its employees for drug use.
Assume that only 0.5% of the employees actually use the drug and that a certain drug test is 99% sensitive and 99% specific
What is the probability that, given a positive drug test result, an employee is actually a drug user? $P(D|+)$

$P(D)$= the probability that the employee is a drug user. This is 0.005
$P(+|D)$= the probability that the test is positive, given that the employee is a drug user. This is 0.99, since the test is 99% sensitive.
$P(+)$= the probability of a positive test event: it is found by adding the probability that a true positive result will appear (= 99% × 0.5% = 0.495) + the probability that a false positive will appear (= 1% × 99.5% = 0.995)

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{0.99 \times 0.005}{0.0495 + 0.0995} = 0.3322$$

Bayes' theorem

- **Bayesian inference** refers to the likelihood that a particular hypothesis is true given some observed evidence (the so-called *posterior probability* of the hypothesis) comes from a combination of the *prior probability* of the hypothesis and the compatibility of the observed evidence with the hypothesis.

- *Probability a priori* – simple probability - derived purely by deductive reasoning
- *Probability a posteriori* – conditional probability assigned after some relevant evidence is taken into account

Higgs, Attwood pp.268-9 & 336-9

Character-based methods: Bayesian inference

Calculate:

$$P(\text{Tree}|\text{Data}) = \frac{P(\text{Data}|\text{Tree}) \times P(\text{Tree})}{P(\text{Data})}$$

P (Tree | Data) is the posterior probability of distribution of trees. Ideally this involves a summation over all possible trees.

In practice, *Monte Carlo Markov Chains* (MCMC) are run to estimate the posterior probability distribution.

Bayesian approaches require you to specify prior assumptions about the model of evolution (user determine probability *a priori* of such parameters as: tree topology, branch lenghts and rates of substitutions)

Bayesian inference is used in MrBayen

Higgs, Attwood pp.268-9 & 336-9

[5] Evaluating trees: bootstrapping

Bootstrapping is a commonly used approach to measuring the robustness of a tree topology.

To bootstrap, make an artificial dataset obtained by randomly sampling columns from your multiple sequence alignment.

Make the dataset the same size as the original.
Do 100 (to 1,000) bootstrap replicates.

Observe the percent of cases in which the assignment of clades in the original tree is supported by the bootstrap replicates.
>70 % is considered significant.

Pevsner, Page 266