In the Name of God, the Merciful, the Compassionate

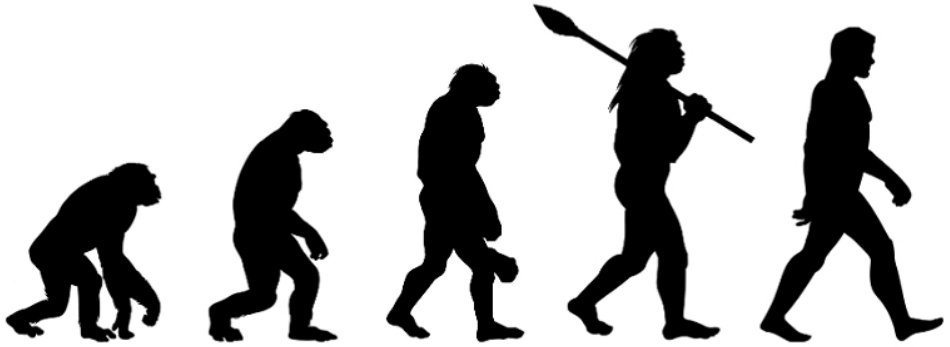
Introduction to Bioinformatics
09 - Phylogenetics Basics

Instructor: Hossein Zeinali

Amirkabir University of Technology

Molecular Evolution

- What is evolution?
  - In the *biological context*, evolution can be defined as the *development* of a biological form from *other preexisting* forms or *its origin* to the current existing form through **natural selections** and **modifications**.
- The driving force behind evolution is natural selection
  - **Unfit forms** are eliminated through *changes of environmental conditions or sexual selection* so that only the **fittest** are selected.
- The underlying mechanism of evolution is genetic *mutations* that occur spontaneously.
  - Mutations provide the biological diversity within a population.



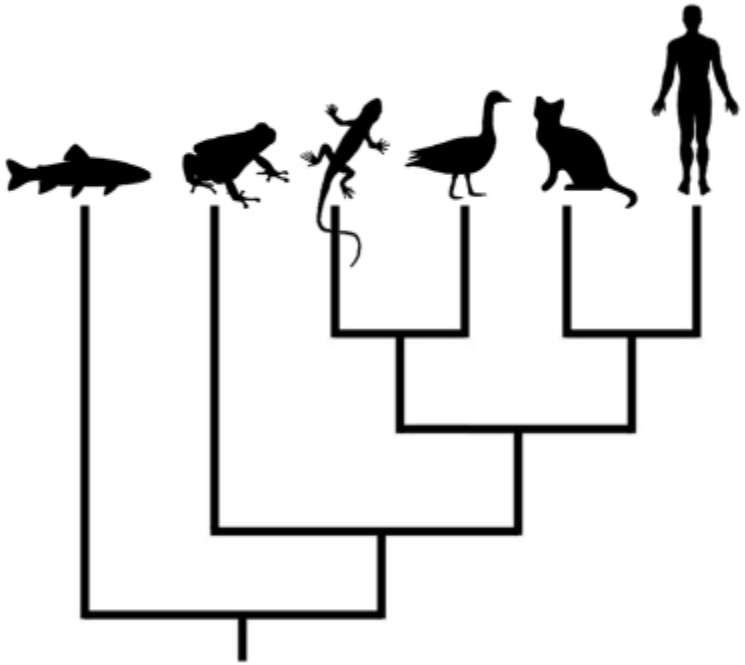
Natural Selection in Other Words

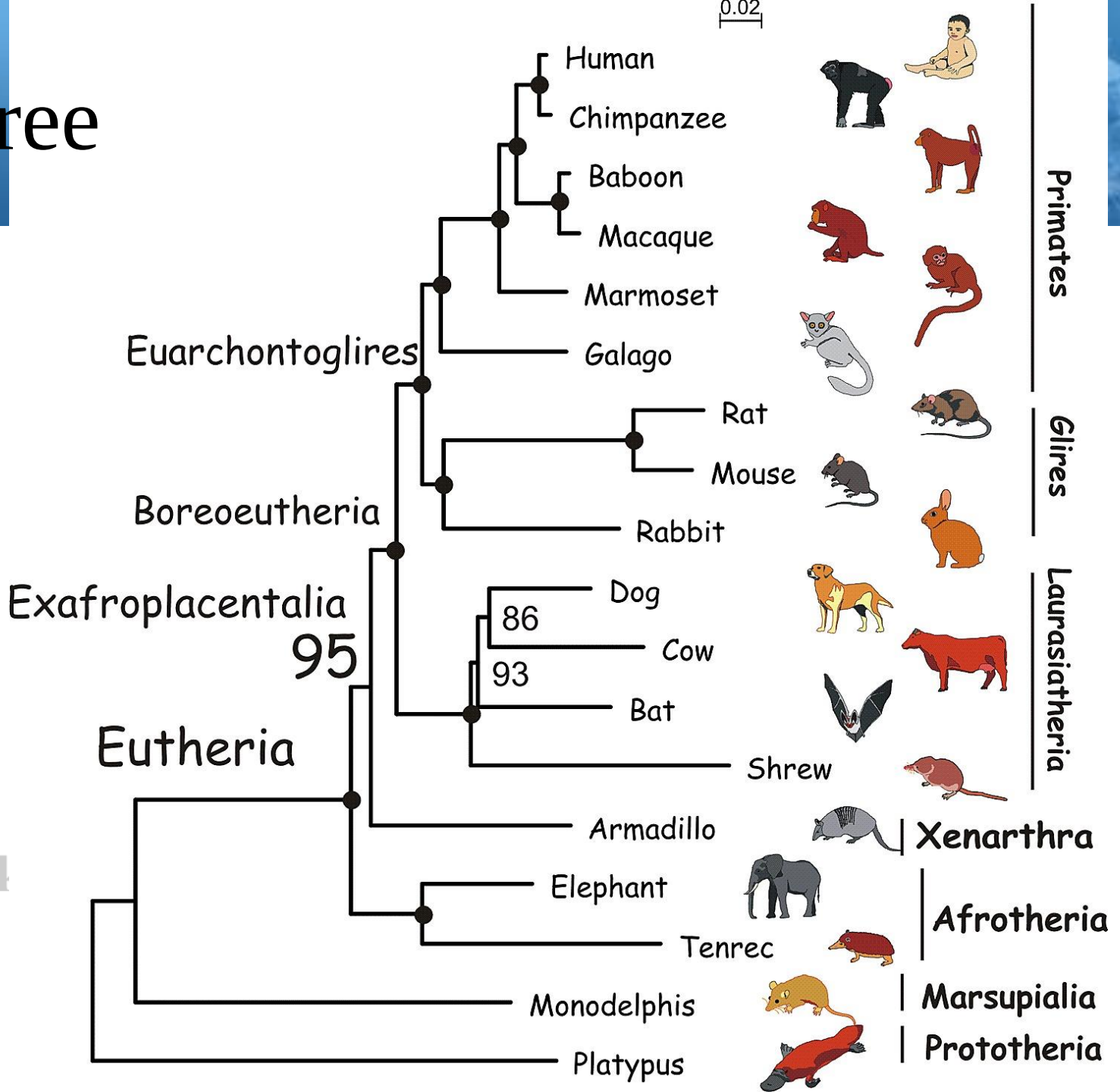
- Species can produce more offspring than the environment can support.
- This leads to competition for resources.
- Genetic variations exist in a population that give some individuals an advantage, others a disadvantage, leading to differential reproductive success.

Phylogenetics

- ***Phylogenetics*** is the study of the evolutionary history of living organisms using treelike diagrams to represent pedigrees of these organisms.
- ***Phylogeny***: The tree branching patterns representing the evolutionary divergence.
- Similarities and differences seen in a multiple sequence alignment are easier to make sense in a phylogenetic tree

Phylogenetics Tree





Data Used in Phylogenetics

- Fossil records
  - Contain morphological information about ancestors of current species and the timeline of divergence
  - Limitations - not available for all species in all areas, morphology determined by multiple genetic factors (e.g. abundance, habitat, geographic range) which make it ambiguous and biased, fossils for microorganisms are especially rare
- Molecular data - DNA and protein sequences
  - Molecular fossils - genes are the medium for recording the accumulated mutations
  - Advantages - lots of data, easy to obtain
  - Limitations - can be difficult to get sequences from extinct species
- Physical, behavior, and developmental characteristics can also be used in phylogenetics

Molecular Phylogenetics

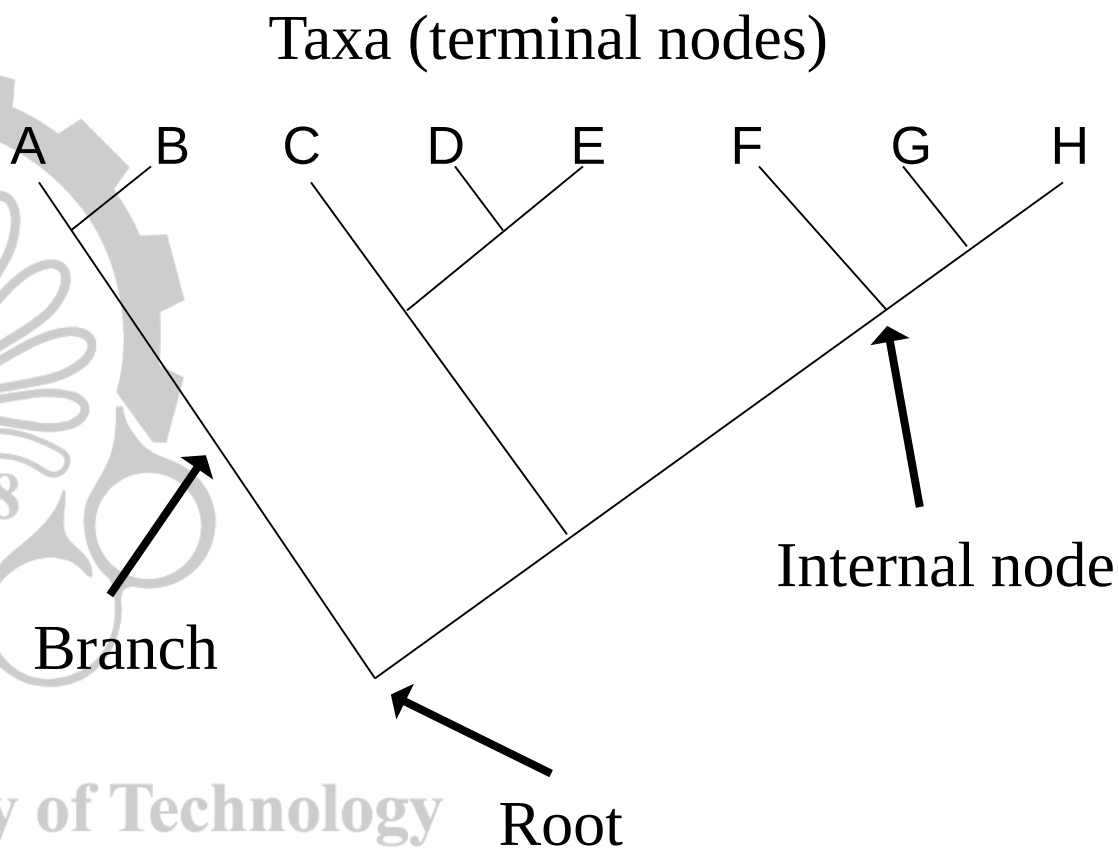
- *Molecular phylogenetics* is the study of evolutionary relationships of genes and other biological macromolecules by *analyzing mutations* in their sequences
  - Developing hypotheses about the evolutionary relatedness of the biomolecules.
- Sequence similarity can be used to infer evolutionary relationships

Assumptions in Molecular Phylogenetics

- To use molecular data to reconstruct evolutionary history requires making a number of reasonable assumptions.
- 1) Sequences used in phylogenetic construction are homologous, i.e. share a common ancestor (origin) and subsequently diverged through time.
- 2) Phylogenetic divergence is bifurcating, i.e. parent branch splits into two daughter branches at any given point.
- 3) Each position in a sequence evolved independently.
- 4) Molecular clock – sequences evolve at constant rates
  - The amount of accumulated mutations is proportional to evolutionary time.
- The variability among sequences is sufficiently informative for constructing unambiguous phylogenetic trees.

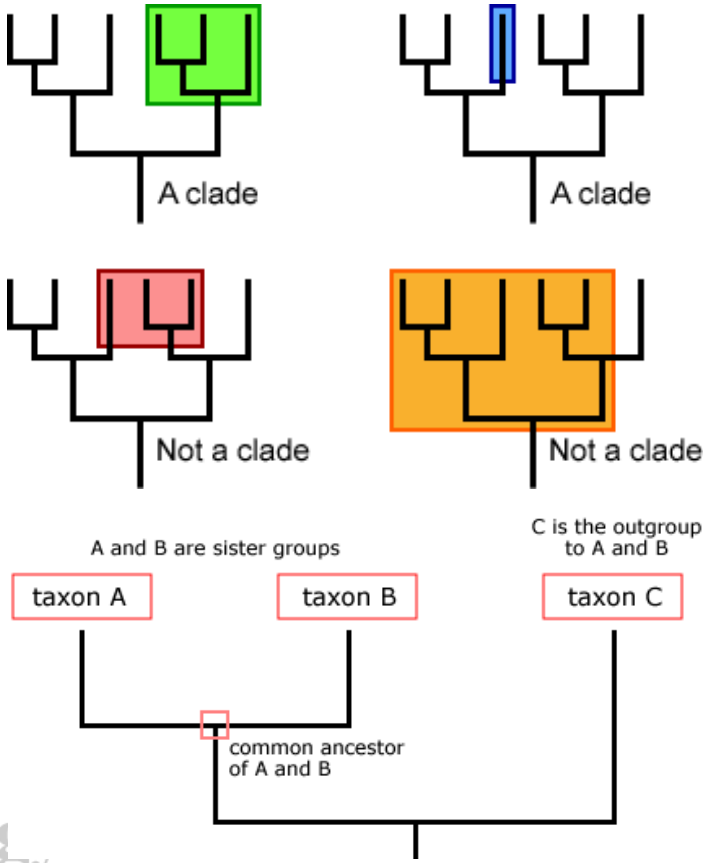
Terminology

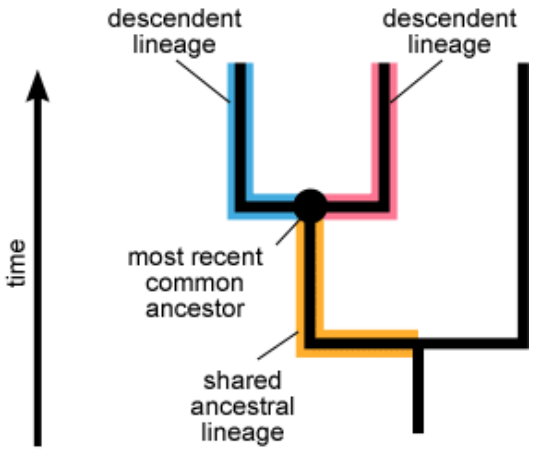
- The lines in the tree are called ***branches***.
- At the tips of the branches are present-day species or sequences known as ***taxa*** (the singular form is ***taxon***)
- The connecting point where two adjacent branches join is called a ***node***.
  - inferred ancestor of extant taxa
- The bifurcating point at the very bottom of the tree is the ***root node***.
  - the common ancestor of all members of the tree.

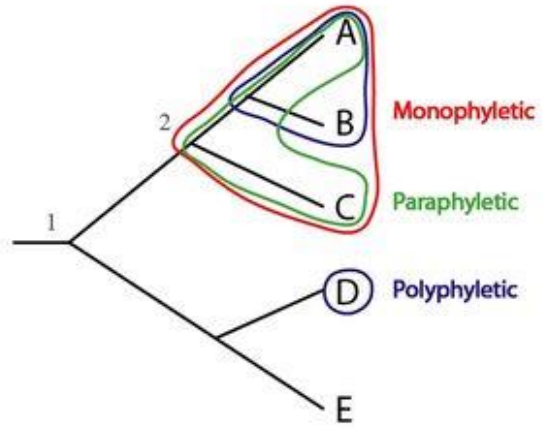


Terminology (Cont.)

- ***Clade*** (*monophyletic group*)= group of taxa descended from a *common ancestor*
- ***Sister taxa*** = two taxa which share a unique common ancestor not shared by any other taxa
- ***Lineage*** = branch path depicting ancestor-descendant relationship
- ***Paraphyletic group*** = group of taxa that share more than one closest common ancestor

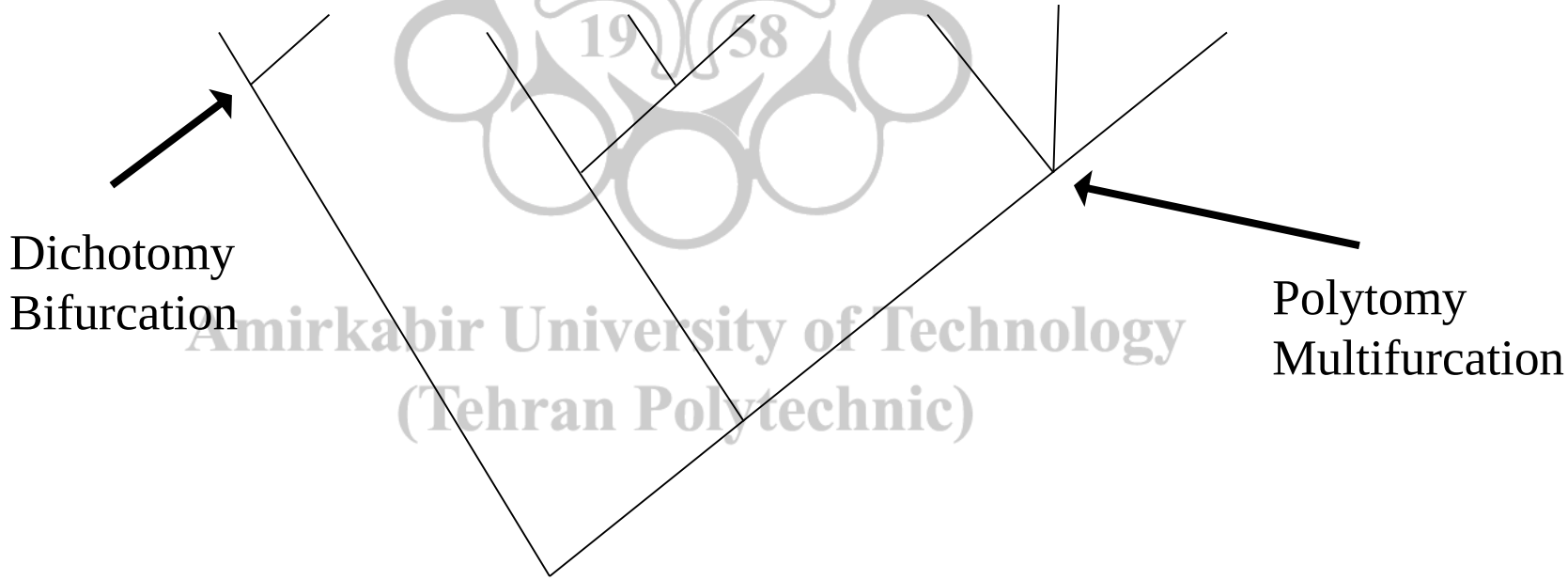






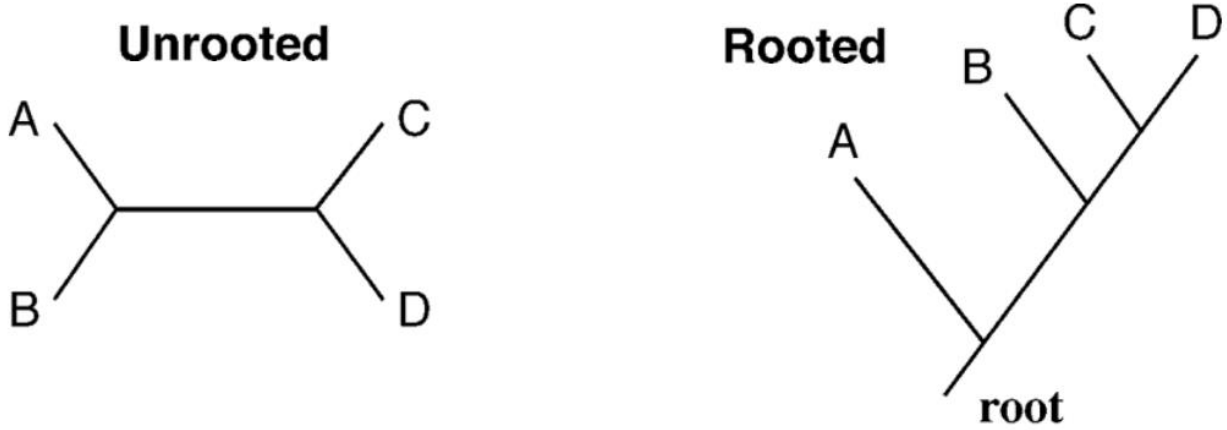
Tree Topology

- The branching pattern in a tree is called *tree topology.*
- When all branches bifurcate on a phylogenetic tree, it is referred to as *dichotomy.*
- If a branch has more than two descendants, it is referred to as *multifurcating node.*



Rooted vs. Unrooted Trees

- A phylogenetic tree can be either rooted or unrooted.
- An *unrooted phylogenetic tree* does not assume knowledge of a common ancestor
  - Only shows the taxa their relative relationships.
  - There is no direction of an evolutionary path.
- In a *rooted tree*, all the sequences under study have a common ancestor or root node.
  - Is more informative than an unrooted one.

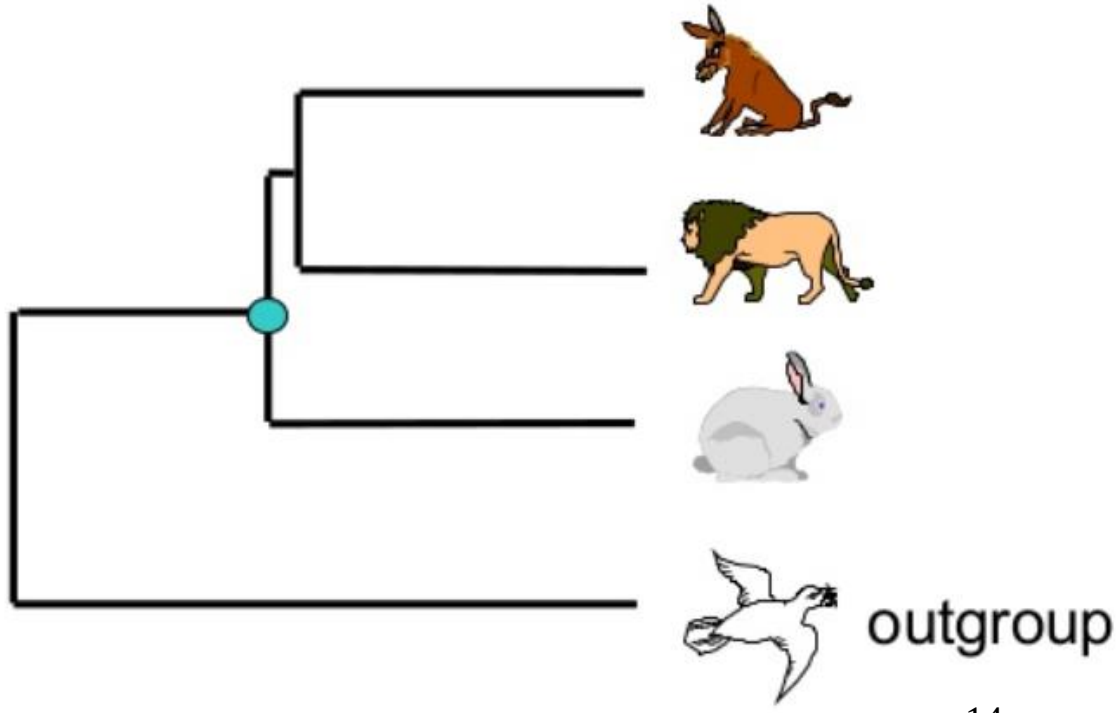


Rooted vs. Unrooted Trees (Cont.)

- Can convert between unrooted and rooted, but first need to determine where the root is.
  - The root of the tree is not known; the common ancestor is already extinct.
- Two ways to define the root:
  - Use an *outgroup*
  - Midpoint rooting – midpoint of the two most divergent groups is assigned to be the root

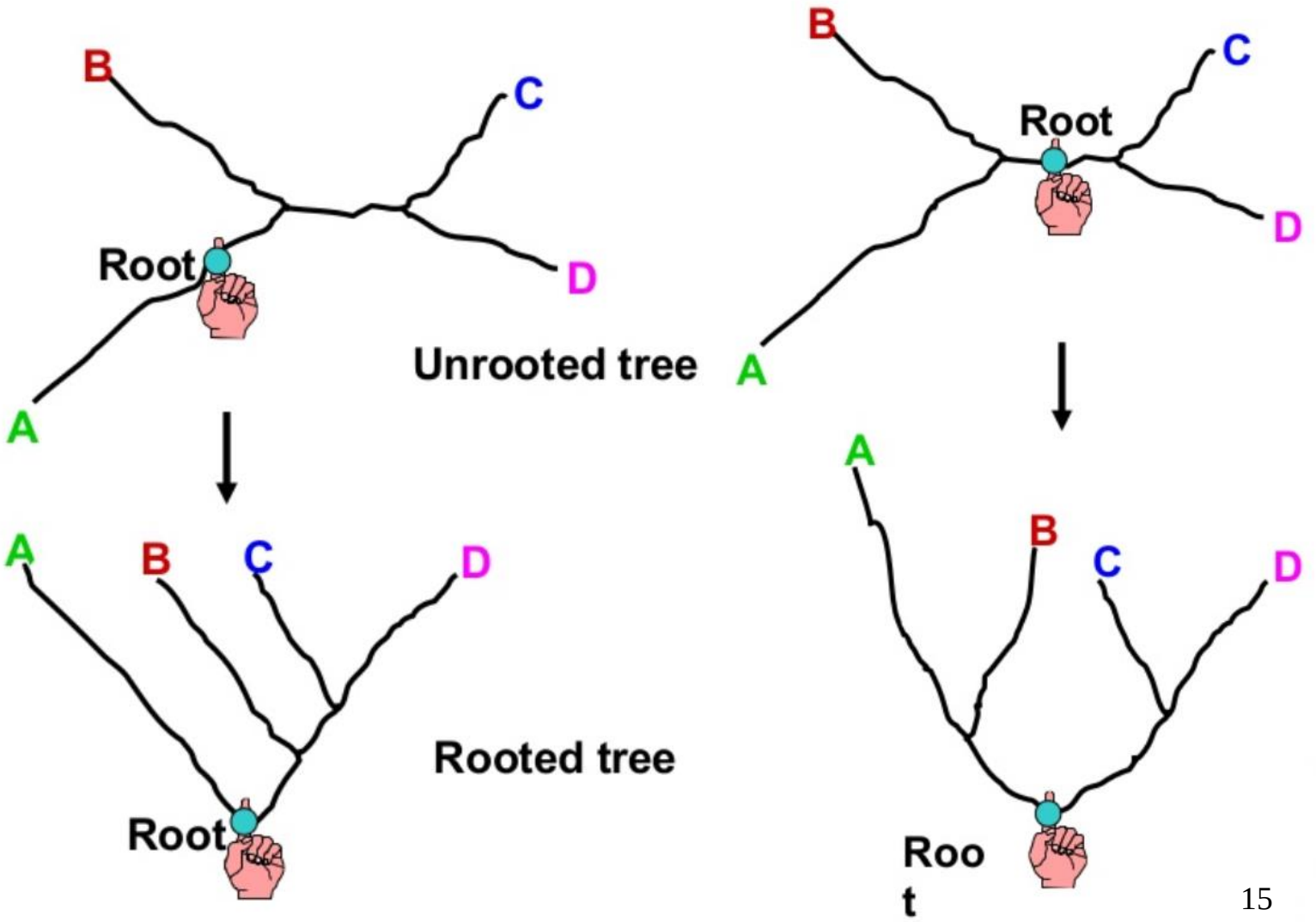
Outgroups

- Outgroup is a sequence that is homologous to the sequences under consideration
- Outgroup is related to the sequences being studied, but is more distantly related.
- Must be distinct from the ingroup, but not too distant.
  - If outgroup is too distantly related, it can lead to errors in tree construction.
- Trick is to find the closest related sequence that is removed from the ingroup



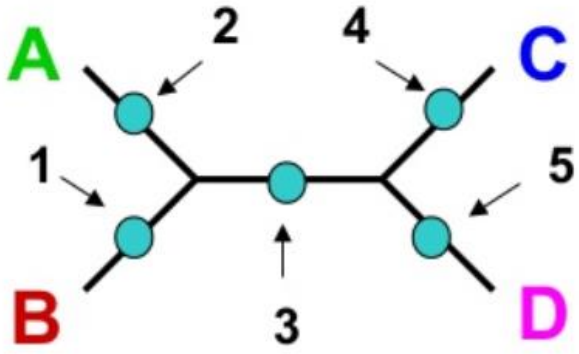
Midpoint Rooting

- Assumes that divergence from root to tips for both branches is equal and follows the “molecular clock” hypothesis.

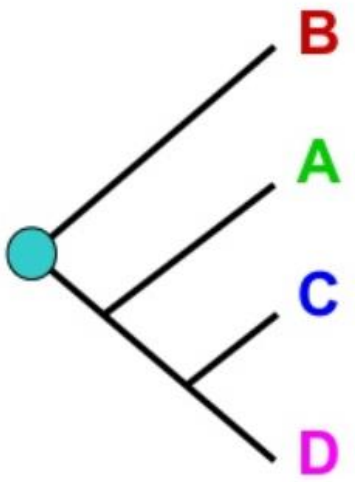


Midpoint Rooting (Cont.)

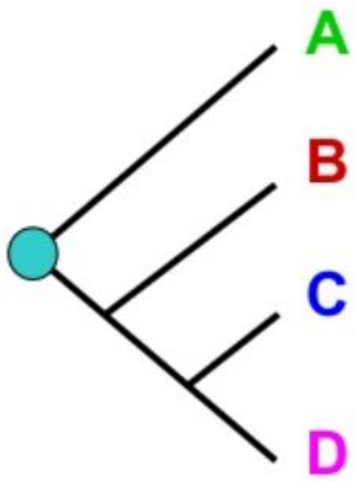
The unrooted tree 1:



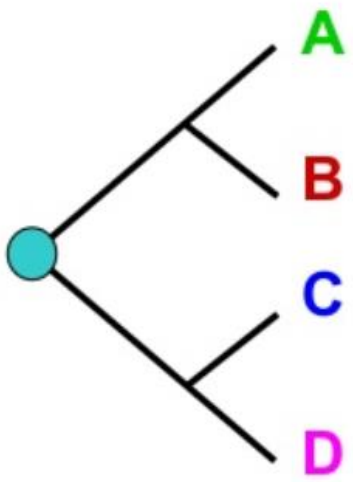
Rooted tree 1a



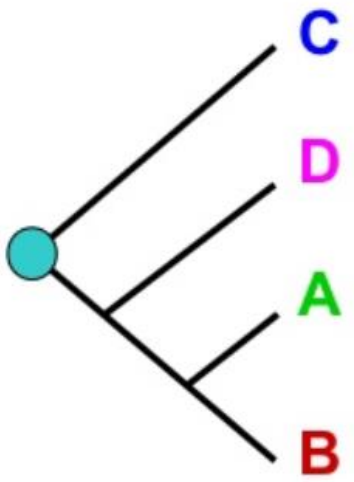
Rooted tree 1b



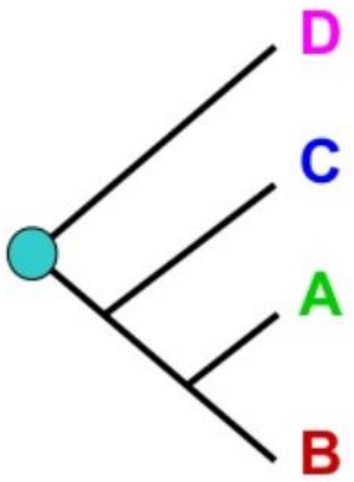
Rooted tree 1c



Rooted tree 1d



Rooted tree 1e

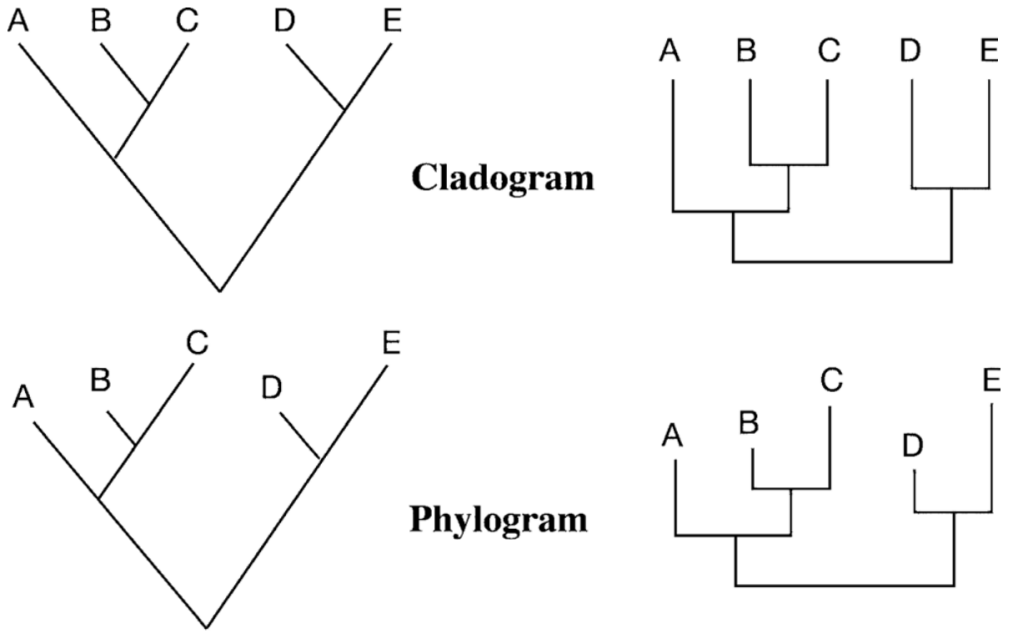


Gene Phylogeny vs. Species Phylogeny

- *Gene phylogeny*:
  - Inferred from a gene or protein sequence
  - Only describes the evolution of that particular gene or encoded protein
- Species evolution is the result of mutations in the entire genome
- Your gene may have evolved differently than other genes in the genome
- To obtain a *species phylogeny*, we need to use a variety of gene families to construct the tree

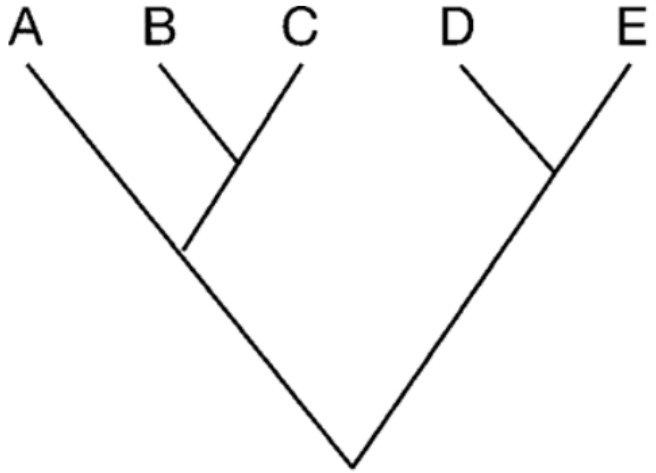
Forms of Tree Representation

- The topology of branches in a tree defines the relationships between the taxa.
  - The branches of a tree can freely rotate without changing the relationships
- In a ***phylogram***, the branch lengths represent the amount of evolutionary divergence.
- In a ***cladogram***, however, the external taxa line up neatly in a row or column.
  - Branch lengths are meaningless

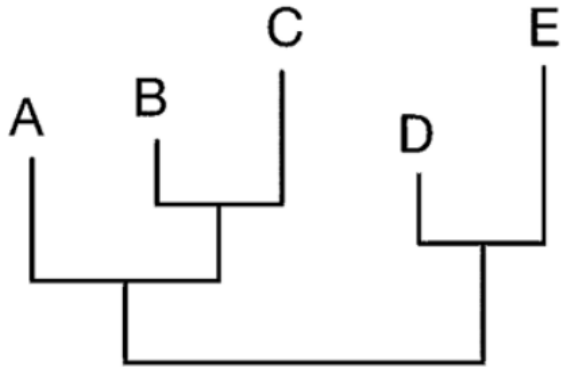


Newick format

- *Newick format* is developed to provide information of tree topology to computer programs without having to draw the tree itself. Is a special text format.
- Each internal node is represented by a pair of parentheses
- The branch lengths in arbitrary units are placed immediately after the name of the taxon.



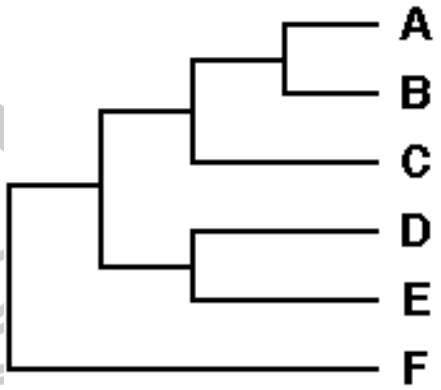
(((B,C),A),(D,E))



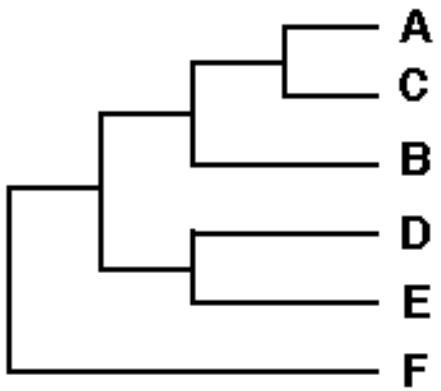
(((B:1,C:2),A:2),(D:1.2,E:2.5))

Consensus Trees

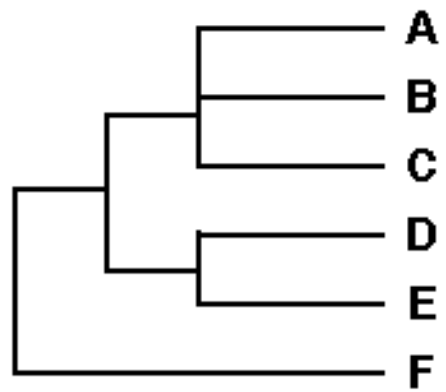
- Multiple trees that are equally optimal
- Strict consensus tree:
  - build consensus tree by collapsing disagreements into a single node
- Majority rule:
  - Nodes that agree by more than 50% of the nodes are retained whereas the remaining nodes are collapsed into multifurcation



TREE A

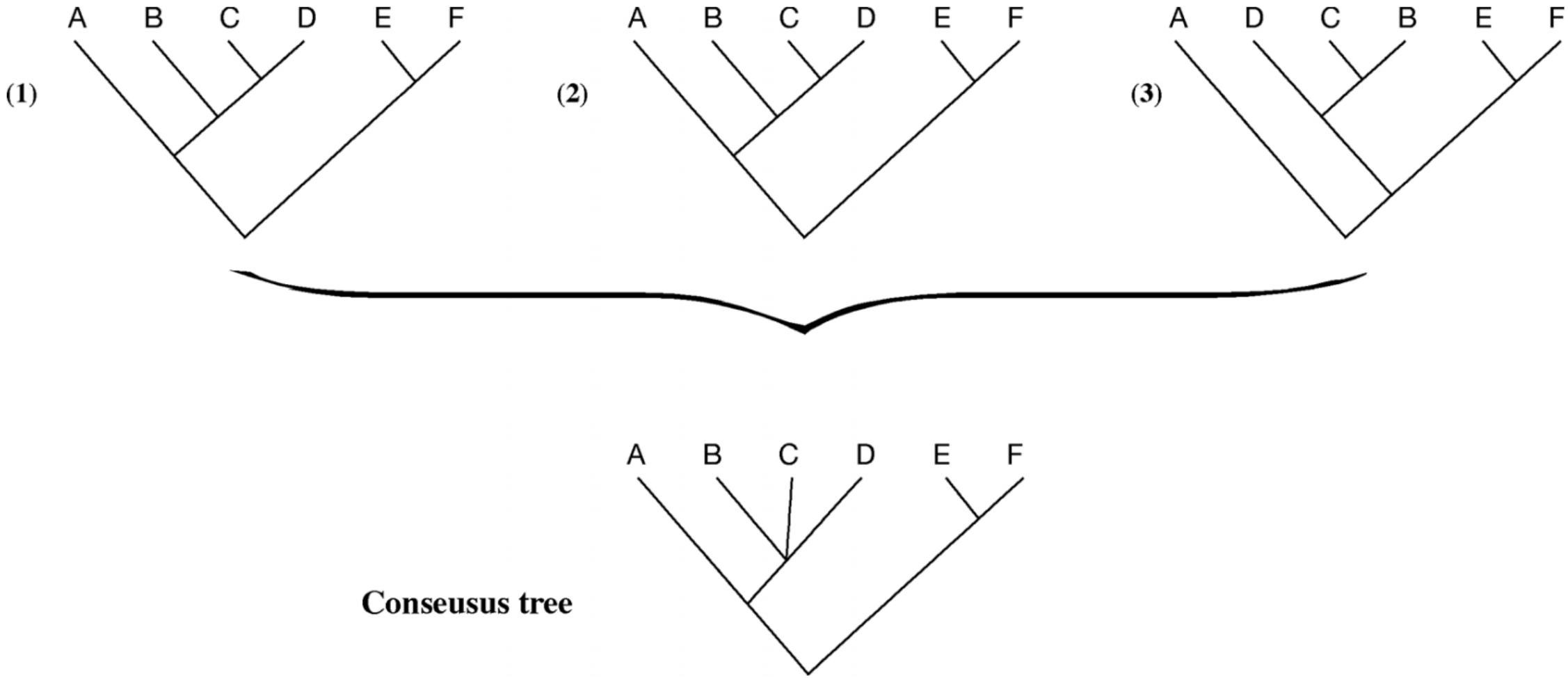


TREE B



CONSENSUS TREE

Consensus Trees (Cont.)



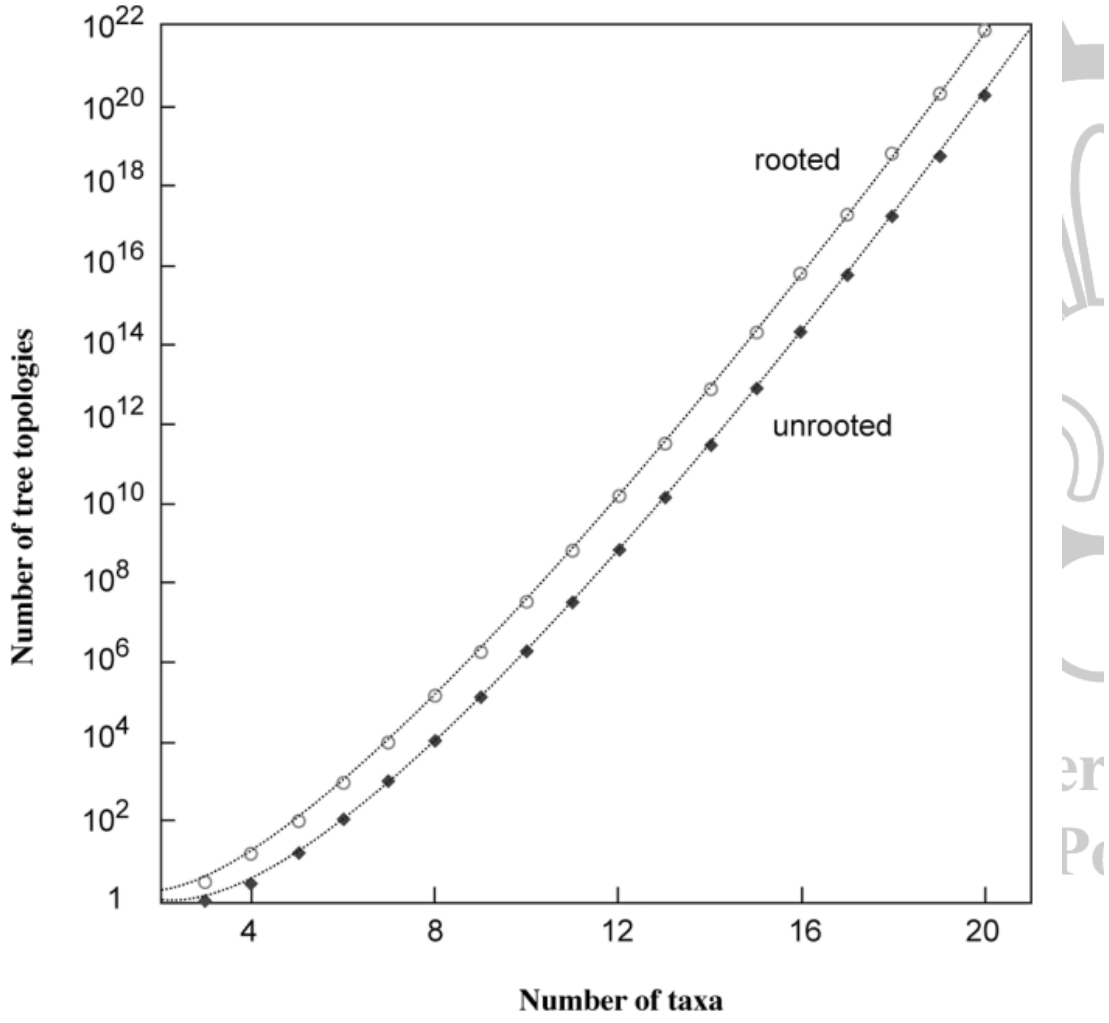
Why Finding a True Tree is Difficult?

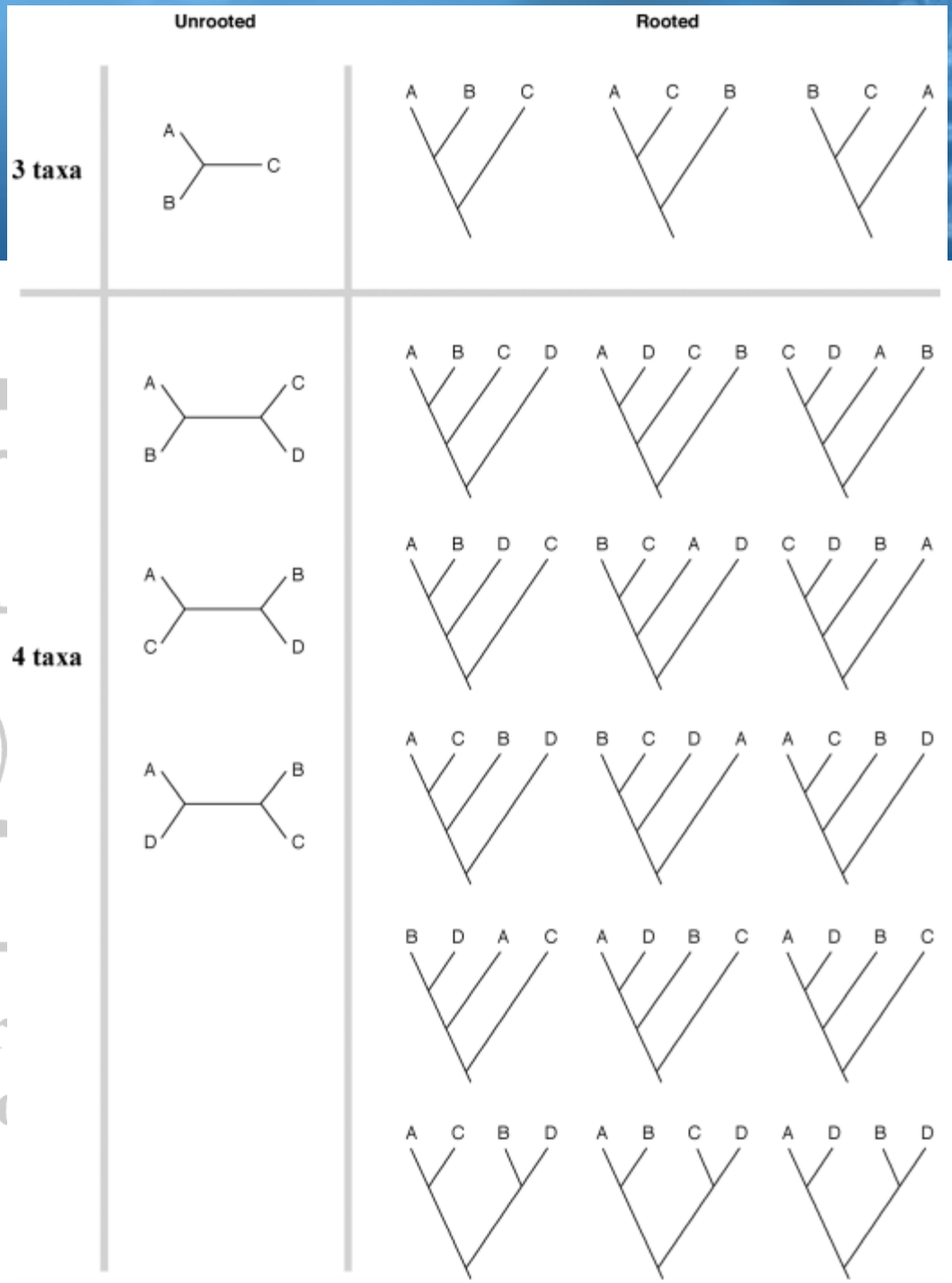
- The number of possible trees grows exponentially with the number of species (or sequences)
- $N_r = (2n - 3)!/2^{(n-2)}(n-2)!$
- $N_u = (2n - 5)!/2^{(n-3)}(n-3)!$
- To find the best tree, you must explore all possibilities

Number of rooted trees

species	number of trees
1	1
2	1
3	3
4	15
5	105
6	945
7	10,395
8	135,135
9	2,027,025
10	34,459,425
11	654,729,075
12	13,749,310,575
13	316,234,143,225
14	7,905,853,580,625
15	213,458,046,676,875
16	6,190,283,353,629,375
17	191,898,783,962,510,625
18	6,332,659,870,762,850,625
19	221,643,095,476,699,771,875
20	8,200,794,532,637,891,559,375
30	4.9518×10^{38}
40	1.00985×10^{57}
50	2.75292×10^{76}

Example:


Number of tree topologies
10^{22}
10^{20}
10^{18}
10^{16}
10^{14}
10^{12}
10^{10}
10^{8}
10^{6}
10^{4}
10^{2}
1
rooted
unrooted
4
8
12
16
20
Number of taxa


Unrooted
Rooted
3 taxa
4 taxa
A B C
A C B
B C A
A B C D
A D C B
C D A B
A B D C
B C A D
C D B A
A C B D
B C D A
A C B D
B D A C
A D B C
A D B C
A C B D
A B C D
A D B D

Tree Building Procedure

1. Choose molecular markers
2. Perform MSA
3. Choose a model of evolution
4. Determine tree building method
5. Assess tree reliability

Choice of Molecular Markers

- For constructing molecular phylogenetic trees, one can use either *nucleotide or protein sequence* data.
- The choice of molecular markers can make a major difference in obtaining a correct tree.
- **Very closely related organisms -** nucleic acid sequence will show more differences
- **For individuals within a species -** faster mutation rate is in noncoding regions of Mitochondrial DAN (mtDNA)
- **More distantly related species -** slowly evolving nucleic acid sequences like ribosomal RNA or protein sequences
- **Very distantly related species -** use highly conserved protein sequences

Advantages of Protein Sequences

- More highly conserved - mutations in DNA may not change amino acid sequence: sixty-one codons encode twenty amino acids.
- Third position in a codon especially has more variation - violates our assumption of independent evolution of all positions in a sequence
- DNA sequences can be biased by codon usage differences between species - causes variations in sequence that are not attributable to evolution
- In alignments, DNA sequences that are not related can show a lot of similarity due to only 4 letters in alphabet, proteins do not have this problem (at least not as much)
- Introducing gaps in alignments of DNA sequences can cause frameshift errors, making alignment biologically meaningless

Advantages of DNA Sequences

- Better for closely related species
- Show *synonymous* and *non-synonymous* mutations, which allows analysis of positive and negative selection events
- *Synonymous substitutions* are nucleotide changes in the coding sequence that do not result in amino acid sequence changes for the encoded protein.
- *Nonsynonymous substitutions* are nucleotide changes that result in alterations in the amino acid sequences
  - Lots of nonsynonymous mutations may mean *positive selection (or adaptive evolution)* for new functions of protein with different amino acid sequence
  - Lots of synonymous mutations may mean *negative selection (*or *purifying selection)* - changed amino acid sequence is detrimental

Multiple Sequence Alignment

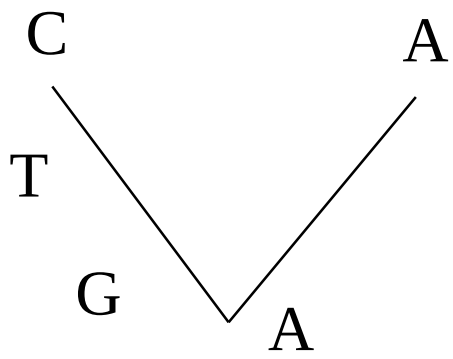
- Most critical step in tree building - cannot build correct tree without correct alignment
- Should build alignments with multiple programs, then inspect and compare to identify the most reasonable one
- Most alignments need manual editing
  - Make sure important functional residues align
  - Align secondary structure elements
  - Use full alignment or just parts

Automatic Editing of Alignments

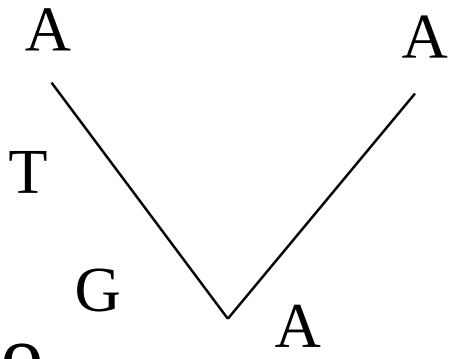
- Rascal and NorMD – correct alignment errors, remove potentially unrelated or highly divergent sequences
- Gblocks – detect and eliminate poorly aligned positions and divergent regions

Multiple Substitutions

- How to measure divergence between sequences?
  - Simple measure – just count the number of substitutions observed between the sequences in the MSA
  - Problem – number of substitutions may not represent the number of evolutionary events that actually occurred.
- *Intermediate mutation*: just because we only see one difference, does not mean that there was only one evolutionary event.



Multiple Substitutions (Cont.)



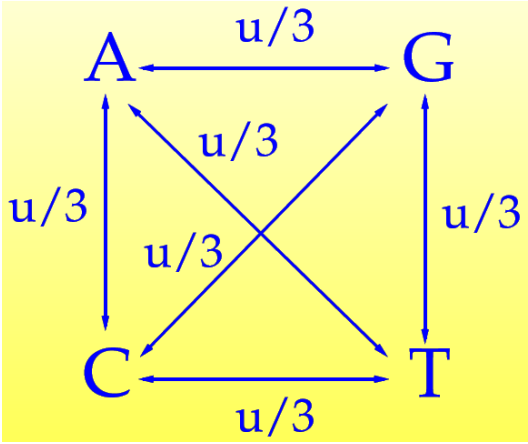
- *Back mutation*: Just because we only see no difference, does not mean that there were no evolutionary events.
- *Parallel mutations*: when both sequences mutate into the same residue.
- *Homoplasy*: multiple substitutions and convergence at individual positions obscure the estimation of the true evolutionary distances between sequences.
  - If not corrected, can lead to the generation of incorrect trees.

Choosing Substitution Models

- The statistical models used to correct homoplasy (multiple substitution problem) are called *substitution models* or *evolutionary models.*
- Focus on DNA models
- Substitution models are used to correct evolutionary distances:
  - Jukes–Cantor Model
  - Kimura Model

Jukes-Cantor Model

- The simplest nucleotide substitution model.
- Jukes-Cantor model assumes all nucleotides are substituted with equal probability.
- Can be used to correct for multiple substitutions
- Formula for deriving evolutionary distances:

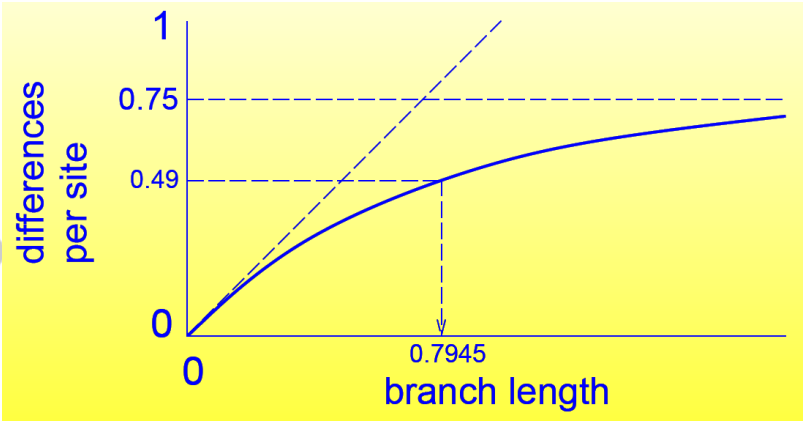


$$d_{AB} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p_{AB}\right)$$

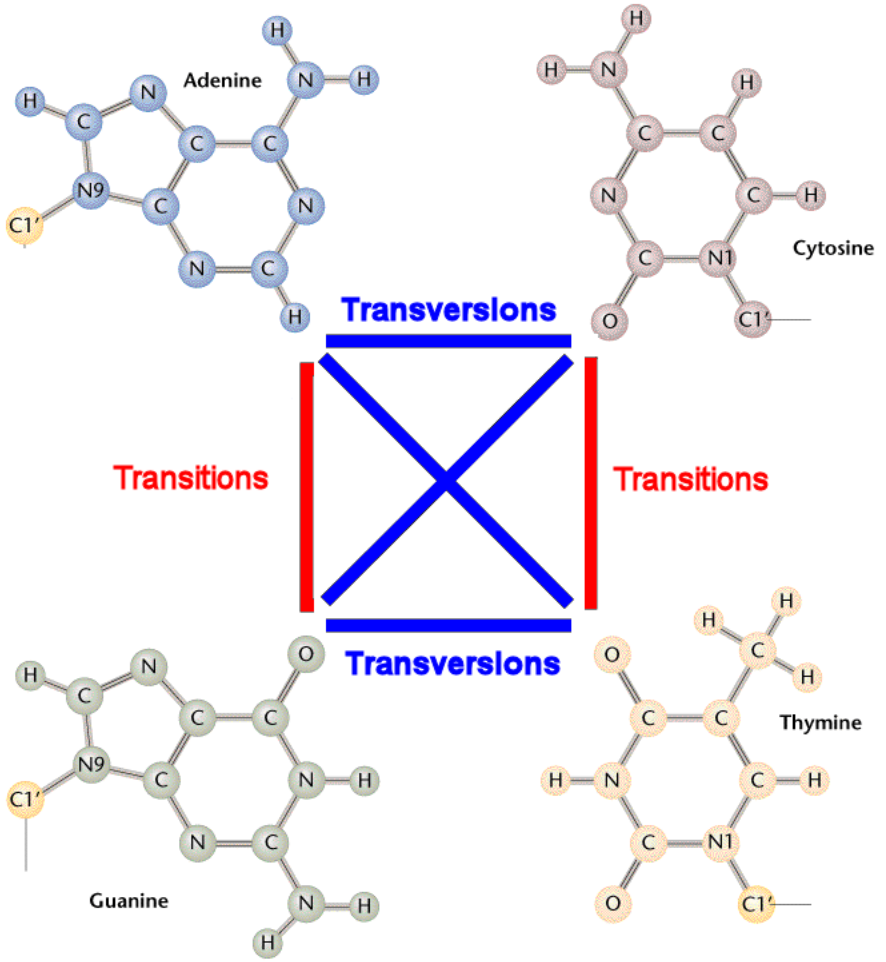
- p_{AB} is the observed sequence distance measured by the proportion of substitutions over the entire length of the alignment.

Jukes-Cantor Model (Cont.)

- For example, if an alignment of sequences A and B is twenty nucleotides long and six pairs are found to be different, the sequences differ by 30%, or have an observed distance 0.3 (p_{AB}).
- So: $$d_{AB} = -\frac{3}{4}\ln\left(1 - \frac{4}{3} \times 0.3\right) = 0.38$$
- The Jukes–Cantor model can only handle reasonably closely related sequences.



Transition vs Transversion mutations



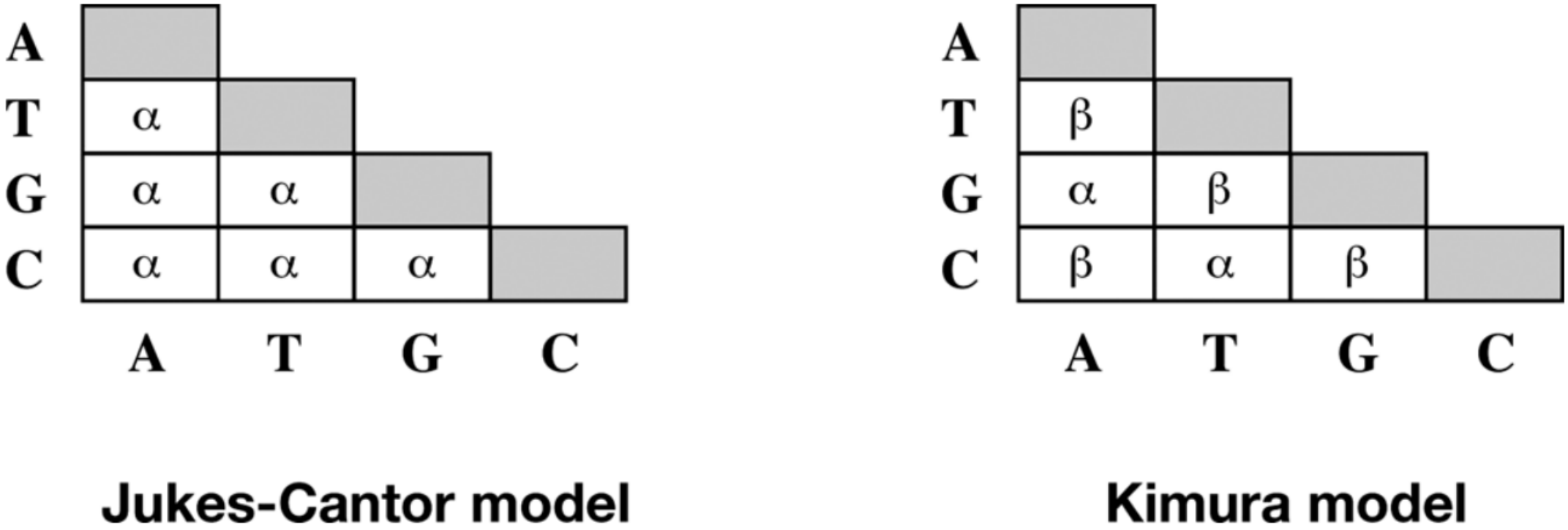
Kimura Model

- It also is called the Kimura two-parameter model.
- Here mutation rates for transitions and transversion are assumed to be different, which is more realistic.
  - Transitions occur more frequently than transversions.
- The Kimura model uses the following formula:

$$d_{AB} = -\frac{1}{2}\ln(1 - 2p_{ti} - p_{tv}) - \frac{1}{4}\ln(1 - 2p_{tv})$$

where p_{ti} is the observed frequency for transition, and p_{tv} the frequency of transversion.

Jukes–Cantor vs Kimura



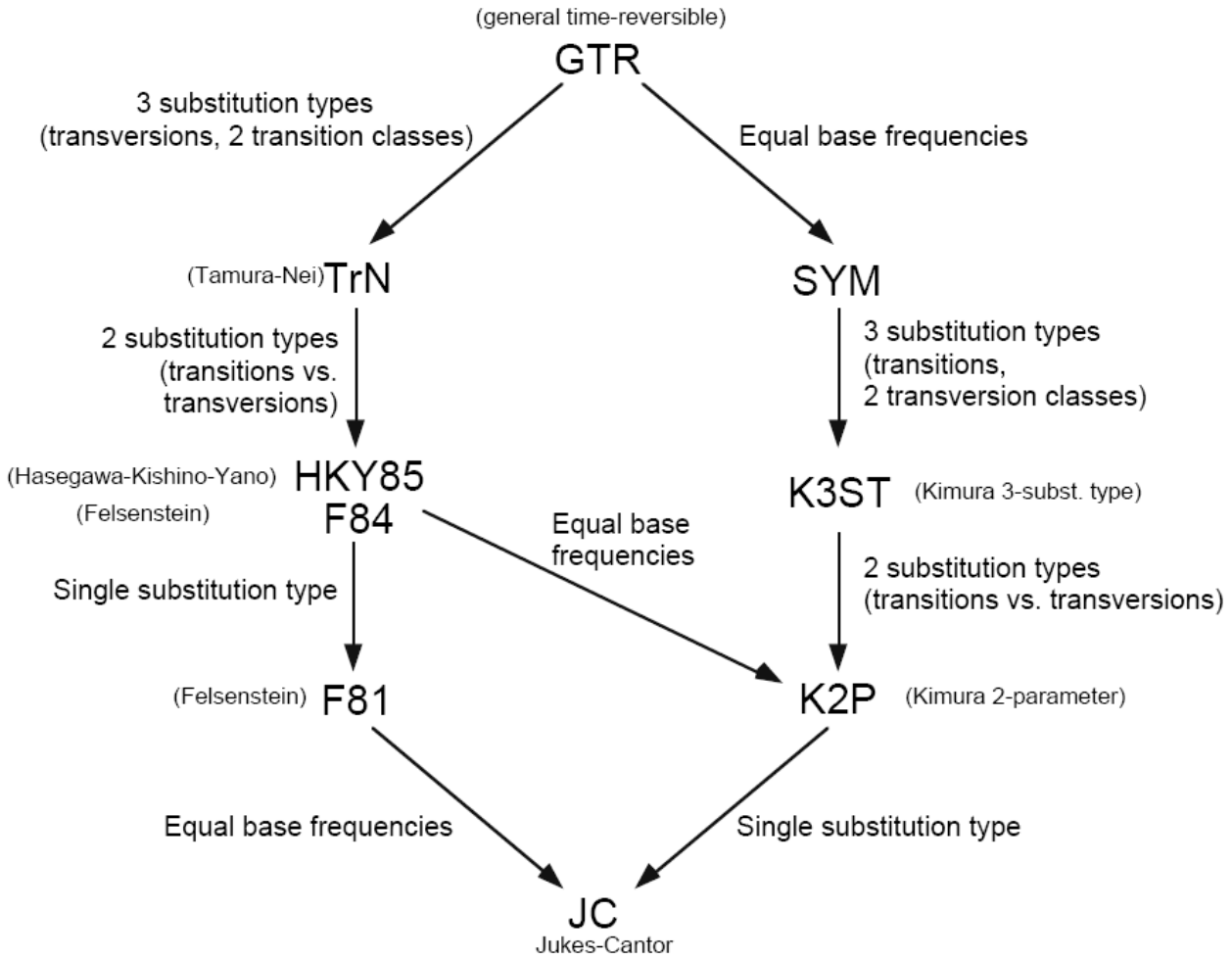
Transitions (α) and Transversions (β) rates

Example

- Assume a comparison of sequences A and B that differ by 30%. If 20% of changes are a result of transitions and 10% of changes are a result of transversions, the evolutionary distance can be calculated:

$$d_{AB} = -\frac{1}{2}\ln(1 - 2 \times 0.2 - 0.1) - \frac{1}{4}\ln(1 - 2 \times 0.1) = 0.40$$

Many Other Models



Evolutionary Models for Protein Sequences

- PAM and JTT substitution matrices already take into account multiple substitutions
- There are also models similar to Jukes-Cantor for protein sequences, but are not covered here.

Differences in mutation rates between positions

- One of our assumptions was that all positions in a sequence are evolving at the same rate
- Bad assumption
  - Third position in a codon changes with higher frequency
  - In proteins, some amino acids can change and others cannot
- This variation is called *among-site rate* heterogeneity
- Many tree building programs have parameters meant to deal with this problem – adds to complexity of getting the correct tree

References

- Mostly used:
  - Essential bioinformatics, Chapter 10 (Phylogenetics Basics)
- Second reference:
  - Bioinformatics and functional genomics, Chapter 7 (Molecular Phylogeny and Evolution)

- IP notice: some slides were selected from Drena Dobbs' slides.

Thanks for your attention