Bioinformatics Algorithms
COS-BIOL-530/630
Lecture04

Days & Times	Room	Meeting Dates
Tu 2:00PM - 3:50PM	Thomas Gosnell Hall (GOS)-2178	01/13/2025 - 04/28/2025
Th 2:00PM - 3:50PM	Thomas Gosnell Hall (GOS)-2178	01/13/2025 - 04/28/2025

Instructor:
Fernando Rodriguez
email: frvsbi@rit.edu
Office: Orange Hall 1311

Phylogenetics
- Lecture04 & Lecture05 -

Announcements

Week 4
Lecture04
Lab04
- Discussion 4
- Activity 4

Week 5
Lecture05
Lab05
Activity 5
Discussion 5

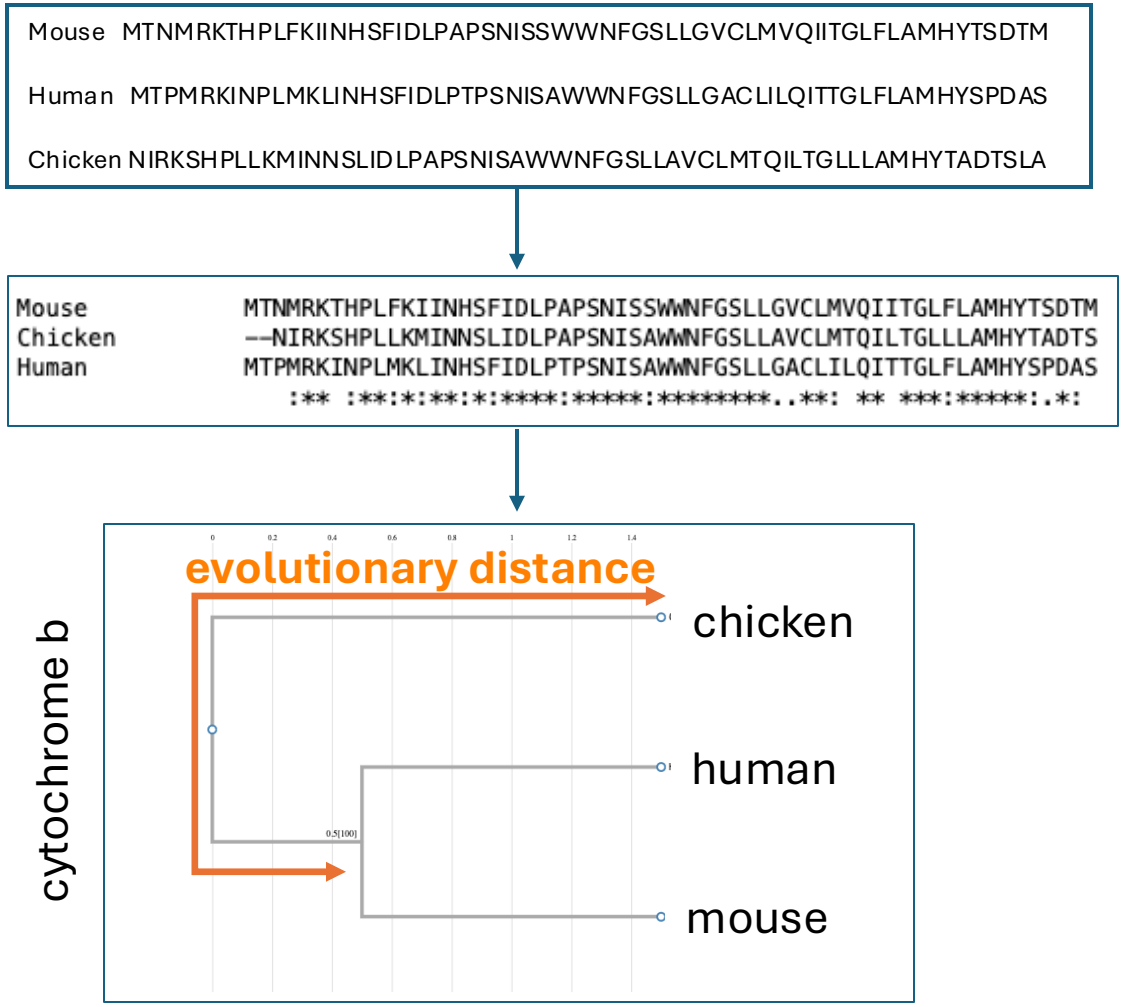
Quiz 4 (week 4-5)

Perusal: Rodriguez *et al*. 2009

Phylogenetics
- Lecture04 & Lecture05 -

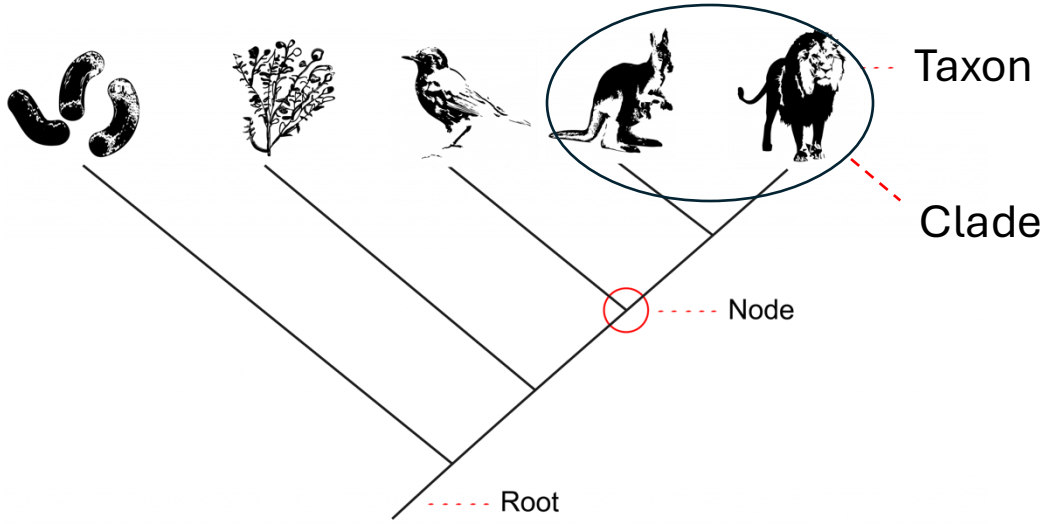
Topics:

- Substitution matrices
- Phylogenetic trees
  - Distance based
  - Character based
- Labs 04 &05: PHYLIP!



Phylogenetics

- Phylogenetic analysis is the means of inferring or estimating evolutionary relationships.
- Phylogenetic analysis are based on the analysis of characters or traits.
    - Morphological data (beak shape in finch).
    - Molecular phylogenetics by comparing nucleotides or amino acids in a sequence (the alignment is the character or trait).
- Different types of hierarchical trees most commonly represent the resulting relationships.

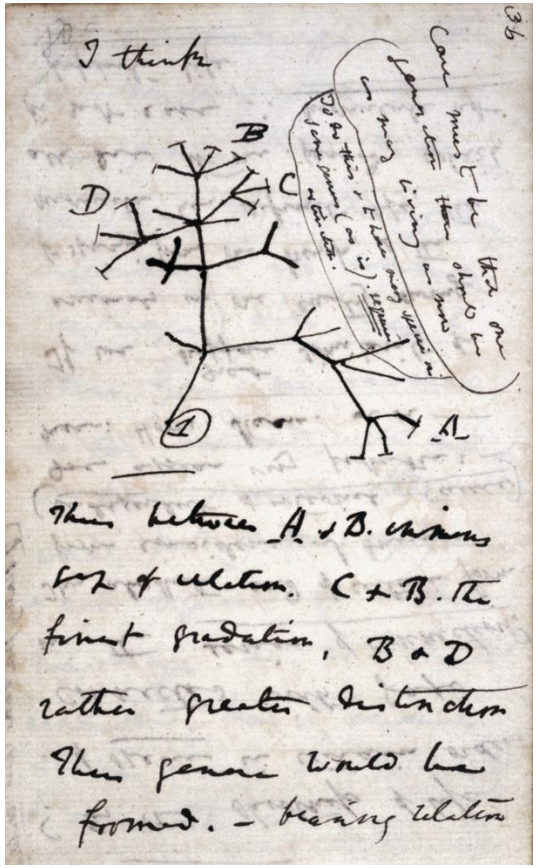


Tree topology: how the nodes and branches connect the taxa.

Phylogenetics

Basic steps for constructing a phylogenetic tree:

- Define a biological question.
- Selecting sequences that are homologous (share a common ancestry)
- Comparison of conserved and variable characters.
- Quantification of the change between sequences.
- Representation of data in a tree.

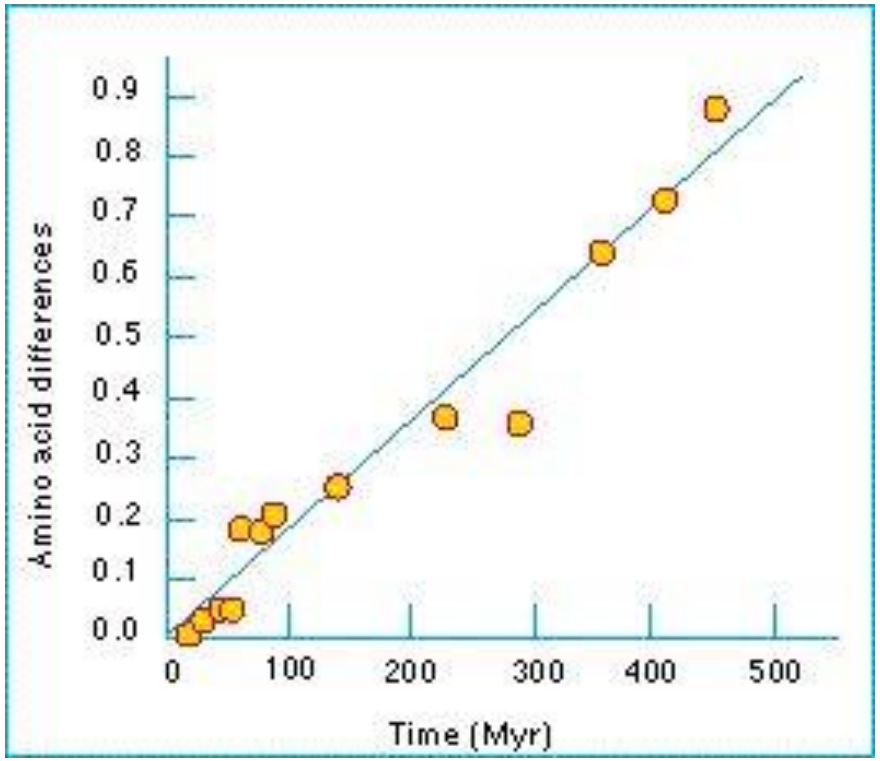


Page from Darwin's notebooks around July 1837 showing his first sketch of an evolutionary tree.

Sequences as Molecular Clocks

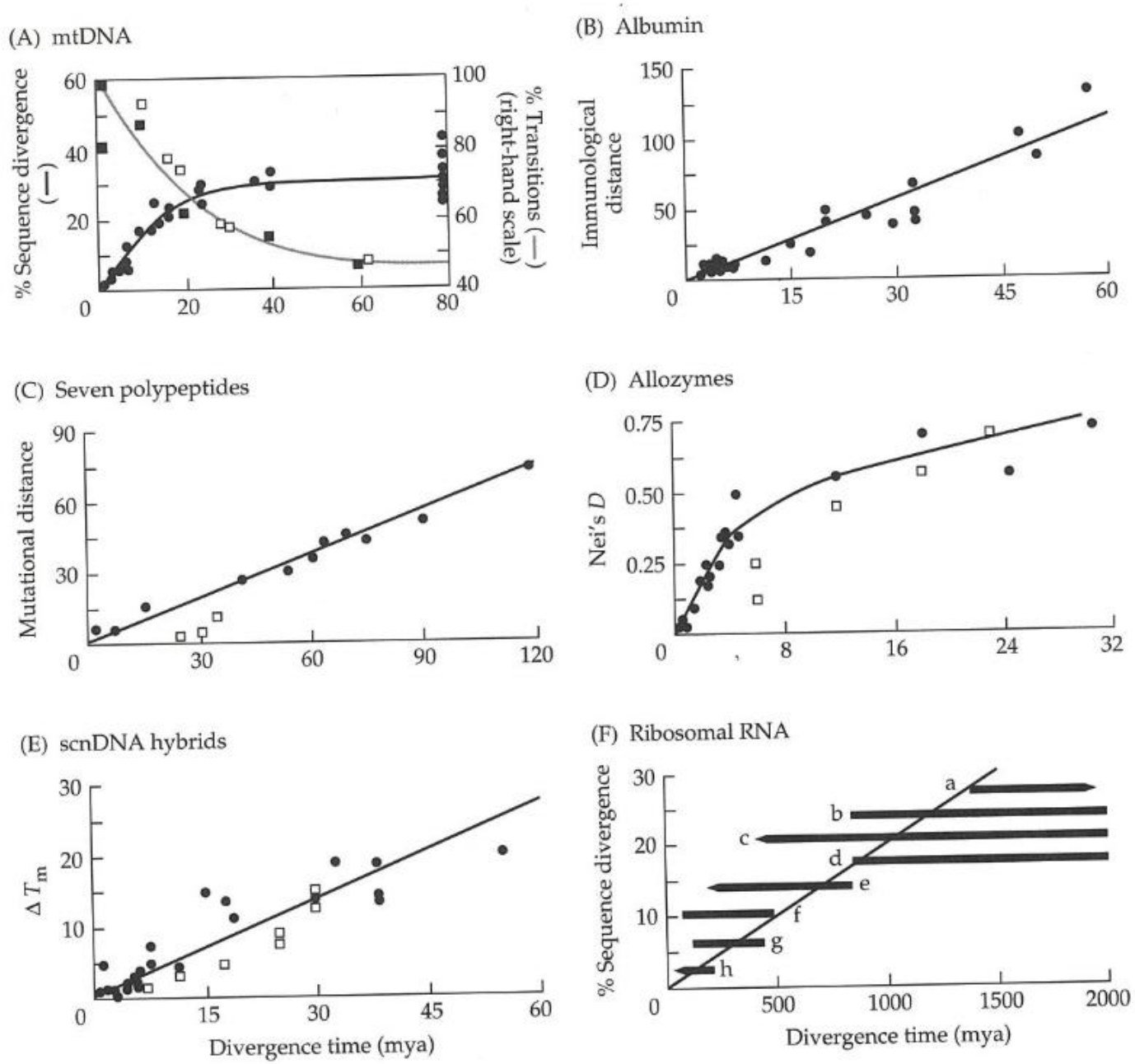
- DNA sequences (genes, regulatory regions) accumulate different types of mutations over time:
    - Mutations
    - Insertions
    - Deletions
    - Expansion of repetitive sequences
    - Gene duplications
- Since protein sequence is intricately linked to functionality, the accumulation of mutations in specific regions can significantly impact protein function, cell viability, and species fitness.
- The more **sequence differences** between organisms, the more time they likely had to independently acquire mutations—and, thus, the more distant their **evolutionary relationship**.

Sequences as Molecular Clocks



The constant rate of evolution of the alpha-globin
Source: *Evolution*, M. Ridley.

- Observations about amino acid changes that occurred during the divergence between species.
- Molecular evolution takes place at an **approximately constant rate**.
- Molecular differences between species are therefore used to infer phylogenetic relations.



But molecular clocks need to be calibrated for various types of molecular genetic data.

Source: *Molecular markers natural history and evolution*, Avise.

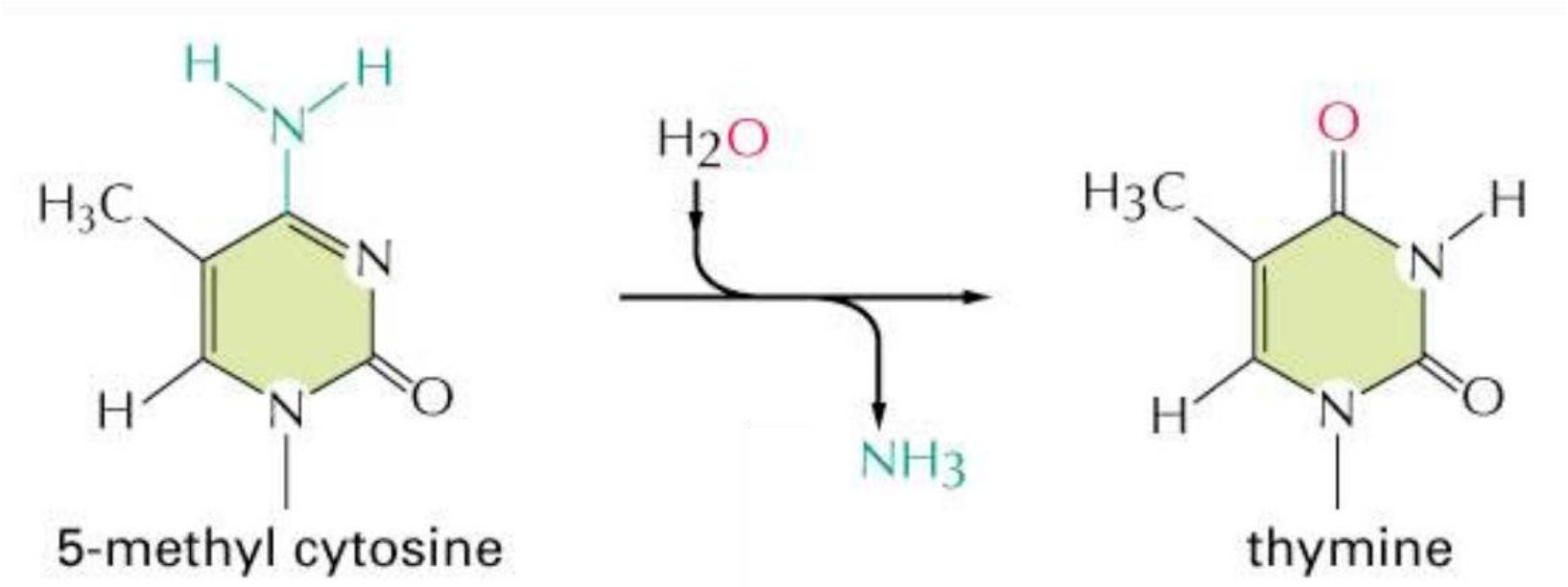
Substitution matrices

- Typically, a mismatch is coded by a single value.
- However, certain nucleic acid mutations are more common than others.
- Transitions are more common than transversions.

Nucleotide scoring matrix from Lecture02

	A	T	G	C
A	5	-4	-4	-4
T	-4	5	-4	-4
G	-4	-4	5	-4
C	-4	-4	-4	5

C to T transition

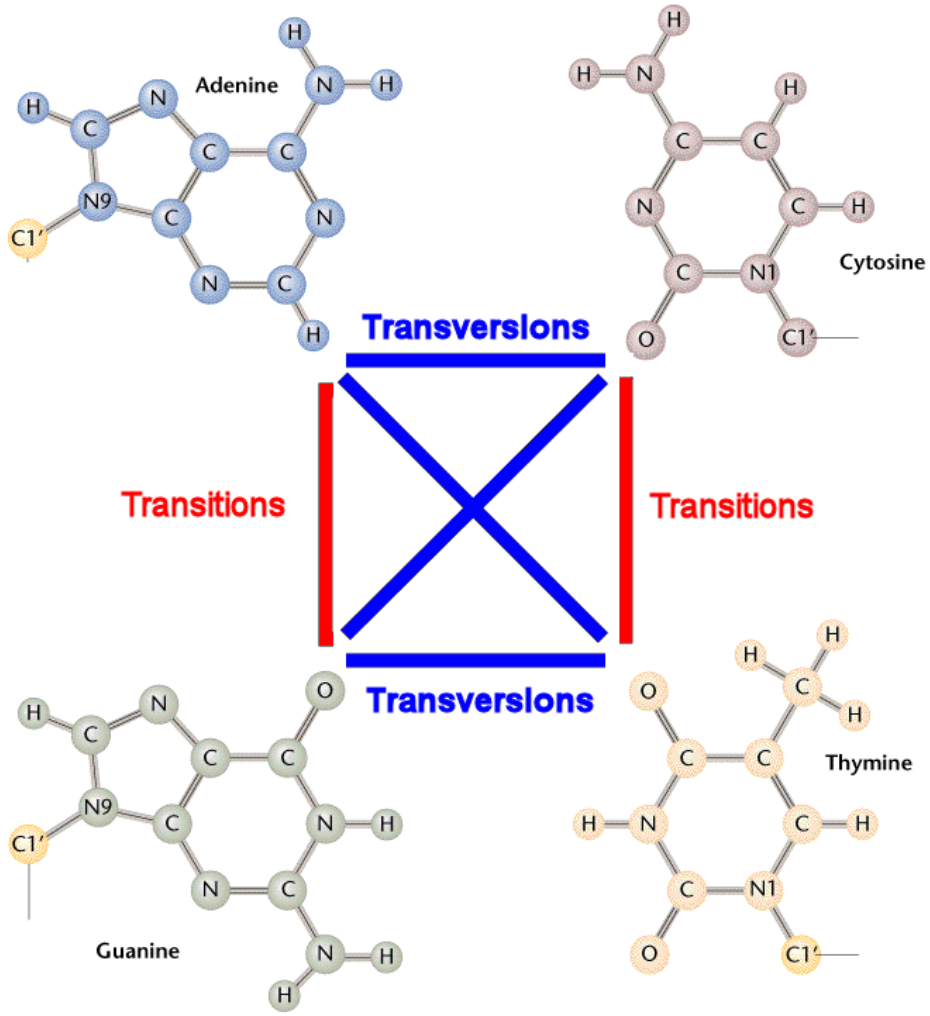


Spontaneous in the presence of water:
The process is called oxidative deamination.
More frequently in the presence of external sources like UV light.

Transitions *vs*. transversions

Types of DNA mutations:

- **Transitions**: are interchange of two-ring of purines (A<->G), or of one-ring pyrimidines (C<->T).
- **Transversions**: are interchanges of purine for pyrimidine bases (exchange of one-ring and two-ring structures)



Source: https://www.mun.ca/biology/scarr/Transitions_vs_Transversions.html

Substitution matrices

- We need to weigh the nucleotide mismatches ratio by knowing the transition/transversion ratio in:

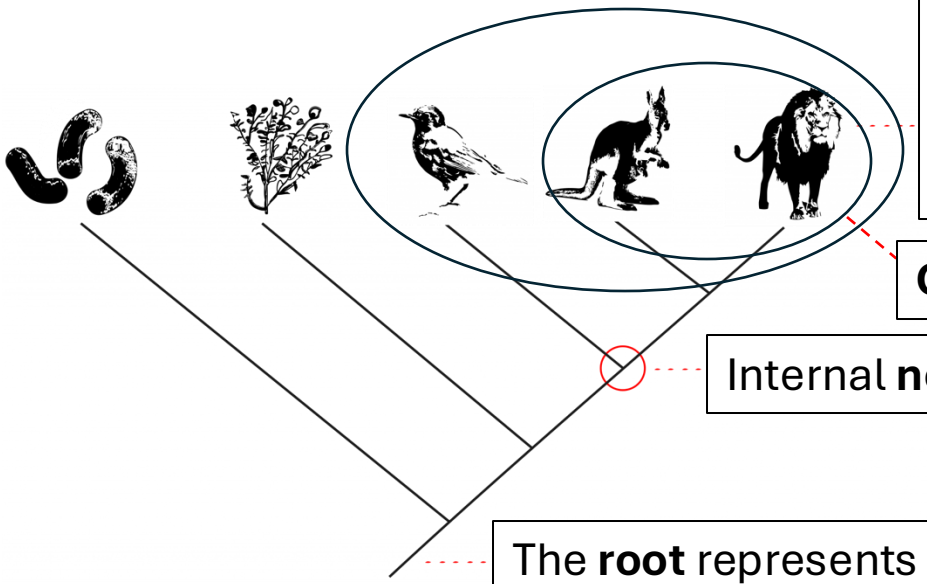
  - Organism / genome
  - Organelle: mitochondria / chloroplast
  - Specific gene

- And give a penalty for:

  - Transitions: small penalty (more frequent substitution)
  - Transversions: bigger penalty (less frequent substitution)

Phylogenetic trees

Terminology



Leaves represent things (genes, individuals/strains, species) being compared.
Taxon (**taxa** plural) is used to refer to these when they represent species and broader classifications of organisms.

Clades are group of species with a common ancestor.

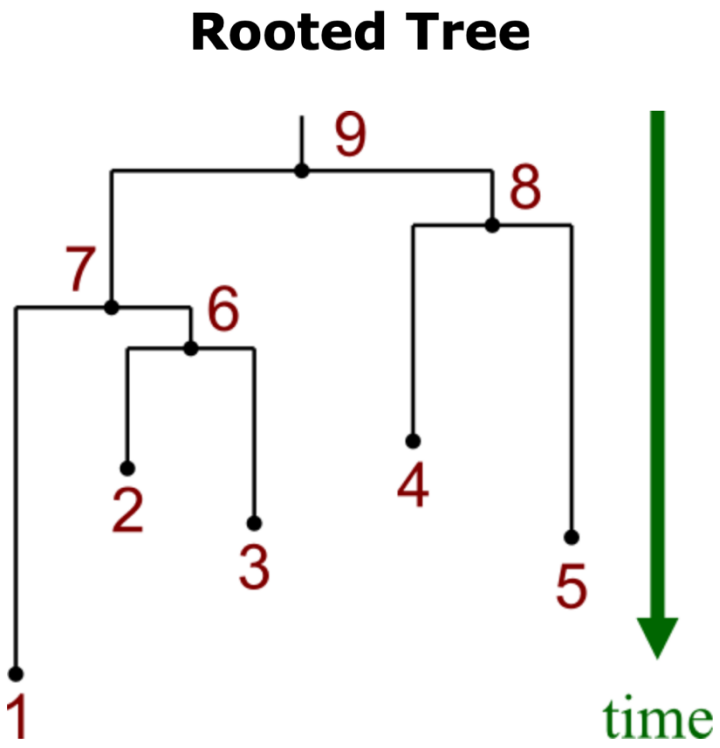
Internal **nodes** are hypothetical ancestral units

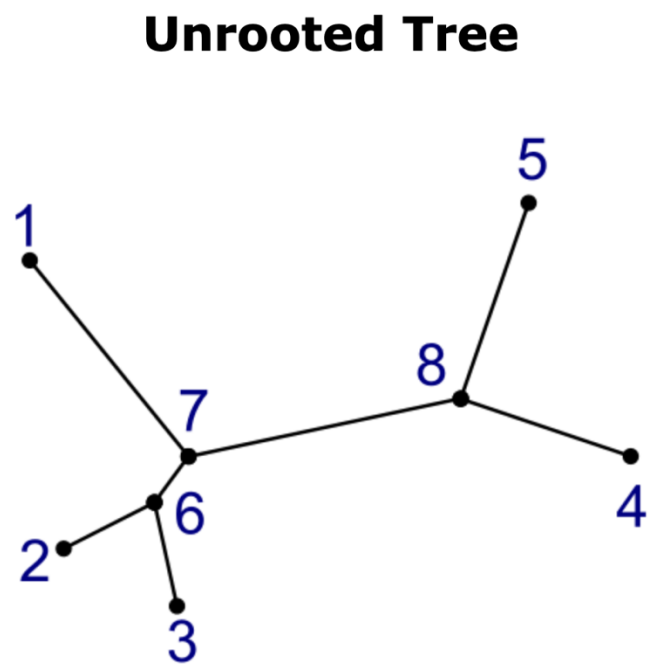
The **root** represents the common ancestor.

In a ***rooted tree***, the path from root to a node represents an evolutionary path.

Phylogenetic trees

rooted & unrooted





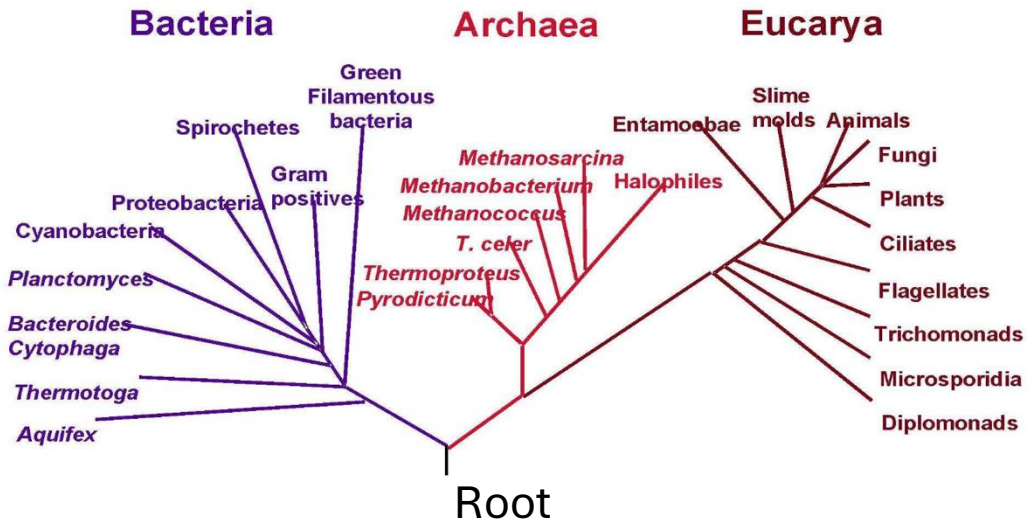
Rooted trees tell you something about evolutionary paths over time.

Unrooted trees only tell you relations.

Phylogenetic trees

outgroups

- For methods that generate unrooted tress, it is common to include an outgroup to effectively provide a root.
- Typically, you need to provide something just outside of the set that is being studied.
- For example, you compare several Eukaryotes and use an Archea bacteria as an outgroup.



The phylogenetic “tree of life”
Source: Woese, Kandler & Wheelis 1990

Phylogenetic trees

Question: Given *n* species/genes/sequences, how many trees is it possible to build?

- A tree construction algorithm would be:

    1. Enumerate every possible tree buildable from the *n* leaves.
    2. Score each one of them.
    3. Return the one with the best score.

- Number of unrooted trees.

$$N = (2n - 5)!/(2^{n-3}(n-3)!)$$

- Number of rooted trees.

$$N = (2n - 5)!/(2^{n-2}(n-2)!)$$

Phylogenetic trees

Number of possible trees

Number of Leaves	Unrooted trees	Rooted trees
2	1	1
4	3	15
6	105	945
8	10,395	135,135
10	2,027,025	34,459,425

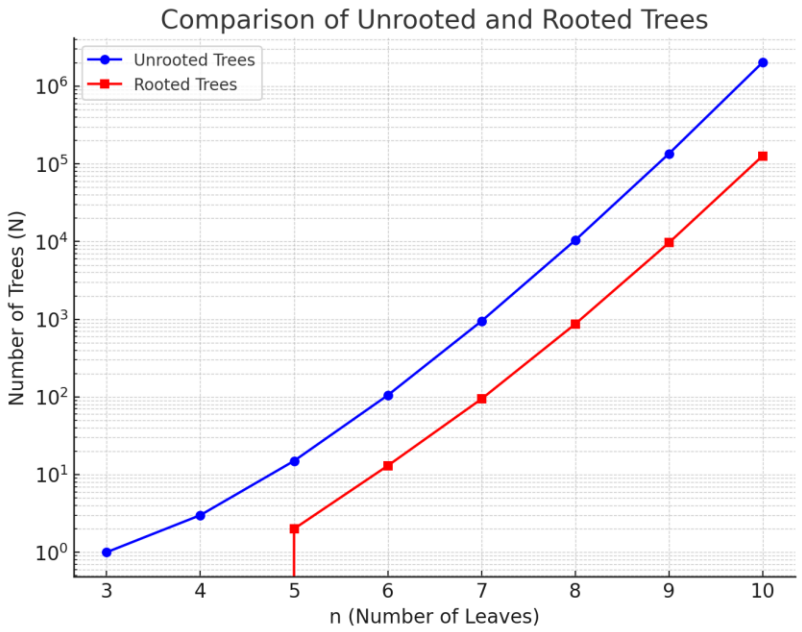
- This is a _____ function?

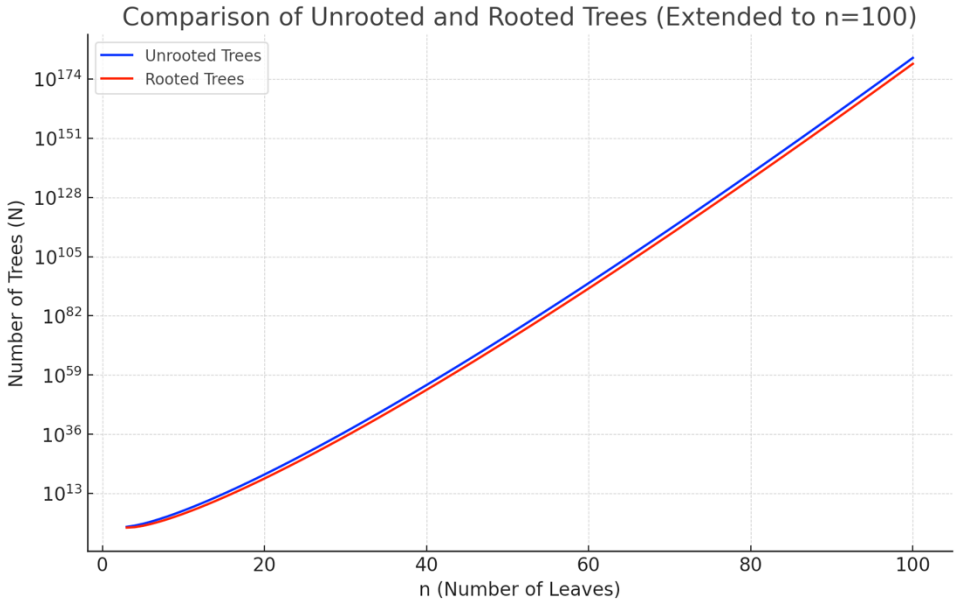
$N = (2n-5)!/(2^{n-3}(n-3)!)$

$N = (2n-5)!/(2^{n-2}(n-2)!)$

Phylogenetic trees

Number of possible trees





- The number of trees grows exponentially
  - logarithmic scale on the y-axis
- Does it look like a reasonable function for an algorith?
- We need better ways (algorithms).

Phylogenetic trees

Methods

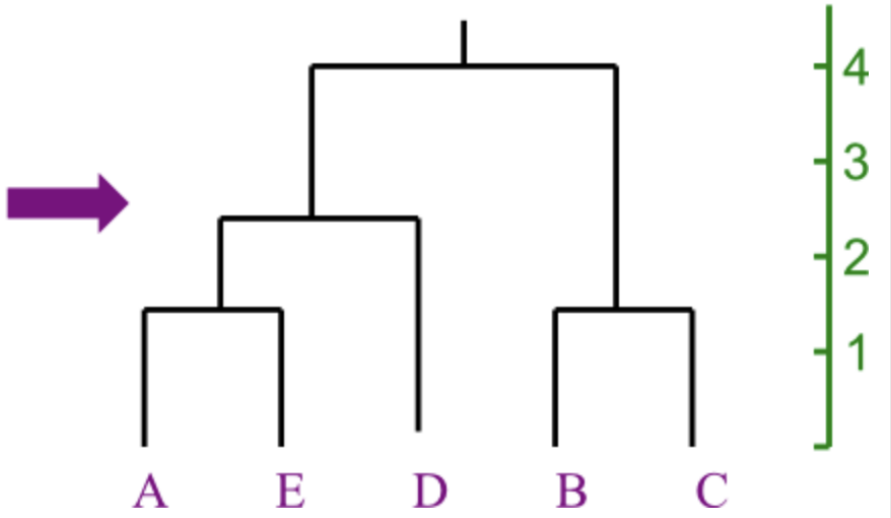
- Trees can be constructed from various types of data:
  - **Distance-based**: measures of distance between species/genes.
  - **Parsimony-based**: explains the character-based data with minimal number of changes (*Occam's Razor*).

Distance-based Methods

- Distance-based measures use pairwise distances between any two sequences.
- Use a substitution model to score pairwise distances.

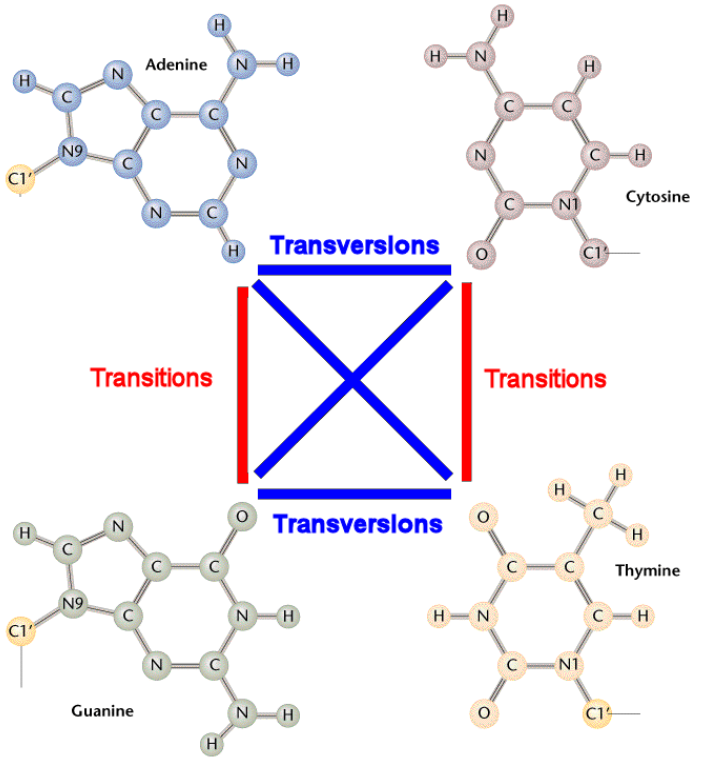
- The distance between objects (sequences) is represented in a matrix (*n* x *n* size).
- Then, we build an edge-weighted tree with the distances between leaves (taxa).

	A	B	C	D	E
A	0	8	8	5	3
B		0	3	8	8
C			0	8	8
D				0	5
E					0



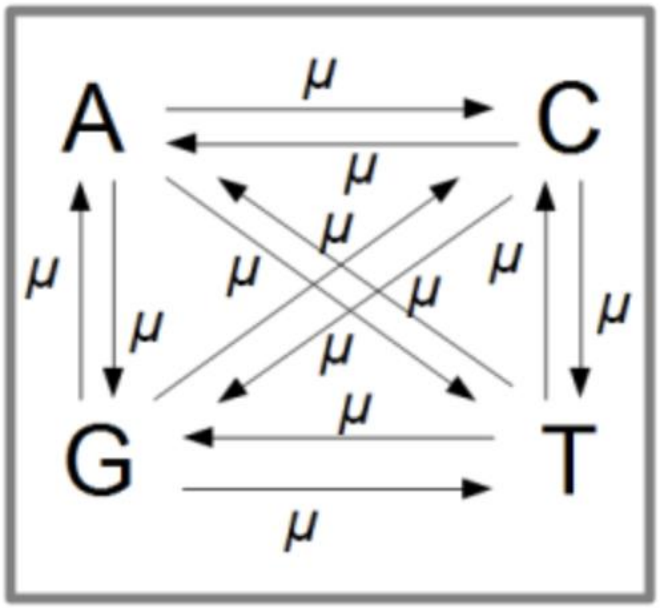
Distance-based Methods

- There are many ways to calculate distances between any two sequences.
- Jukes-Cantor
  - Distance is proportional to time.
  - Assumes molecular clock.
- Kimura two-parameter
  - Distance is proportional to time.
  - Assumes molecular clock.
  - Transversion rate < transition rate



Distance-based Methods

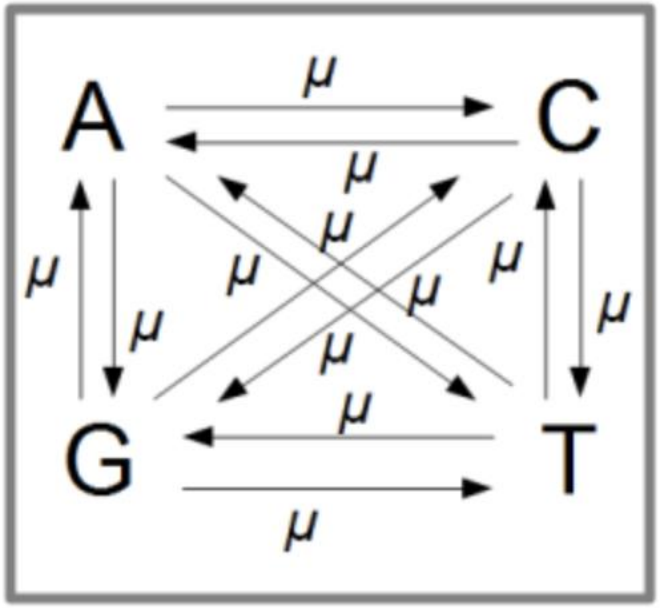
Jukes-Cantor Model of DNA substitution



- Mutation rates (μ) in all directions are equal.
- But there is also a mutation and back mutation.
- Estimates the true number of substitutions between lineages, including back mutations

Distance-based Methods

Jukes-Cantor Model of DNA substitution

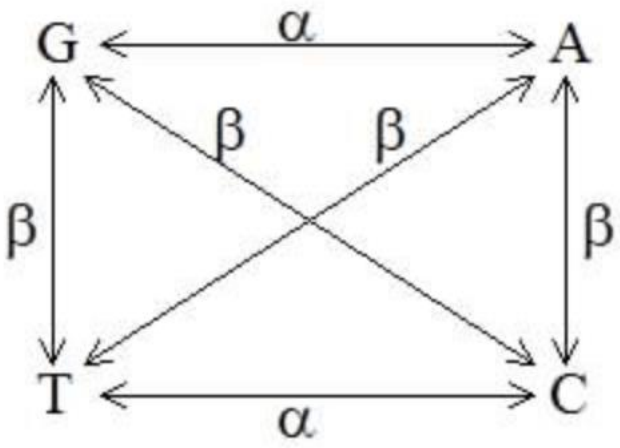


But it makes a few assumptions:

- If all nucleotides are equally frequent in the sequences
- All mutations are equally likely to happen
- Ignores gaps

Distance-based Methods

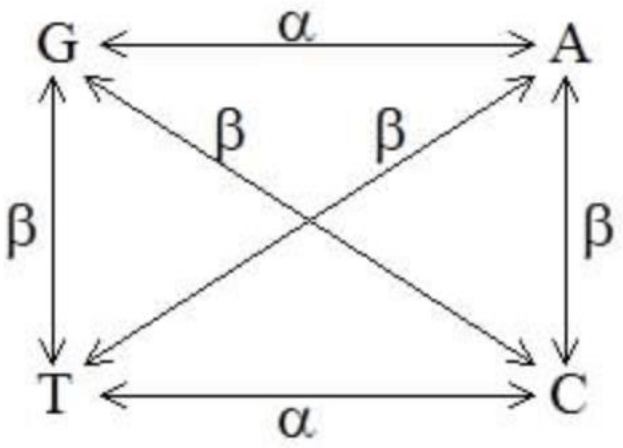
Kimura two-parameter Model of DNA substitution



- Nucleotide bases fall into two categories:
  - Purines: A & G (two-rings base)
  - Pyrimidines: C & T (one-ring base)
- In nature, the number of transitions observed is more frequent than transversions.
- Kimura's model incorporates these different rates (α, β).

Distance-based Methods

Kimura two-parameter Model of DNA substitution

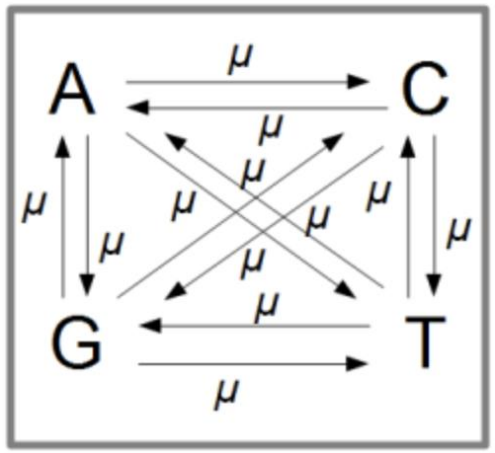


- Assumes all bases are equally frequent.
- Ignores gaps.

- There are other distance-based models.
- Many start with Kimura and build onto it.
- As more data became available, models with even more parameters are build.

Distance-based Methods

Calculating distances

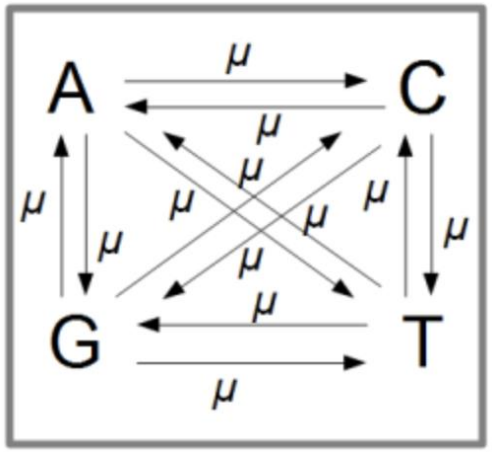


- For a multiple alignment, we can calculate the distance between each pair of sequences.
- Then, construct a matrix, whose elements (dij) are the distances between each pair of leaves, i and j.
- Eg. for Jukes-Cantor:
  D: fraction of sites that differ between two sequences (transitions + transversions)
  d: evolutionary distance

$$d = -\frac{3}{4} ln\left(1 - \frac{4}{3} D\right)$$

https://www.megasoftware.net/web_help_7/hc_jukes_cantor_distance.htm

Distance-based Methods



Source: *Bioinformatics and Molecula Evolution*, by Higgs & Attwood

BOX 4.1 Solution of the Jukes–Cantor model

Consider a single site in a sequence. We define $P_{ij}(t)$ to be the probability that the site is in state j at time t given that the site in the ancestral sequence was in state i at time 0. Here i and j label the base (A, C, G, or T) that is present at that site. Figure 4.2 shows a site that was initially in state A, that is in some state k at time t, and that is in state A a short time interval δt after that. The probability of this situation is (by definition) $P_{AA}(t+\delta t)$. This can be written as a sum over the four possible bases that could exist at time t. If base k is a C, then there is a probability $P_{AC}(t)$ that an A changes to a C in time t multiplied by the probability $\alpha\delta t$ that a substitution from a C to an A happens in the interval δt. The terms for the cases where k is G or T are similar to this. The term for the case where k is an A is the probability $P_{AA}(t)$ that the base is an A after time t, multiplied by the probability that there is no substitution in the interval δt. The total rate of change is 3α; therefore the probability that there is no substitution is $1 - 3\alpha\delta t$. Adding these four terms, we obtain:

$$P_{AA}(t+\delta t) = \alpha\delta t\,(P_{AC}(t) + P_{AG}(t) + P_{AT}(t)) + (1 - 3\alpha\delta t)P_{AA}(t) \qquad (4.5)$$

For small δt we know that $P_{AA}(t+\delta t) = P_{AA}(t) + \delta t \dfrac{dP_{AA}}{dt}$

(see Section M.6). After substituting this into the above equation, the term $P_{AA}(t)$ can be subtracted from both sides, and we can then cancel out the δt, which results in the following differential equation for P_{AA}:

$$\frac{dP_{AA}}{dt} = \alpha(P_{AC} + P_{AG} + P_{AT}) - 3\alpha P_{AA} \qquad (4.6)$$

We also know that $P_{AC} + P_{AG} + P_{AT} = 1 - P_{AA}$: this just says that the sum of the probabilities that base k is a C, G, or T is equal to the probability that base k is **not** an A. Putting this into Eq. (4.6), we obtain:

$$\frac{dP_{AA}}{dt} = -4\alpha P_{AA} + \alpha \qquad (4.7)$$

This is an equation that involves only the single unknown function P_{AA}. It is of a standard form that allows us to guess that the solution must be of the form $P_{AA}(t) = Ae^{-4\alpha t} + B$, where A and B are two constants to be determined. By substituting this trial solution back into Eq. (4.7), following the method described in Section M.8, we find that the equation is satisfied if $B = 1/4$. By using the initial condition that $P_{AA}(0) = 1$, we find that $A = 3/4$. Therefore the solution is:

$$P_{AA}(t) = \frac{3}{4}e^{-4\alpha t} + \frac{1}{4} \qquad (4.8)$$

Due to the symmetry of the model, $P_{CC}(t)$, $P_{GG}(t)$ and $P_{TT}(t)$ are all equal to $P_{AA}(t)$. Also, $P_{AC}(t)$, $P_{AG}(t)$, $P_{AT}(t)$ are all equal to one another, hence:

$$P_{AC}(t) = \frac{1}{3}(1 - P_{AA}(t)) = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t} \qquad (4.9)$$

The functions $P_{AA}(t)$ and $P_{AC}(t)$ are plotted in Fig. 4.3(a). Both functions tend to $1/4$ for long times. This is because after a long time the site is equally likely to be any of the four bases, irrespective of which base it was at time 0.

The probability that the base at time t is different from the base at time 0 is equal to $P_{AC}(t) + P_{AG}(t) + P_{AT}(t) = 3P_{AC}(t)$. When we are comparing two species with a time t since the common ancestor, the total time separating the species is $2t$. Therefore, the probability that these two species have a different base is $D = 3P_{AC}(2t)$, which gives us an $8\alpha t$ in the exponential of Eq. (4.9). Thus we obtain the Eq. (4.2) for D in the main text.

Distance-based Methods

Substitutions between Protein sequences

	A	B	C	D	E
A	0	8	8	5	3
B		0	3	8	8
C			0	8	8
D				0	5
E					0

We use protein alignments to compute a distance matrix, under different models of amino acid replacement.

- The proportion of different amino acids between two sequences:

  $P = n / L$
  - n: number of amino acids that differ between the two sequences.
  - L: number of positions at which differences could be observed in the aligned sequences.

- Models of amino acid substitution: Estimating the number of substitutions is more complicated than DNA.
  - Back mutations, some substitutions occur more frequently.
  - The path of substitution from one to another is not always the same length.

  - PAM matrix: he point accepted mutation (PAM) substitution model, also known as the Dayhoff substitution model, is an amino acid substitution model derived from empirical observation of mutations among closely related proteins (used in PHYLIP).

Small in-class activity!!

Lab04

- Article: Rodriguez *et al*. 2009
  - https://pubmed.ncbi.nlm.nih.gov/18796461/
  - In Perusal: https://app.perusall.com/courses/bioinformatics-algorithms-spring25

- We are going to recreate the phylogenetic analysis:
  - Read Materials and Methods & Results
    - Feel free to post questions & comments in Perusal
  - Download sequences from GenBank (multi fasta file)
    - There are 20 sequences (haplotypes)
    - 2 sequences as outgroups
  - Have everything ready:
    - Files (fasta)
    - software: PHYLIP, MEGA, Tree viewer (Lab00)

Construction Phylogenetic Trees

Once we have a distance matrix, it can be used as input to a clustering method.

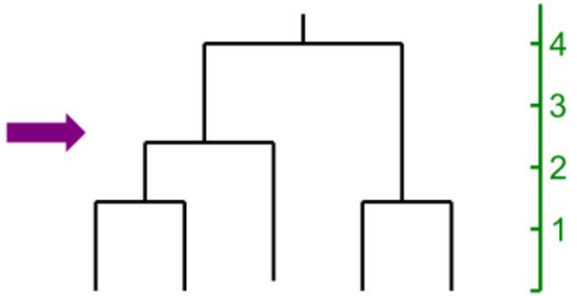
- Join the closest two clusters to form a single larger cluster.
- Recalculate distances between all the clusters.
- Repeat steps 1 and 2 until all leaves (species/genes) have been connected to a single cluster.
- We will discuss two clustering methods that are relevant in phylogenetics:
    - UPGMA
    - Neighbor-Joining

Construction Phylogenetic Trees

UPGMA

- Unweighted Pair Group Method using arithmetic Averages (UPGMA):
  - Find the two closer leaves/clusters and merge them: create a new node in tree for the merged cluster.
  - Recalculate distances between all clusters: in the UPGMA method, the distance between two cluster is defined as the mean of the distances between species of the two clusters.
  - Repeat until all clusters are joined

	A	B	C	D	E
A	0	8	8	5	3
B		0	3	8	8
C			0	8	8
D				0	5
E					0



Construction Phylogenetic Trees

UPGMA

	A	B	C	D	E
A	0	8	8	5	3
B		0	3	8	8
C			0	8	8
D				0	5
E					0

Let's build a tree using UPGMA:

- Which two leaves are closest to each other?

Construction Phylogenetic Trees

UPGMA

	AE	B	C	D
AE	0	8	8	5
B		0	3	8
C			0	8
D				0

Let's build a tree using UPGMA:

- Which two leaves are closest to each other?
- Recalculate the distance matrix, so that AE is a member.
- Repeat…

Distance AE – D:

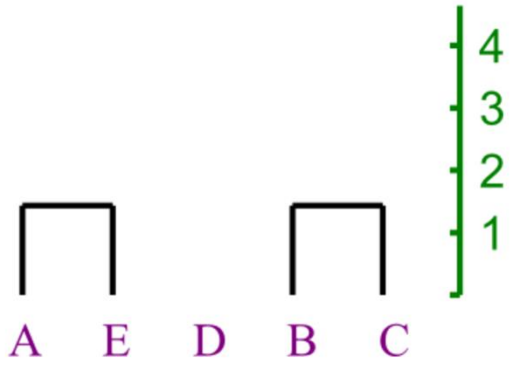
$d(AD) + d(ED)/2$

$(5 + 5)/2$

Construction Phylogenetic Trees

UPGMA

	AE	BC	D
AE	0	8	5
BC		0	8
D			0



Let's build a tree using UPGMA:

- Which two leaves are closest to each other?
- Recalculate the distance matrix, so that AE is a member.
- Repeat. Merge B and C now. Recalculate measures.
- Then, AE is closest to D

Construction Phylogenetic Trees

UPGMA

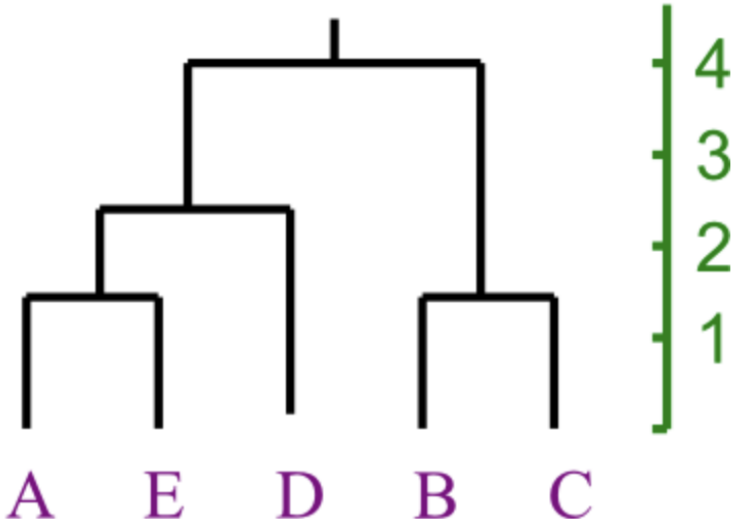
	AED	BC
AED	0	8
BC		0

Let's build a tree using UPGMA:

- Which two leaves are closest to each other?
- Recalculate the distance matrix, so that AE is a member.
- Repeat. Merge B and C now. Recalculate measures.
- Then, AE is closest to D.
- At the end, merge the last two clusters.

Construction Phylogenetic Trees

UPGMA

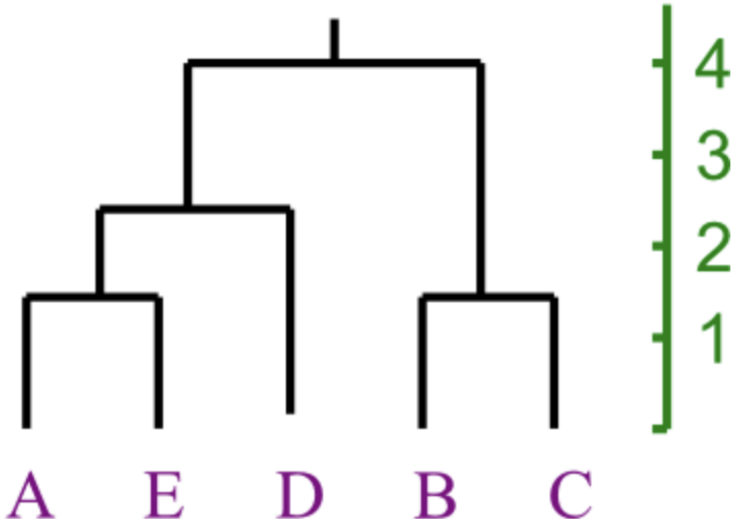


Let's build a tree using UPGMA:

- Which two leaves are closest to each other?
- Recalculate the distance matrix, so that AE is a member.
- Repeat. Merge B and C now. Recalculate measures.
- Then, AE is closest to D
- At the end, merge the last two clusters.

Construction Phylogenetic Trees

UPGMA



A few things to consider on UPGMA trees:

- UPGMA assumes that all lineages evolve at the same rate (per the molecular clock hypothesis).
- It creates a tree where all **leaves are equidistant** from the root.
- If the lineages evolve at different rates (which they do in reality), the UPGMA tree may not fit the distance data well.
- **Ultrametricity property**: pick three leaves from the UPGMA tree, the two larger of the three distances between these leaves will be equal.

Construction Phylogenetic Trees

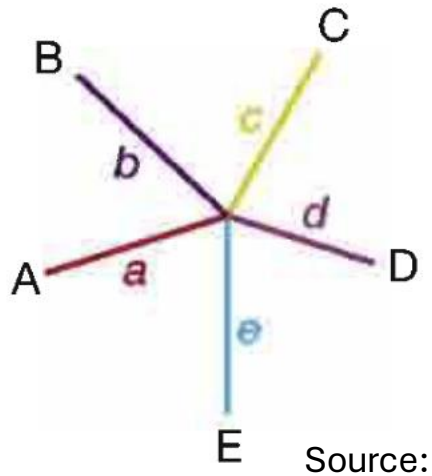
Neighbor Joining

- Neighbor Joining (NJ) is a distance-based method, and like UPGMA construct a tree by **iteratively joining subtrees**.
- But doesn't make the molecular clock assumption (all lineages evolve at the same rate).
- Produces unrooted trees.
- Assumes **additivity**: distance between pair of leaves is sum of lengths of edges connecting them.
- NJ can also form a rooted tree if you do have an outgroup.

Construction Phylogenetic Trees

Neighbor Joining

- Neighbor Joining starts with a “star-like” tree.



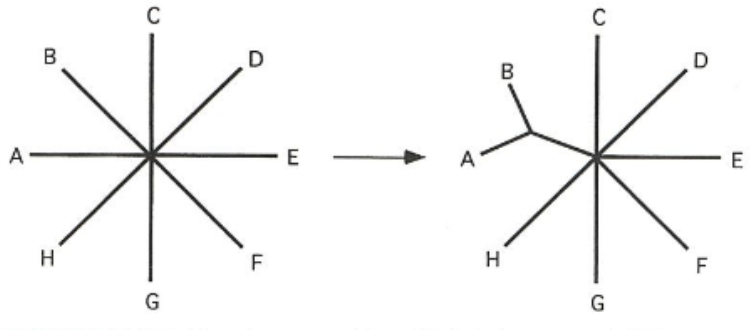
Source: *Bioinformatics,* Mount.

- All terminal nodes are joined to a single internal node
- The tree is then modified, splitting off neighbors until all nodes are bifurcating.

Construction Phylogenetic Trees

Neighbor Joining

- NJ acts by decomposing an unresolved “star” tree in several iterative steps.
- The algorithm first identifies the pair of distinct sequences (annotated as taxa, genes, or proteins) with the shortest distance between them, according to the selected substitution model.
- These taxa are then joined to a newly created node that is connected to the central node.
- The distance from each taxon to the node is calculated, and this value is used to identify the next most closely related sequence, which is then used to create a new node (hence the “joining of neighbors”).
- This process is repeated until all of the taxa are resolved into nodes throughout the tree.

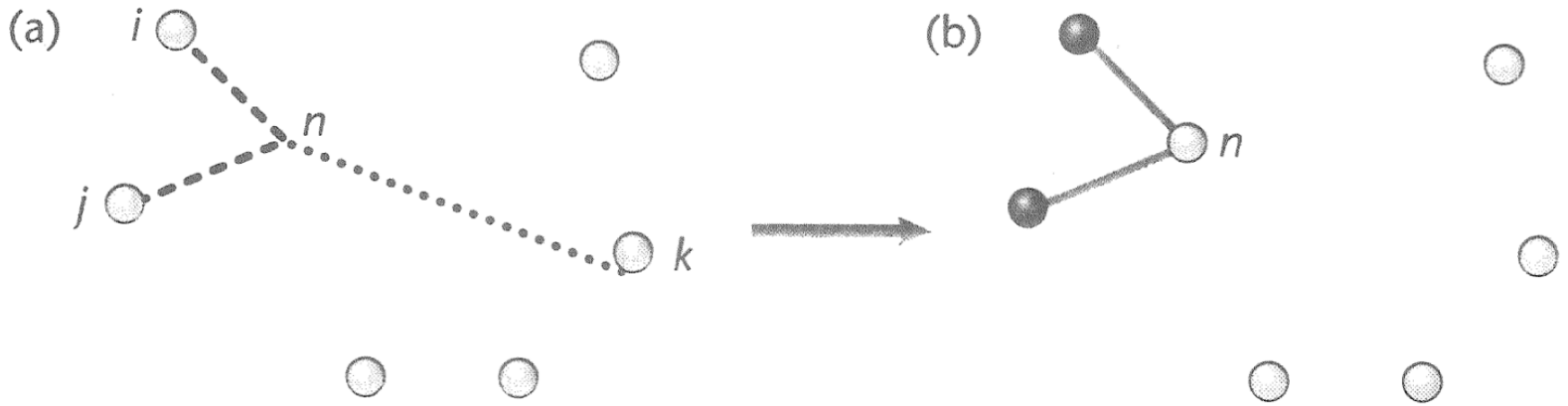


Start decomposition

Construction Phylogenetic Trees

Neighbor Joining

- We know the distance between all the nodes (distance matrix).
- Neighbor Joining identifies and joins the two closest neighbors (i, j) and create a new internal node (n).
- The original nodes (I, j) are now eliminated from the problem.
- Now, we need to find the next two closest neighbors.

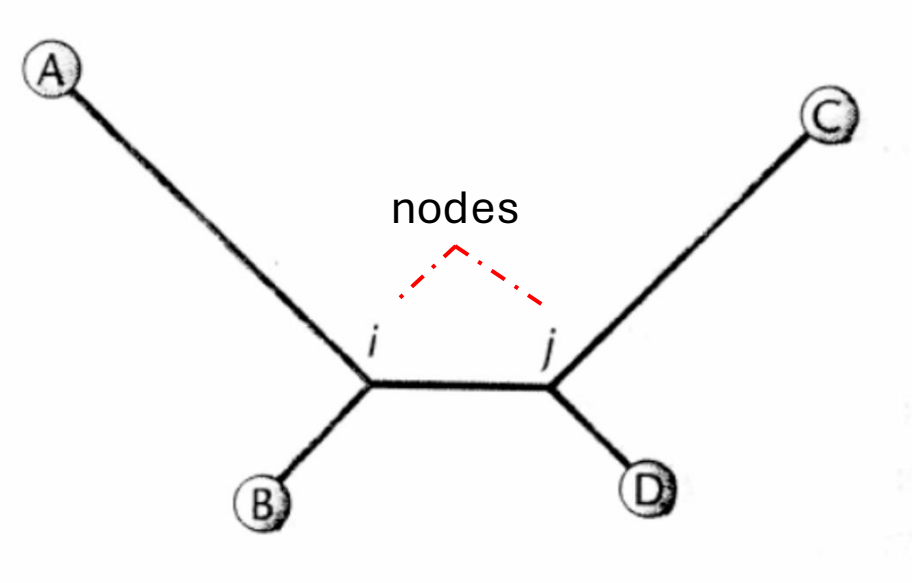


Source: *Bioinformatics and Molecular Evolution*, Higgs and Attwood

Construction Phylogenetic Trees

Neighbor Joining

- Unlike UPGMA, the two nodes with the shortest total distance are not necessary neighbors (!).



B and D have the shortest total branch length but are not neighbors as two internal nodes separate them.

Source: *Bioinformatics and Molecular Evolution*, Higgs and Attwood

Construction Phylogenetic Trees

Neighbor Joining

BOX 8.1
Calculation of distances in the neighbor-joining method

When we introduce the new node, n, we need to know its distance from all the other nodes. To calculate these distances, we will assume the distances are additive. Consider one particular node, k, among the nodes still to be connected. We may write (from Fig. 8.7a):

$$d_{ij} = d_{in} + d_{jn};\ d_{ik} = d_{in} + d_{kn};\ d_{jk} = d_{jn} + d_{kn} \qquad (8.1)$$

We may solve these simultaneous equations to find the distance of k from the new node:

$$d_{kn} = \frac{1}{2}(d_{ik} + d_{jk} - d_{ij}) \qquad (8.2)$$

This equation can be used for each of the unconnected nodes, k. We now need the distances d_{in} and d_{jn}. From the simultaneous Eqs. (8.1), we can write:

$$d_{in} = \frac{1}{2}(d_{ij} + d_{ik} - d_{jk}) \qquad (8.3)$$

The problem with this is that it depends on one particular node, k. We could call any of the disconnected nodes k and we would get a slightly different answer for d_{in} for each one. To avoid this problem, we define the quantities

$$r_i = \frac{1}{N-2}\sum_k d_{ik};\ r_j = \frac{1}{N-2}\sum_k d_{jk} \qquad (8.4)$$

Here, N is the total number of white nodes. The best estimate of d_{in} is to average Eq. (8.3) over all the $N-2$ disconnected nodes k. This gives:

$$d_{in} = \frac{1}{2}(d_{ij} + r_i - r_j) \qquad (8.5)$$

Obviously, because of additivity, $d_{jn} = d_{ij} - d_{in}$. We have now reached the situation in Fig. 8.7(b), where we have a reduced set of $N-1$ white nodes and therefore we can repeat the whole procedure again.

However, we did not yet say how to choose the two neighboring nodes, i and j, to connect in the first place. It might seem natural to choose the two with the smallest distance, d_{ij}, as we would with UPGMA. Unfortunately, there is a problem that is illustrated in Fig. 8.3(a). The closest two species in this tree are B and D, because there are short branches connecting them, **but** B and D are not neighbors, so it would be wrong to connect them. The NJ method has a way of getting around this. We calculate modified distances,

$$d^*_{ij} = d_{ij} - r_i - r_j \qquad (8.6)$$

and we choose the two species with the smallest d^*_{ij} rather than the smallest d_{ij}. It has been shown (Saitou and Nei 1987, see also Durbin *et al.* 1998) that if the distance matrix is additive, then the two species chosen by this rule must be neighbors.

Source:
Bioinformatics and Molecular Evolution,
Higgs &
Attwood

Construction Phylogenetic Trees

UPGMA

Neighbor Joining

Lab04

- Article: Rodriguez *et al*. 2009
  - https://pubmed.ncbi.nlm.nih.gov/18796461/
  - In Perusal: https://app.perusall.com/courses/bioinformatics-algorithms-spring25

- We are going to recreate the phylogenetic analysis:
  - Read Materials and Methods & Results
    - Feel free to post questions & comments in Perusal
  - Download sequences from GenBank (multi fasta file)
    - There are 20 sequences (haplotypes)
    - 2 sequences as outgroups
  - Have everything ready:
    - Files (fasta)
    - software: PHYLIP, MEGA, Tree viewer (Lab00)