Chapter 10

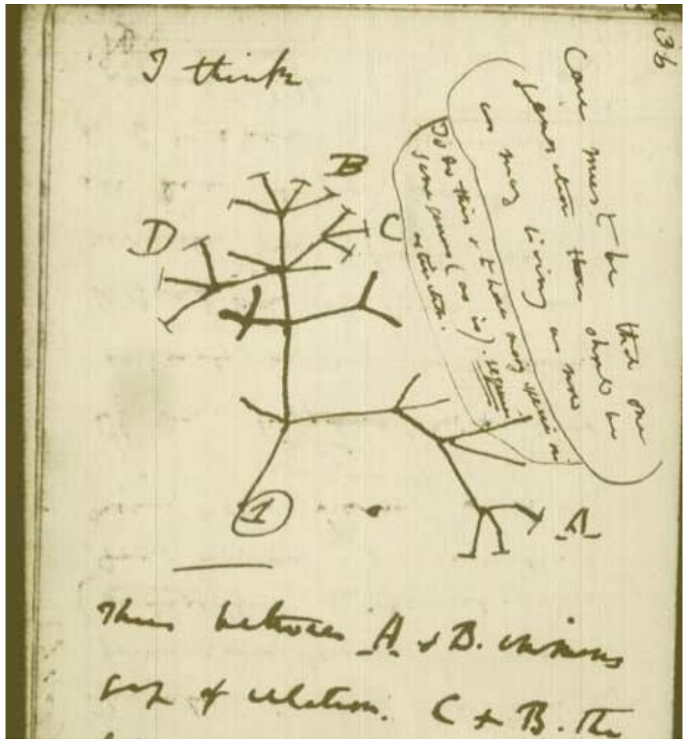
Phylogenetics Basics

Overview

1. Introduction
2. Introduction to Biological Databases
3. Pairwise Sequence Alignment
4. Database Similarity Searching
5. Multiple Sequence Alignment
6. Profiles and Hidden Markov Models
7. Protein Motifs and Domain Prediction
8. Gene Prediction
9. Promoter and Regulatory Element Prediction
10. Phylogenetics Basics
11. Phylogenetic Tree Construction Methods and Programs
12. Protein Structure Basics
13. Protein Structure Visualization, Comparison and Classification
14. Protein Secondary Structure Prediction
15. Protein Tertiary Structure Prediction
16. RNA Structure Prediction
17. Genome Mapping, Assembly and Comparison
18. Functional Genomics
19. Proteomics

Phylogenetics

- *Phylogenetics*: the study of the evolutionary history of organisms
- Evolutionary history is typically represented as a family tree or *phylogeny*:



Tree of life
Charles Darwin
(1837)

- Traditional analysis based on morphology (physical characteristics) and fossil records

The trouble with morphology and fossil records

- Fossil records are usually fragmentary (*i.e.* incomplete), due to low abundance, restricted habitats, poor conservation, *etc.*
- Interpretation of qualitative morphological traits can be ambiguous, making determination of phylogenetic relationships unreliable
- Essentially non-existent for microorganisms

Using molecular data for phylogenetics

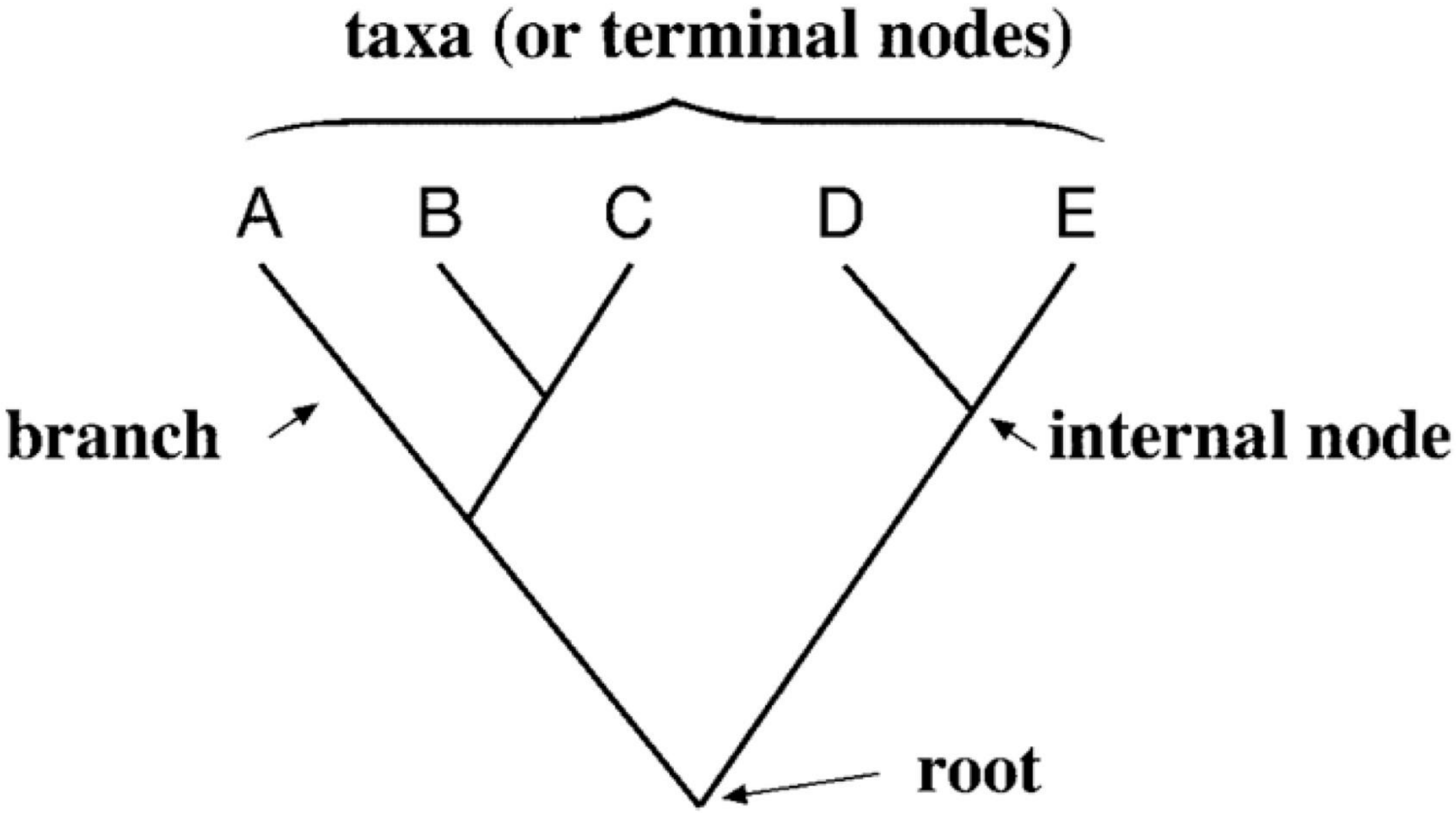
- Genomes accumulate mutations over time, turning genes into "molecular fossils"
- Comparative analysis of homologous genes from related organisms allows reconstructing the evolutionary history of the genes (and that of the organisms)
- Advantages: molecular data are more readily available than fossils, there is no sampling bias, data are easier to interpret and quantitative methods can be used to construct trees in a more objective manner

Assumptions in molecular phylogenetics

- Molecular sequences that are used in phylogenetic construction are indeed homologous, *i.e.* they really share a common origin ("*vertical evolution*")
- Each position (nucleotide / amino acid) in a sequence evolved independently
- Models usually based on the concept that phylogenetic divergence happens through bifurcation, *i.e.* the parent branch splits into two daughter branches

Tree terminology

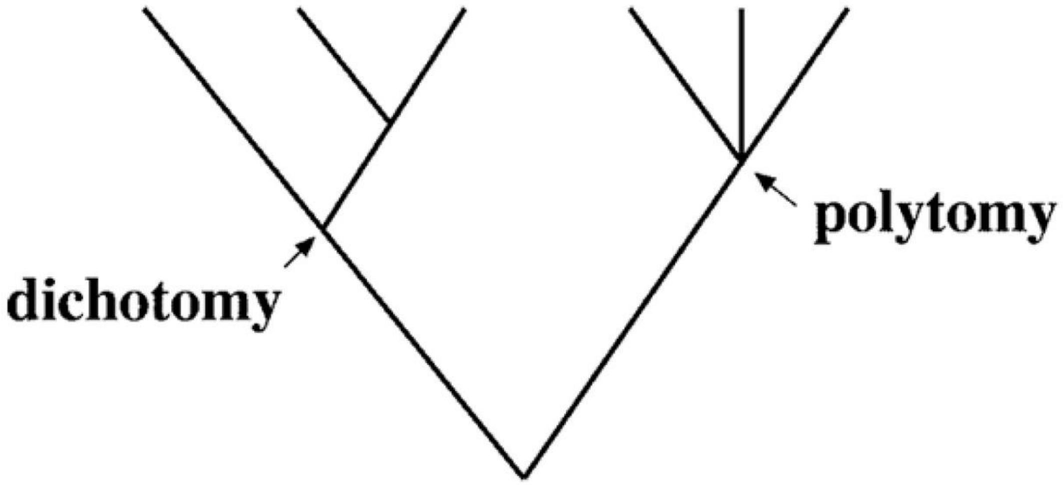
- Lines in the tree = ***branches***
- Tips of the branches = present-day species or ***taxa***
- Connecting point where two adjacent branches join = ***node*** (represents an inferred ancestor or extant taxa)
- Bifurcating point at the very bottom of the tree = ***root*** node (represents the common ancestor of all members of the tree)



- Group of taxa descended from a single common ancestor = ***clade*** or ***monophyletic group*** (*e.g.* A, B and C)
- Group of taxa with more than one closest common ancestor = ***paraphyletic*** (*e.g.* B, C and D)

Dichotomy *vs* polytomy

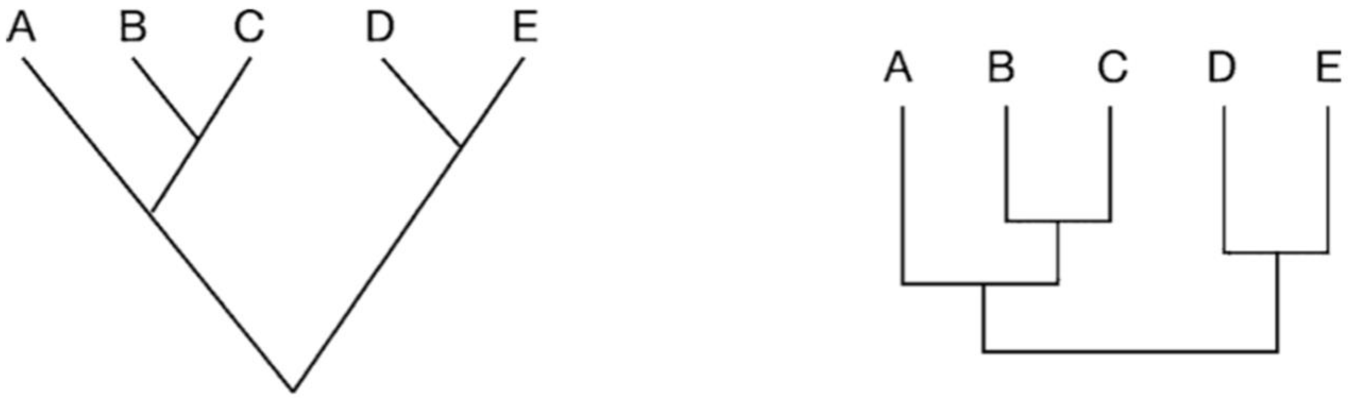
- All branches bifurcate = *dichotomy*
- Branch points with more than two descendants = multifurcating node
- Phylogeny with multifurcating branches = *polytomy*
  - Either ancestral taxon giving rise to more than two immediate descendants = *radiation*
  - Or an unresolved phylogeny in which the exact order of bifurcations cannot be determined precisely



Tree representations: cladogram *vs* phylogram

Cladogram = unscaled trees:

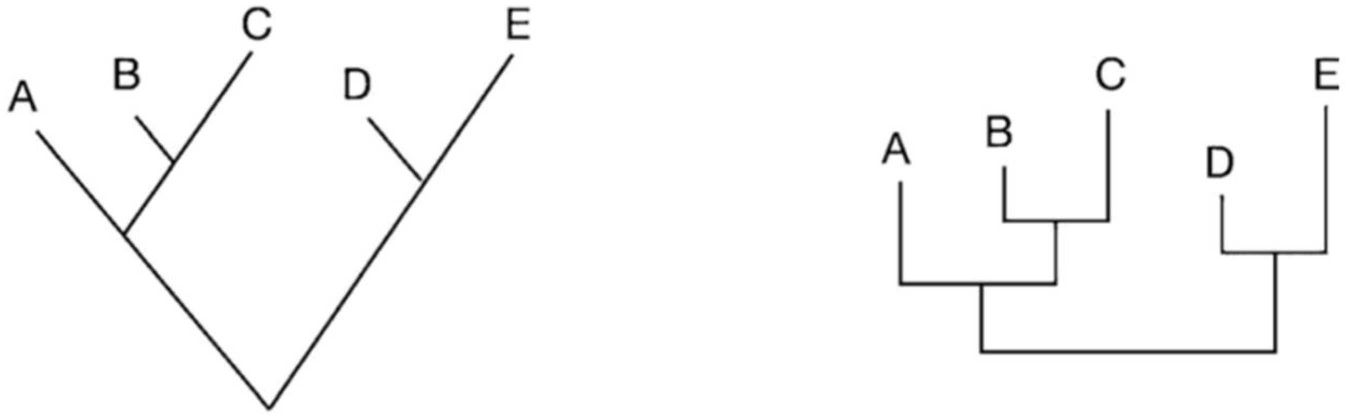
- Branch lengths have no phylogenetic meaning
- Topology of the tree shows the relative ordering of the taxa



Cladogram

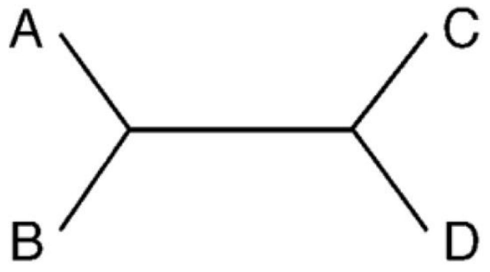
Phylogram = scaled trees:

- Branch lengths represent the amount of evolutionary divergence

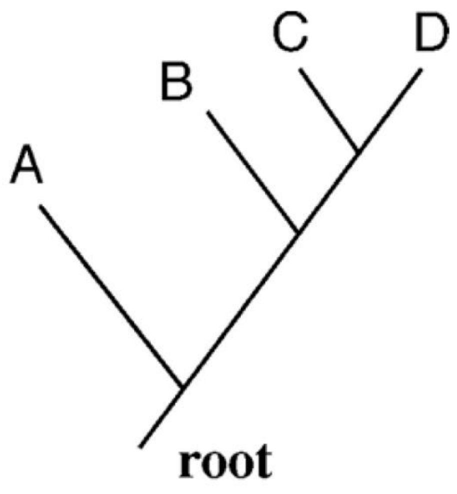


Phylogram

Unrooted *vs* rooted trees



Unrooted



Rooted

- Tree not assuming knowledge of a common ancestor, but only positioning the taxa to show their relative relationships = *unrooted* tree
  - ➢ No direction of an evolutionary path in an unrooted tree

- Tree in which all sequences under study have a common ancestor or root node from which a unique evolutionary path leads to all other nodes = *rooted* tree
  - ➢ More informative than an unrooted tree

Rooting a tree

Tree-building methods typically produce unrooted trees

Defining the root of a tree requires some kind of prior information:

1. Use of an *outgroup*, *i.e.* a sequence that is homologous to the sequences under consideration, but separated from those sequences at an early evolutionary time (*e.g.* a bird sequence as an outgroup for the phylogenetic analysis of mammals)
2. Midpoint rooting approach: midpoint of the two most divergent groups in a phylogram is assigned as the root (corresponds to a *molecular clock assumption*)

Molecular clock assumption

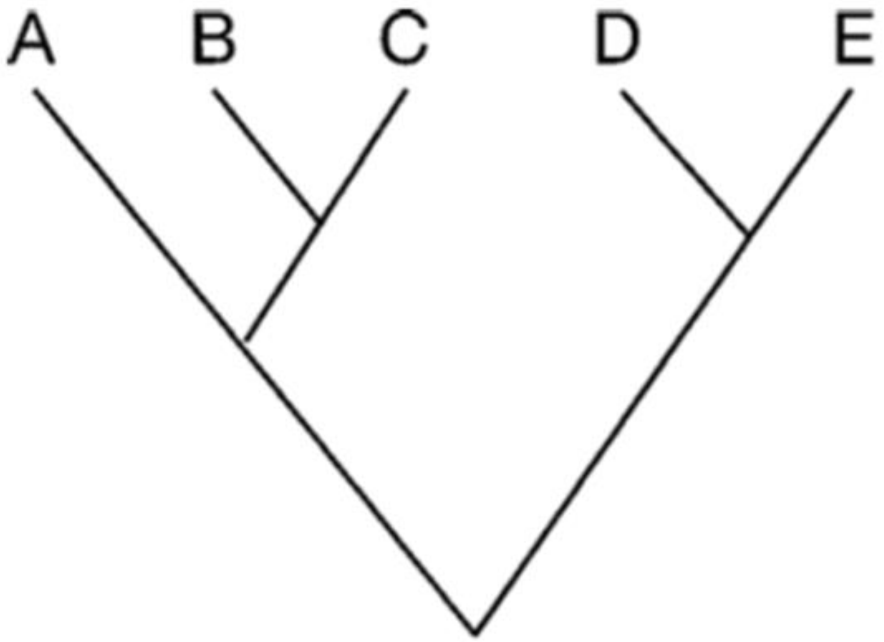
- Sequences evolve at a constant rate
- Number of accumulated mutations proportional to evolutionary time
  - Number of sequence differences can be used to estimate divergence time
- Rarely holds true in reality! Intricate modelling needed to "calibrate" the molecular clock if we really want to determine <u>when exactly</u> the nodes occurred

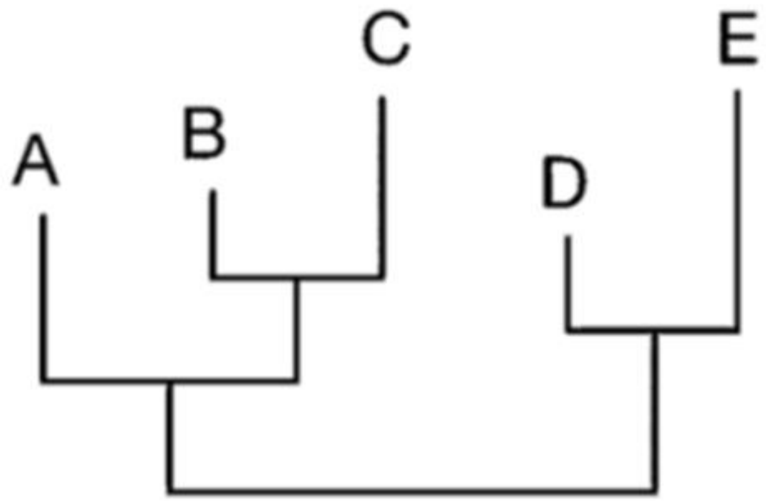
Tree representation file format

Newick format:

- Special text format to define a tree topology
- Trees are represented by taxa included in nested parentheses
- Each internal node is represented by a pair of parentheses that enclose all member of a monophyletic group separated by a comma
- For scaled trees the branch lengths is placed immediately after the name of the taxon separated by a colon

Tree representation file format





(((B,C),A),(D,E))

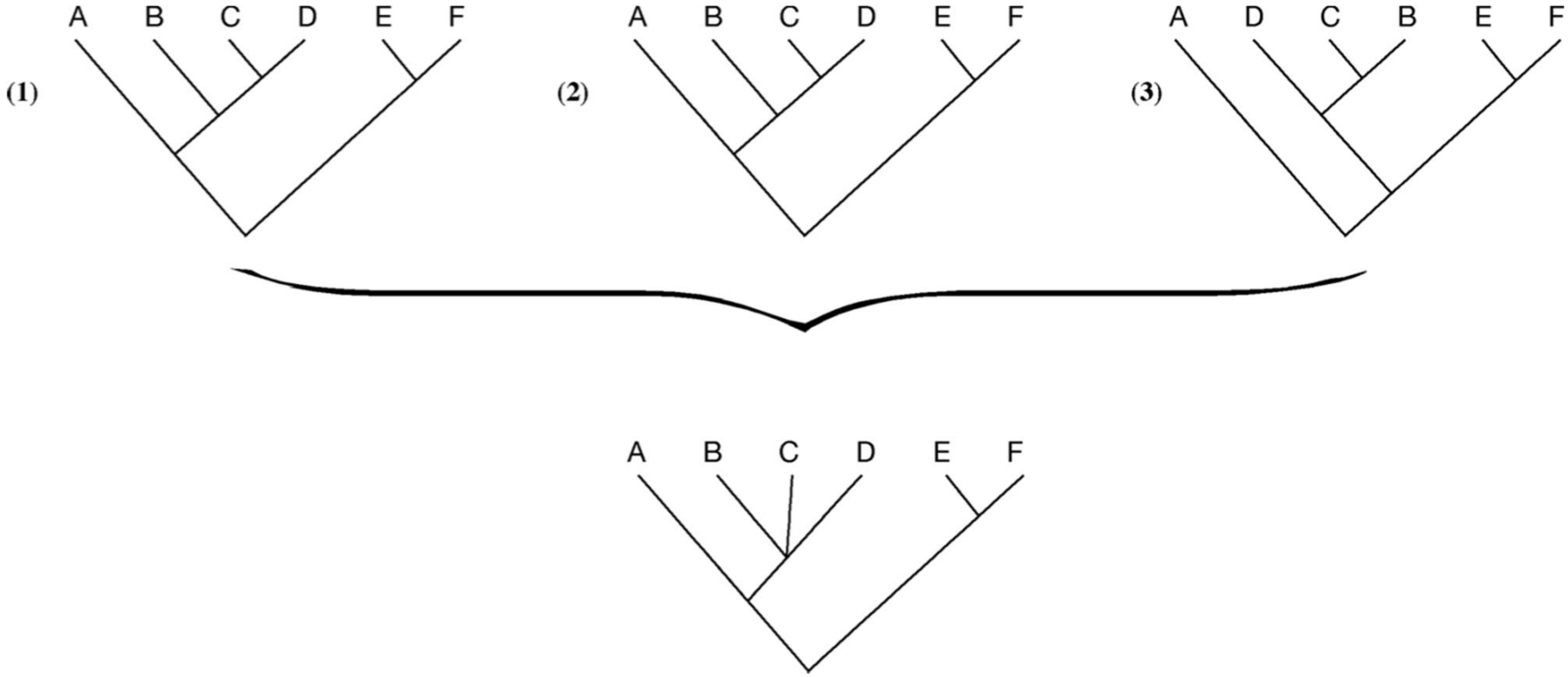
(((B:1,C:2),A:2),(D:1.2,E:2.5))

Newick format

Consensus trees

- Tree-building methods may result in several equally optimal trees
- Consensus tree can be built by showing the commonly resolved bifurcating portions and collapsing the ones that disagree among the trees (resulting in a polytomy)
  - Strict consensus tree: all conflicting nodes are collapsed into polytomies
  - Majority rule based consensus tree: conflicting nodes agreed by more than 50% are retained

Consensus trees



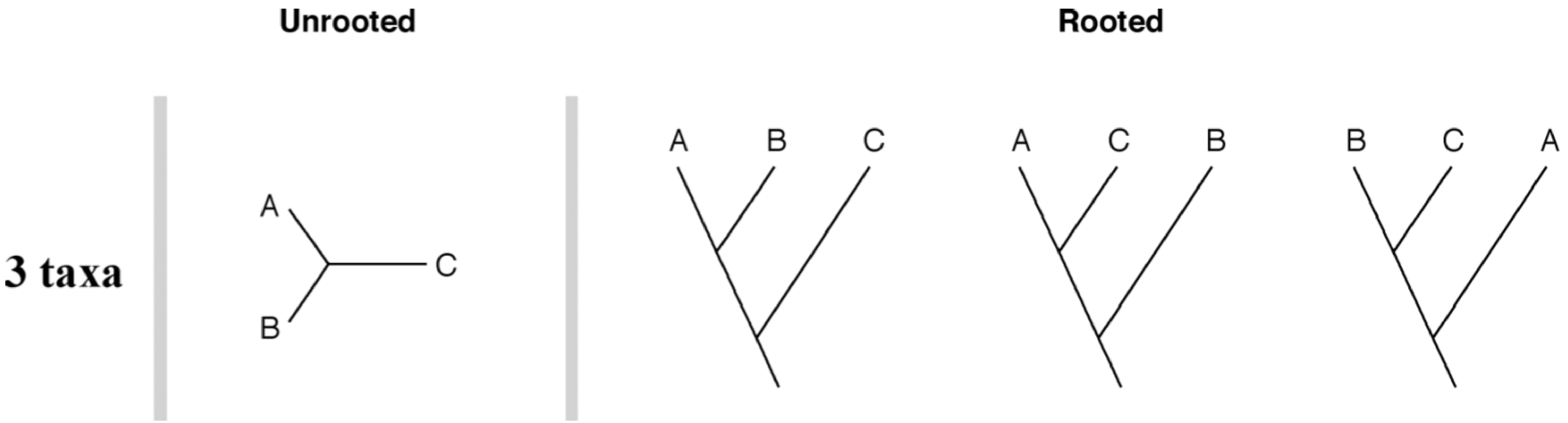
Consensus tree for trees (1), (2) and (3)

Why finding the best tree is difficult

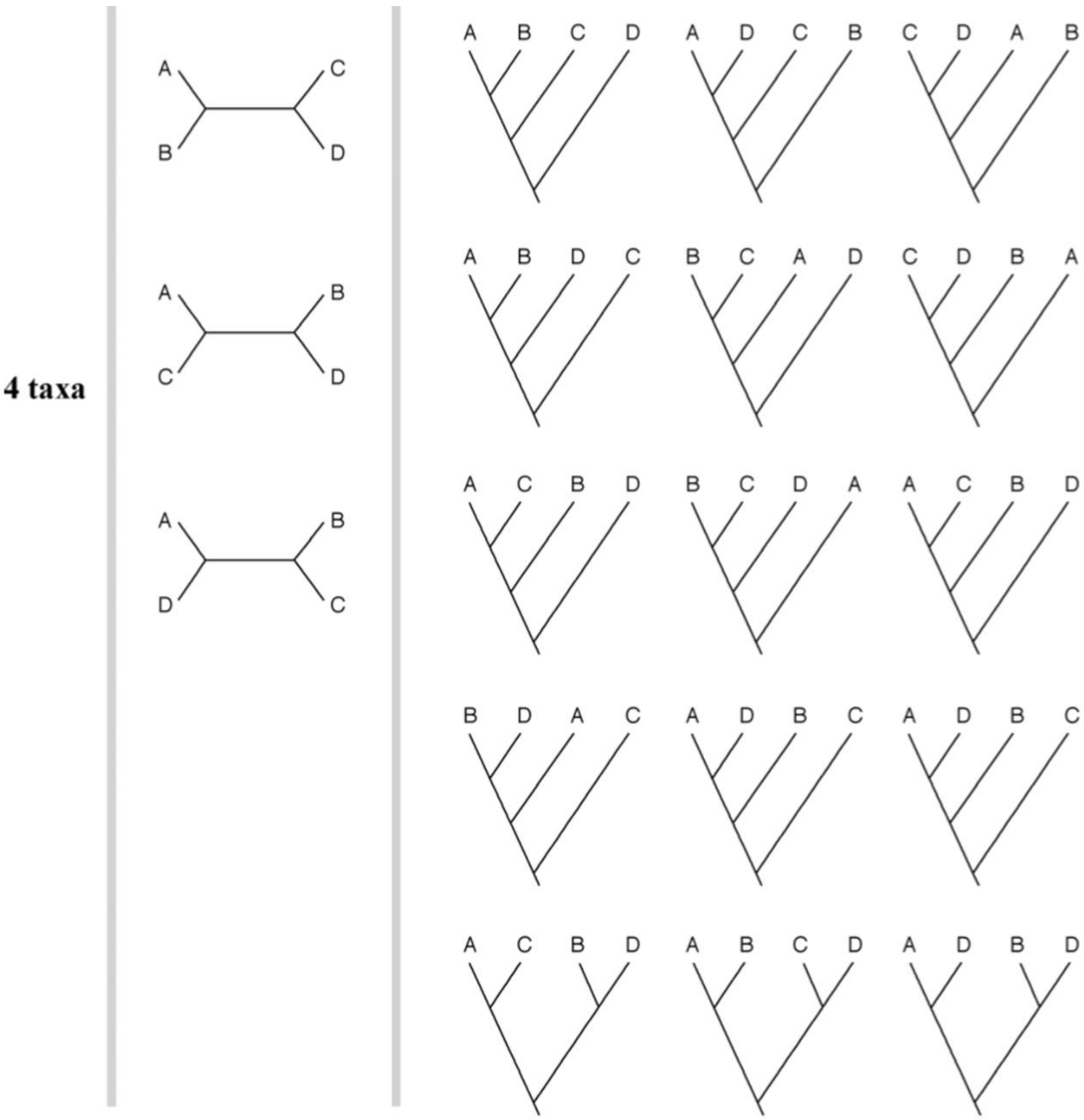
Number of potential tree topologies is enormously large even with a moderate number of taxa (n):

Nr. of rooted trees (N_R): $N_R = (2n - 3)! / 2^{n-2} (n - 2)!$

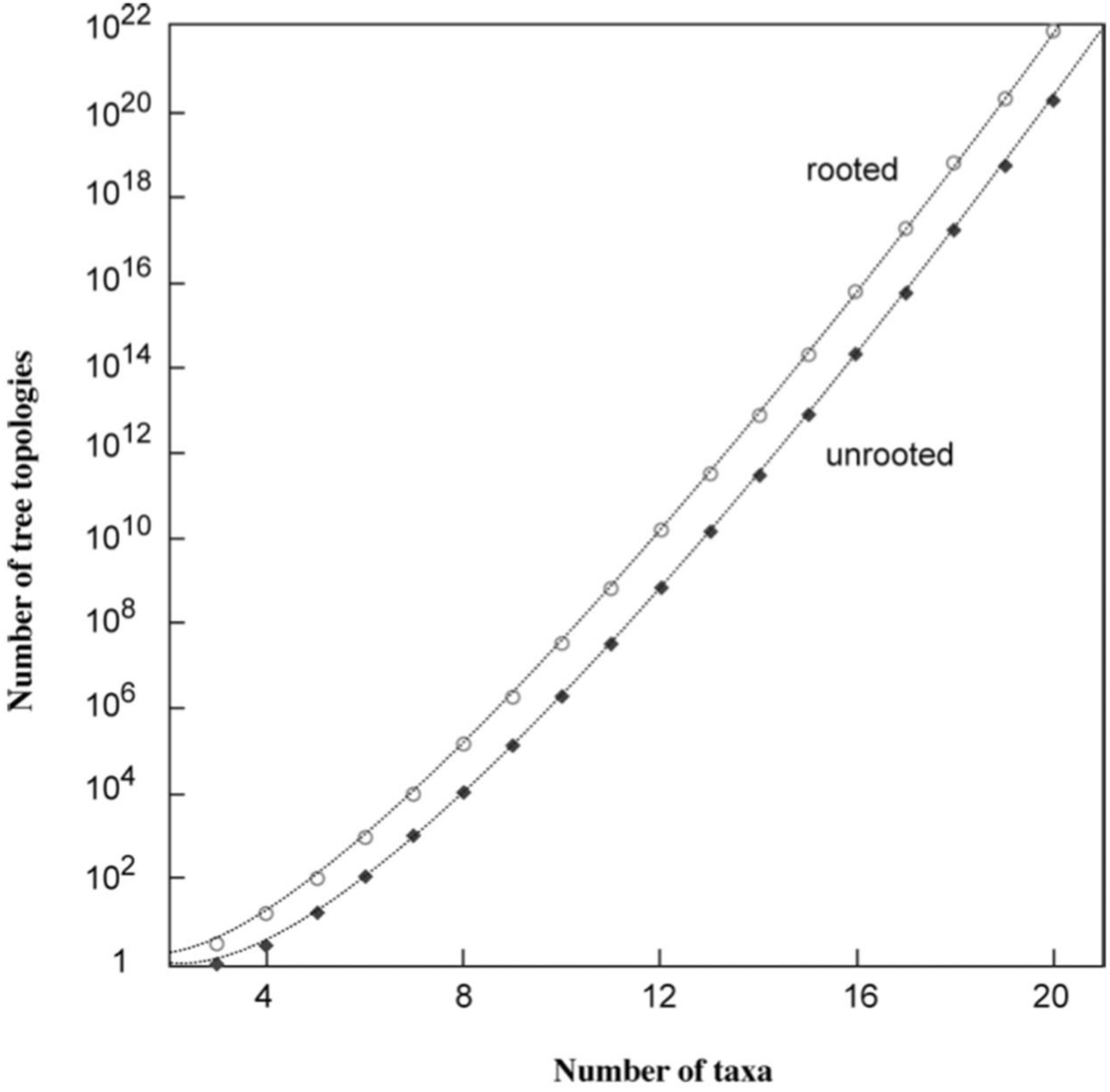
Nr. of unrooted trees (N_U): $N_U = (2n - 5)! / 2^{n-3} (n - 3)!$



Why finding the best tree is difficult



Why finding the best tree is difficult



Gene phylogeny vs species phylogeny

- One possible objective of building phylogenetic trees is to reconstruct the evolutionary history of <u>species</u>
- However, gene phylogeny only describes the evolution of the gene in question (or the protein encoded by it)
- One gene may evolve more or less rapidly than other genes in the genome or may have an entirely different evolutionary history, *e.g.* owing to horizontal gene transfer events
- To obtain species phylogeny, phylogenetic trees from a variety of gene families need to be constructed and compared (*i.e.* checked for consistency)

Phylogenetics: procedure

1. Choosing molecular markers
2. Performing multiple sequence alignment
3. Choosing a model of evolution
4. Determining a tree building method
5. Assessing tree reliability

Choice of molecular markers

- In principle, either nucleotide or protein sequence data can be used
- For studying very closely related organisms more rapidly evolving sequences should be used
  - *E.g.* noncoding regions of mitochondrial DNA for studying individuals within a population
- For widely divergent groups of organisms slowly evolving sequences are appropriate
  - *E.g.* ribosomal RNA or protein sequences

Protein *vs* nucleic acid sequences

In most cases, protein sequences are to be preferred over nucleotide sequences:

- Significant difference in evolutionary rates among the three nucleotide positions, plus positions are not strictly independent
- Preferential codon usage differs per organism, leading to bias in DNA sequences
- Protein sequences allow for a more sensitive alignment than DNA sequences (alignment errors are highly detrimental to tree construction!)

Protein *vs* nucleic acid sequences

In a few special cases nucleic acid sequences may be more informative:

- Very closely related sequences (DNA: more mutations to work with, because of *synonymous* mutations)
- Comparing *synonymous* and *non-synonymous* mutation rates:
    - If the non-synonymous mutation rate is higher than the synonymous rate, a protein is actively evolving and acquiring new functions (*positive selection*)
    - If the synonymous mutation rate is higher, protein function is critical and amino acid substitutions are not well-tolerated (*negative* or *purifying selection*)

Alignment

- This is the most critical step in phylogenetics as **only the correct alignment will produce the true phylogenetic tree!**
- Incorrect alignment leads to comparison of unrelated positions and therefore to systematic errors in the final tree, or even to a completely inaccurate tree
- State-of-the-art alignment programs have to be used
- Manual editing of alignment is often critical to ensure alignment quality (removal of ambiguously aligned regions!)
- If structural information is available, it can be used to optimise alignments

Multiple substitutions

Measurement of evolutionary distance by counting the number of mutations leads to <u>underestimation</u>

- Possible intermediate mutations not accounted for, *e.g.* A→C may in reality be the result of A→T→G→C
- Also: back- and parallel mutations can occur
- *Homoplasy* = multiple substitutions and convergence at individual sequence positions obscuring estimates of evolutionary distances
- Statistical models (corrections) are needed to avoid incorrect trees due to homoplasy

Choosing substitution models

Substitution models = evolutionary models, statistical models to correct homoplasy

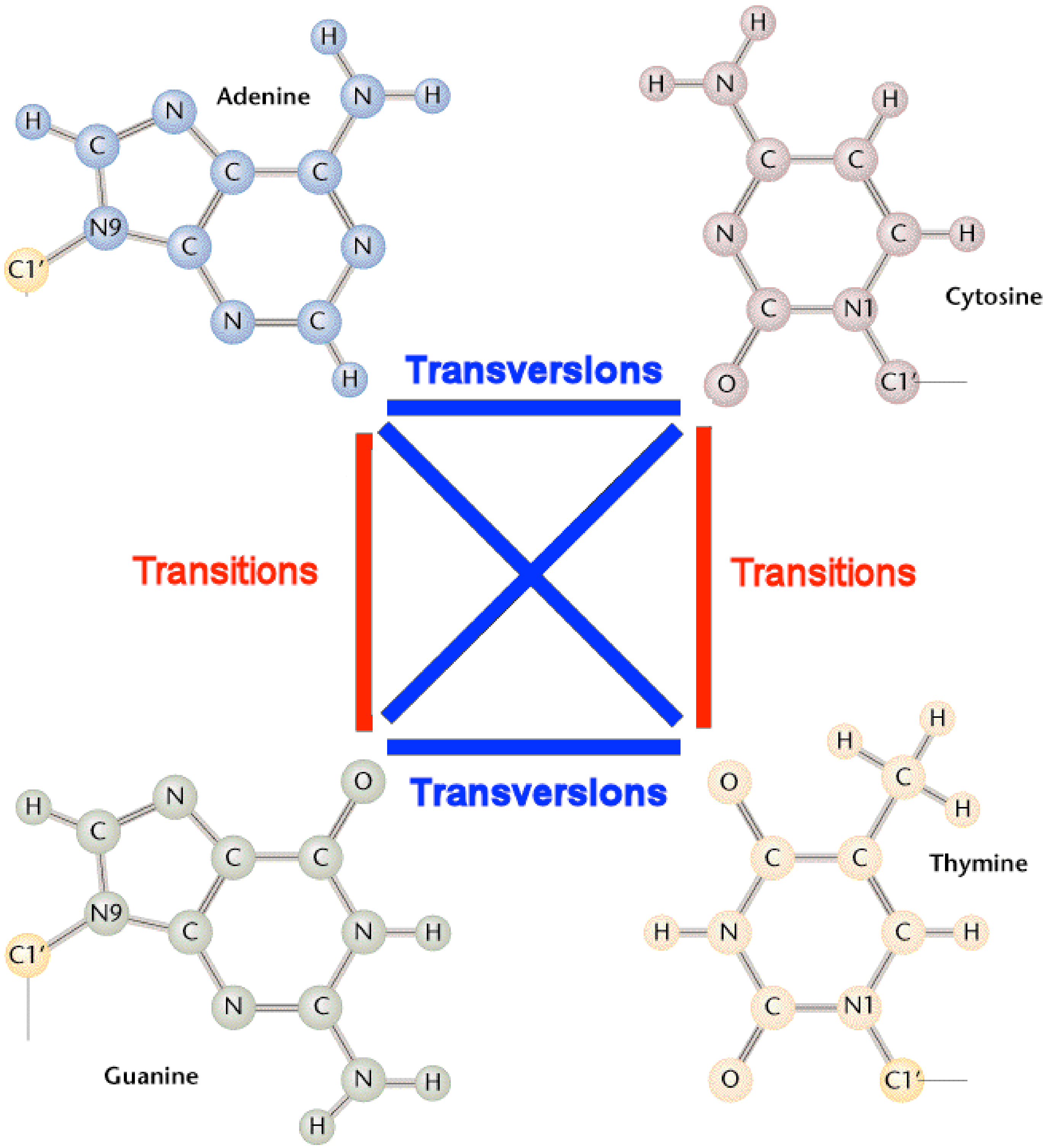
Only reasonably similar sequences can be used in phylogenetic comparisons:

- If there are too many multiple substitutions at a particular position, the position becomes "saturated"
- Saturated postions can not be correct by statistical models, *i.e.* true evolutionary distances cannot be derived

Jukes-Cantor model for DNA/RNA

- Assumes that all nucleotides are substituted with equal probability
- Evolutionary distance between sequences A and B with a substitution portion of p_{AB} is given by:
  $d_{AB} = – 3/4 \ln (1 – 4/3\ p_{AB})$
- *E.g.* sequences differing 30%, *i.e.* $p_{AB} = 0.3$:
  $d_{AB} = – 3/4 \ln (1 – 4/3 \times 0.3) = 0.38$
- $p_{AB} = 0.75 \rightarrow d_{AB} = \infty$
- Jukes-Cantor model can only handle reasonably closely related sequences

Transitions *vs* transversions



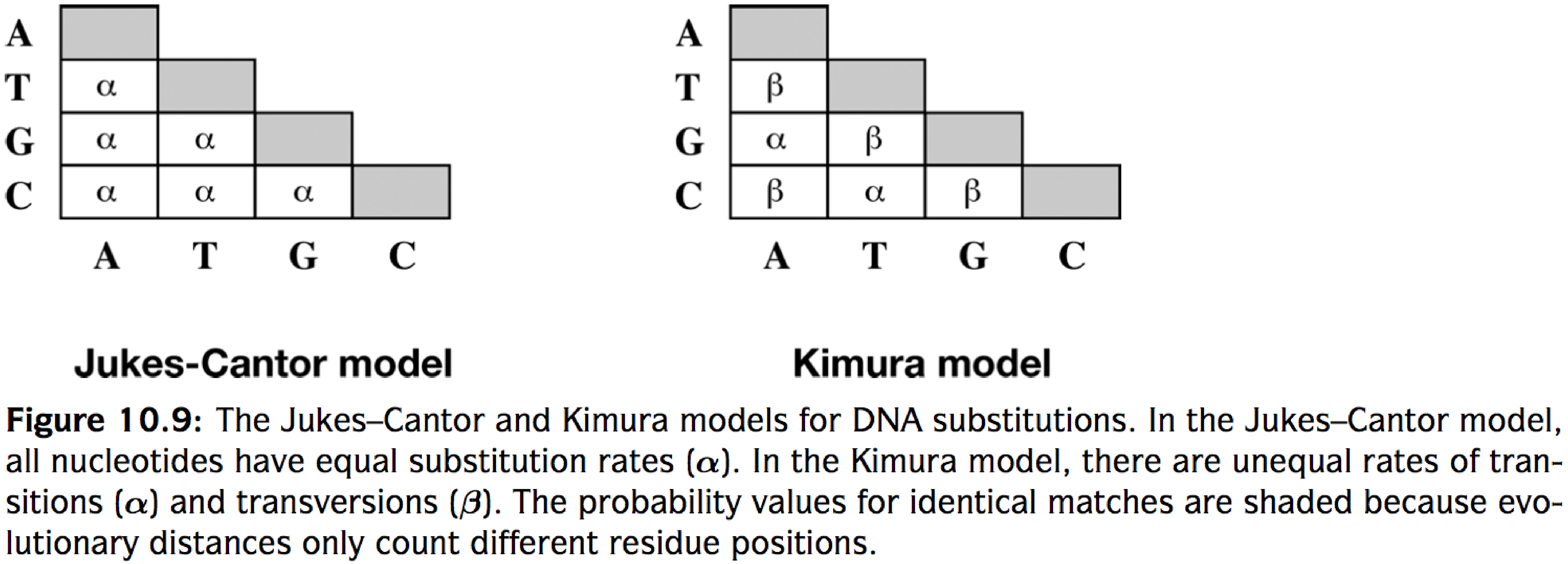
Transverslons

Transitions

Transitions

Transverslons

Jukes-Cantor *vs* Kimura model



A	▨			
T	α	▨		
G	α	α	▨	
C	α	α	α	▨
	A	T	G	C

Jukes-Cantor model

A	▨			
T	β	▨		
G	α	β	▨	
C	β	α	β	▨
	A	T	G	C

Kimura model

Figure 10.9: The Jukes–Cantor and Kimura models for DNA substitutions. In the Jukes–Cantor model, all nucleotides have equal substitution rates (α). In the Kimura model, there are unequal rates of transitions (α) and transversions (β). The probability values for identical matches are shaded because evolutionary distances only count different residue positions.

Kimura model for DNA/RNA

- Assumes that transitions occur more frequently than transversions:
  $d_{AB} = -\, 1/2 \ln (1 - 2\, p_{ti} - p_{tv}) - 1/4 \ln (1 - 2\, p_{tv})$
  p_{ti} = observed frequency of transition
  p_{tv} = observed frequency of transversion

- *E.g.* sequences A and B differing by 30%, 20% due to transitions, 10% due to transversions:
  $d_{AB} = -\, 1/2 \ln (1 - 2 \times 0.2 - 0.1) - 1/4 \ln (1 - 2 \times 0.1) = 0.40$

Substitution models for proteins

- For protein sequences amino acid substitution matrices as PAM or JTT can be used, as they already take multiple substitutions into account
- As simple alternative a protein equivalent of the Kimura model can be used:
  $d = -\ln(1 - \mathrm{p} - 0.2\,p^2)$

Among-site variations

Substitution models (wrongly) assume that different positions in a sequence are evolving at the same rate:

- In DNA sequences, the 3rd codon position mutates faster than the other two
- For protein sequences some amino acid positions change less often than others because of functional constraints
- = *among-site rate heterogeneity*:
  causes artifacts in tree construction

Modelling among-site variation: gamma distribution

Distribution of positions in a sequence with more invariant rates and more variable rates can be described by a site-dependent γ-distribution

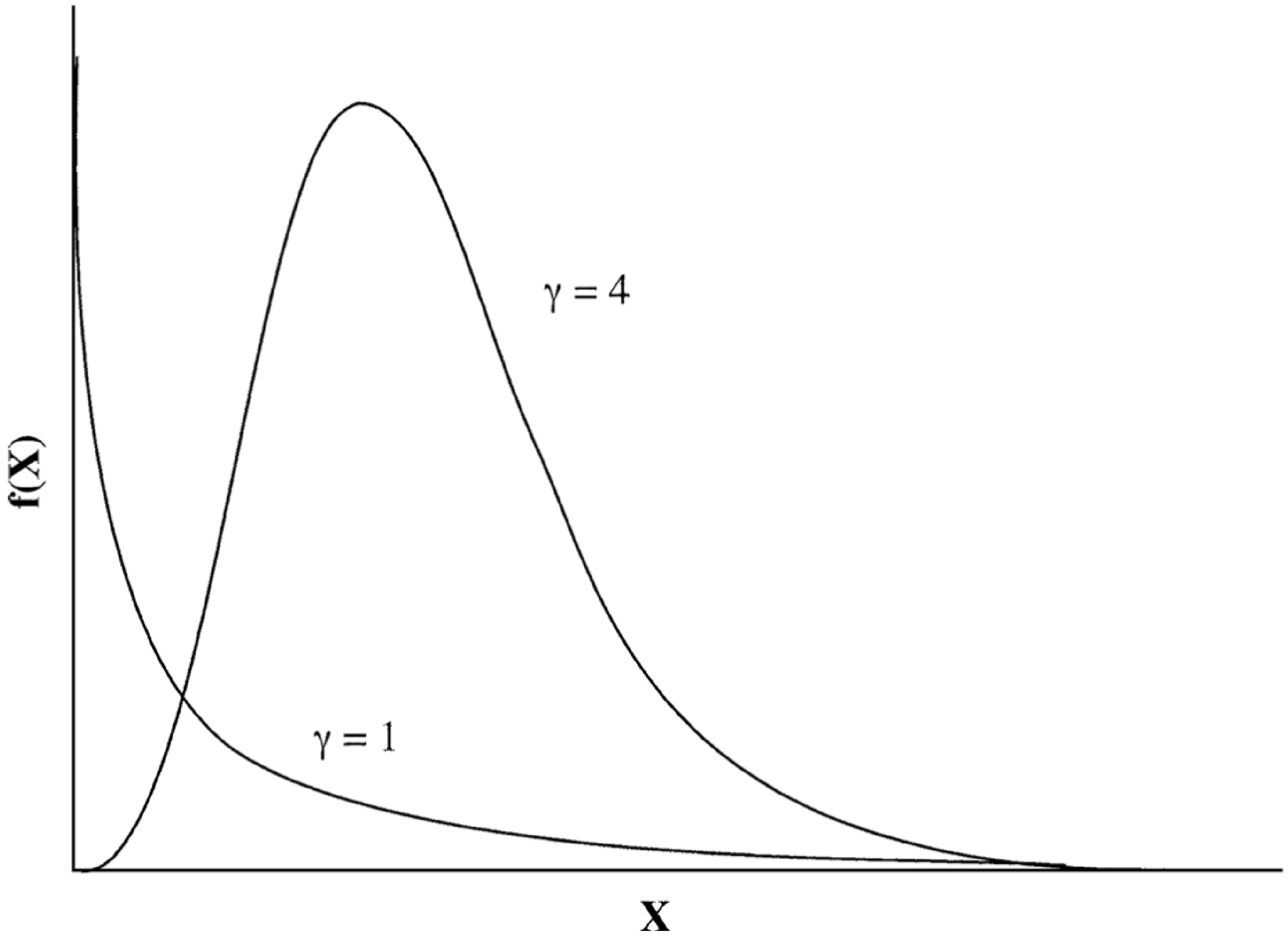


Figure 10.10: Probability curves of γ distribution. The mathematical function of the distribution is $f(x) + (x^{\gamma-1}\ e^{-x}) / \Gamma(\gamma)$. The curves assume different shapes depending on the γ-shape parameter (γ).

Modelling among-site variation: gamma distribution

- Adjusted Jukes-Cantor model:
  $d_{AB} = 3/4\ \alpha\ (\ (\ 1 - 4/3\ p_{AB}\ )^{-1/\alpha} - 1$
  α: correction factor derived from γ correction
- Adjusted Kimura model:
  $d_{AB} = \alpha/2\ (\ (\ 1 - 2\ p_{ti} - p_{tv}\ )^{-1/\alpha} - 1/2\ (\ 1 - 2\ p_{tv}\ )^{-1/\alpha} - 1/2\ )$