Phylogenetic analysis

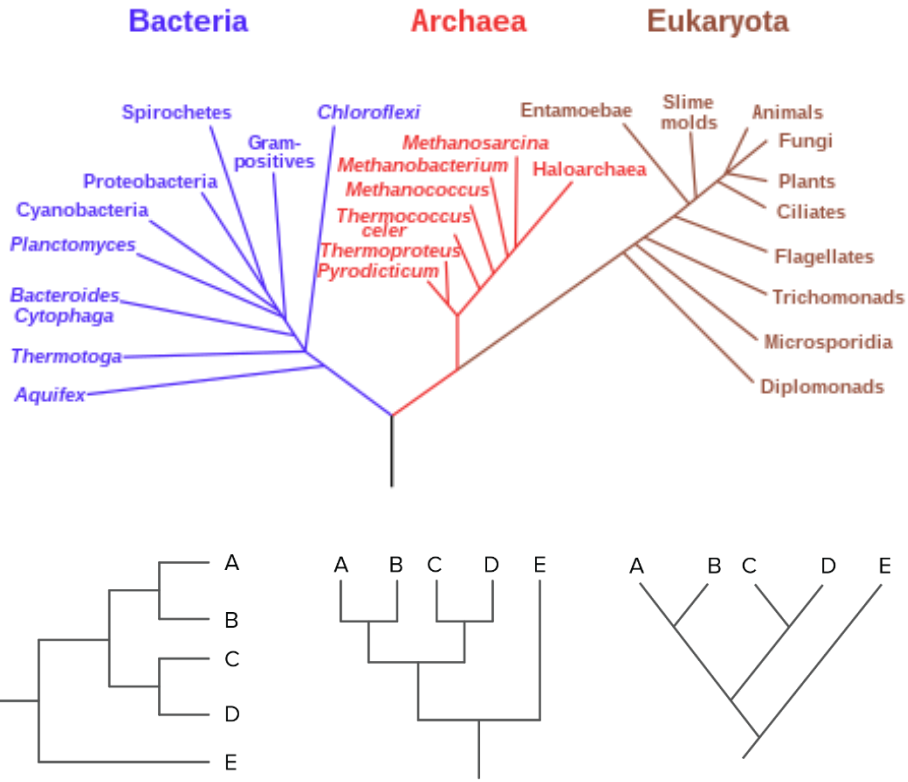
INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Dr. Jaspreet Kaur Dhanjal
Assistant Professor, Center for Computational Biology
Email ID: jaspreet@iiitd.ac.in

September 02, 2025

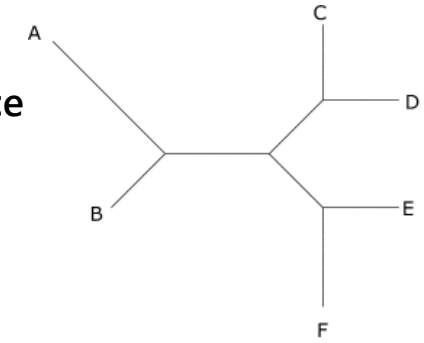
Introduction

- Phylogeny is the inference of evolutionary relationships
- All forms of life share a common origin
  - a tree represents graphical relation between organisms, species, or genomic sequence
  - aim is to deduce the correct trees for all species of life
  - to estimate the time of divergence between organisms since the time they last shared a common ancestor

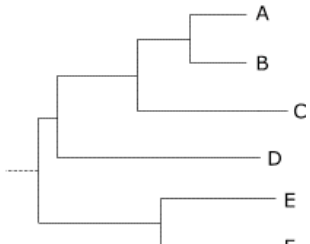


Terminology

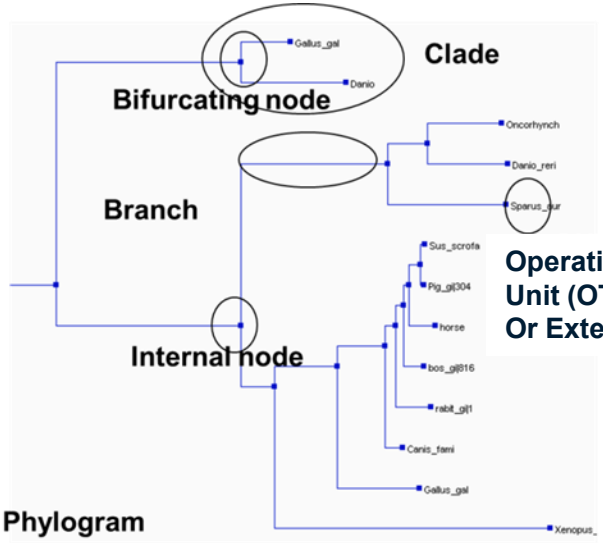
- Root: origin of evolution
- Leaves: current organisms, species, or genomic sequence
- Branches: relationship between organisms, species, or genomic sequence
- Branch length: evolutionary time



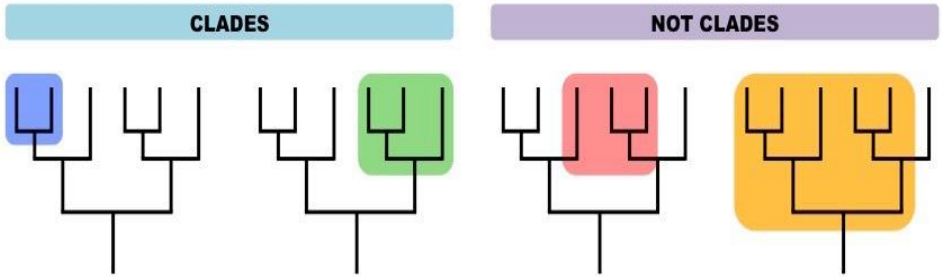
Unrooted tree



Rooted tree



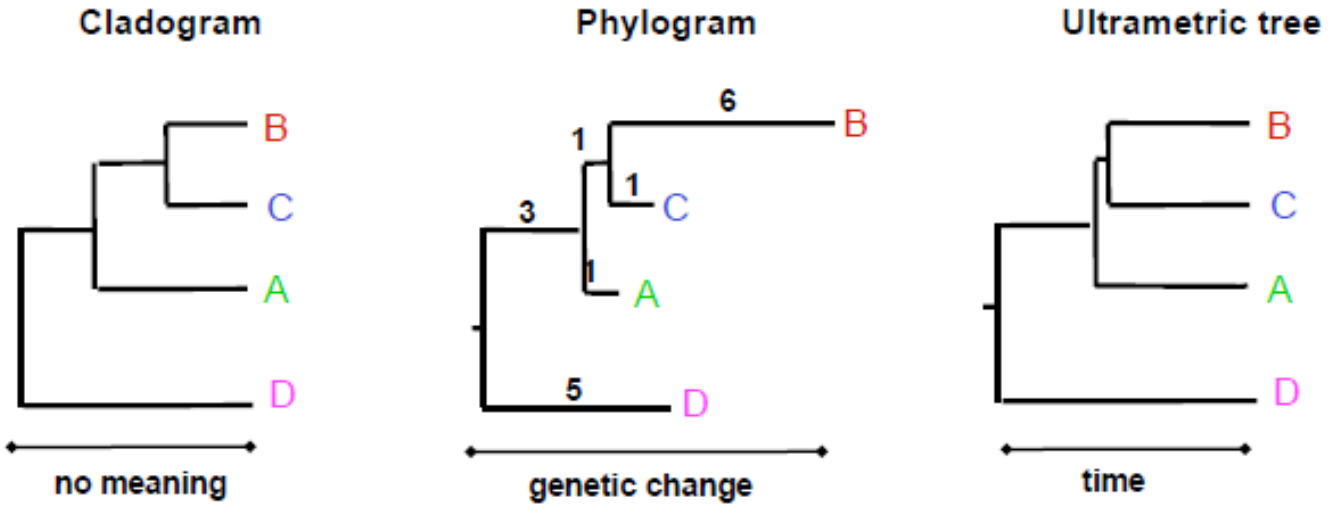
Operational Taxonomic Unit (OTU)
Or External nodes



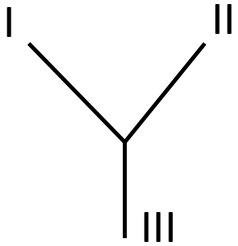
Types of rooted trees

- ***Cladograms***: Branch length have no meaning
- ***Phylograms***: Branch length represent evolutionary change
- ***Ultrametric***: Branch length represents time,

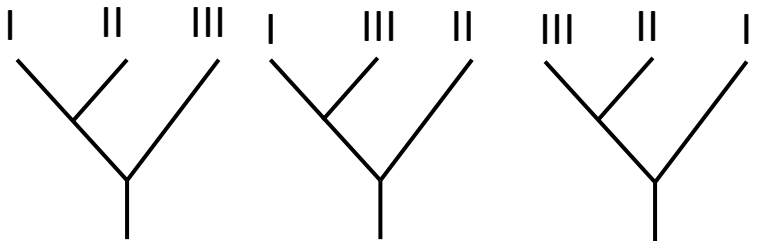
  The length from the root to the leaves are the same



Rooted and Unrooted trees



Unrooted tree



Rooted trees

# of rooted trees = $\frac{(2n-3)!}{2^{n-2}(n-2)!}$

# of unrooted trees = $\frac{(2n-5)!}{2^{n-3}(n-3)!}$

Number of possible trees

#no. of species /UTO (n)	#rooted trees	#unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
10	3.44×10^{7}	2.03×10^{6}
15	2.13×10^{14}	7.91×10^{12}
20	8.20×10^{21}	2.21×10^{20}

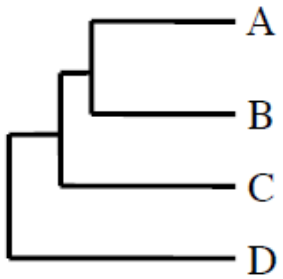
Newick format for representing trees

(, (, (,)))

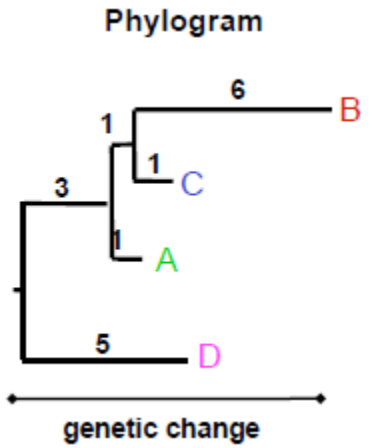
Trees can be represented in "parenthesis notation".

Each set of parentheses represents a branch-point (bifurcation), the comma separates left and right lineages.

(D,(C,(A,B))) =



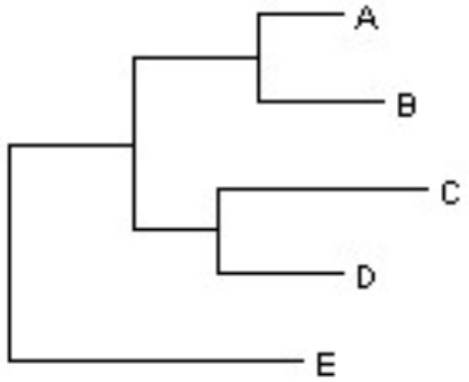
Parenthesis notation can contain sequence labels too.



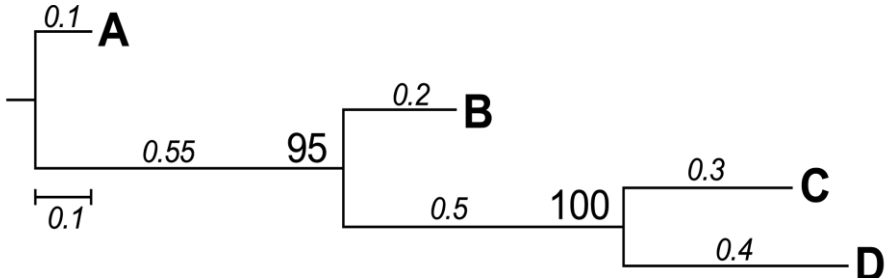
(D:5,(A:1,(C:1,B:6):1):3)

parenthesis notation can have both labels and distances.

Newick format for representing trees



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



Newick:

```
(A:0.1,(B:0.2,(C:0.3,D:0.4)100:0.5)95:0.55);
```

Phylogenetic tree construction

1. Selection of sequences for analysis
2. Multiple sequence alignment
3. Tree building
4. Tree evaluation

Selection of sequences for analysis

DNA

- Higher phylogenetic signal
- Synonymous *vs.* non-synonymous substitutions (detect negative and positive selection)

Protein

- Phylogenetic signal less predominant than in DNA
- Better to construct a tree for evolutionary distant species or genes

RNA

- rRNA often used for constructing species trees

Types of tree building methods

1. *Character-based methods*: Use the aligned sequences directly during tree inference

Taxa	Characters
Species A	ATGGCTATTCTTATAGTACG
Species B	ATCGCTAGTCTTATATTACA
Species C	TTCACTAGACCTGTGGTCCA
Species D	TTGACCAGACCTGTGGTCCG
Species E	TTGACCAGTTCTCTAGTTCG

2. *Distance-based methods*: Transform the sequence data into pairwise distances, and then use the matrix during tree building, ignoring characters

	A	B	C	D	E
Species A	----	0.20	0.50	0.45	0.40
Species B	0.23	----	0.40	0.55	0.50
Species C	0.87	0.59	----	0.15	0.40
Species D	0.73	1.12	0.17	----	0.25
Species E	0.59	0.89	0.61	0.31	----

Distance based methods

Distance based methods:

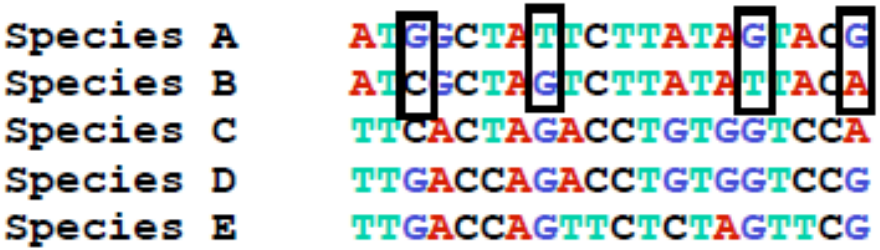
- Calculate the distances between molecular sequences using some distance metric
- A clustering method (UPGMA, Neighbor Joining) is used to infer the tree from the pairwise distance matrix
- Treat the sequence from a horizontal perspective, by calculating a single distance between entire sequences

Advantage:

- Fast
- Allow using evolutionary models

Distance calculation

- Align pairs of sequences and count the number of differences (Hamming distance)
- For an alignment of length N with n sites at which there are differences: D= (n/N*100)



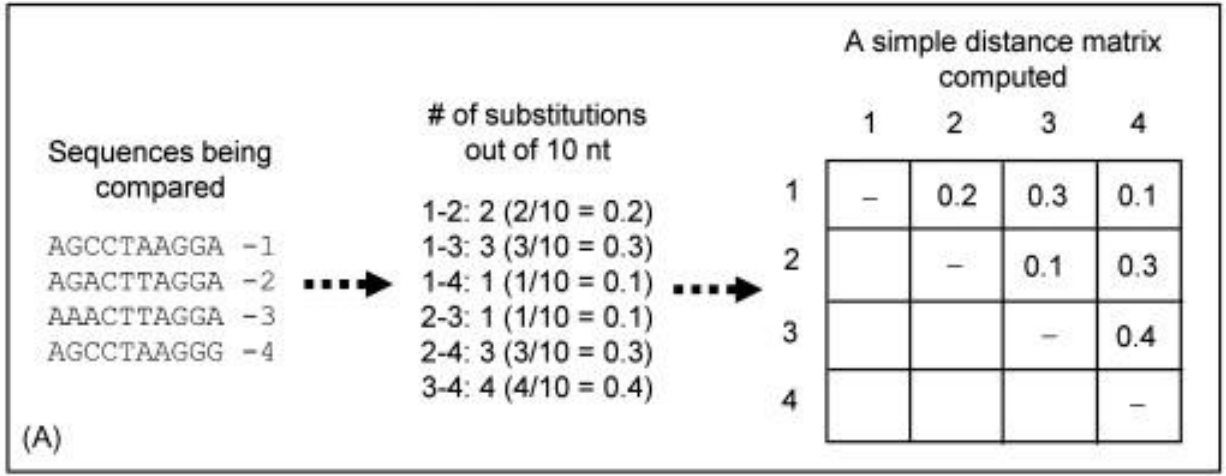
D(A,B) = 4/20

UPGMA

- Abbreviation of "Unweighted Pair Group Method with Arithmetic Mean"
- Originally developed for numeric taxonomy in 1958 by Sokal and Michener
- Simplest algorithm for tree construction, so it's fast!

- How to construct a tree with UPGMA?
  - Prepare a distance matrix
  - ***Step 1***:
    Cluster a pair of leaves (taxa) by shortest distance
  - ***Step 2***:
    Recalculate a new average distance with the new cluster and other taxa, and make a new distance matrix
    Repeat step 1 and step 2 until there are only two clusters

UPGMA - Example



UPGMA – Example 1

	A	B	C	D
A	0			
B	3	0		
C	5	4	0	
D	7	1	2	0

Matrix 1

B and **D** are the closest (**1** unit apart). Hence, **B** and **D** are clustered (**BD**) and the distance matrix is recalculated

$d(A,BD) = \{d(A,B)+d(A,D)\}/2 = (3+7)/2 = 5$

$d(BD,C) = \{d(B,C)+d(C,D)\}/2 = (4+2)/2 = 3$

	A	BD	C
A	0		
BD	5	0	
C	5	3	0

Matrix 2

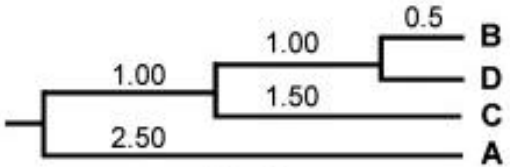
$d(A,BDC) = \{d(A,B)+d(A,D)+d(A,C)\}/3 = (3+7+5)/3 = 5$

Because this is **unweighted,** all pairwise distance are assumed to contribute equally. If this were **weighted**, the calculation would be $\{d(A,BD)+d(A,C)\}/2 = (5+5)/2 = 5$. In this example, the results are the same, but they may be different in other situations

Clustering Process Repeated

	A	BDC
A	0	
BDC	5	0

Matrix 3

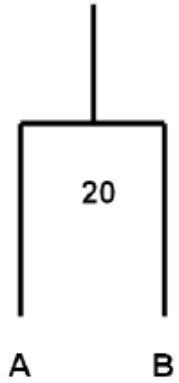


The Tree

(B) UPGMA Method

UPGMA – Example 2

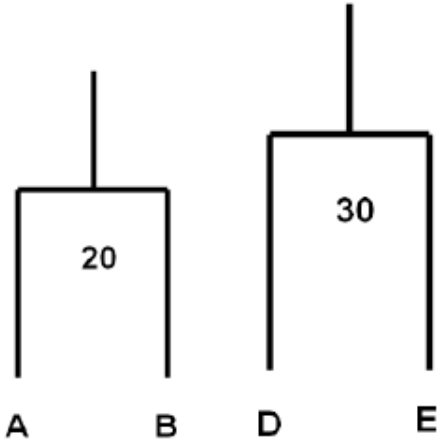
	A	B	C	D	E
A	0				
B	20	0			
C	60	50	0		
D	100	90	40	0	
E	90	80	50	30	0



- New average distance between AB and C is:
  - C to AB = (60 + 50) / 2 = 55
- Distance between D to AB is:
  - D to AB = (100 + 90) / 2 = 95
- Distance between E to AB is:
  - E to AB = (90 + 80) / 2 = 85

UPGMA – Example 2

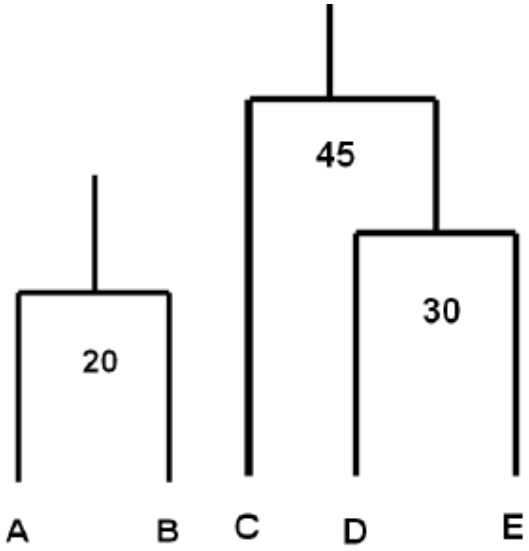
	AB	C	D	E
AB	0			
C	55	0		
D	95	40	0	
E	85	50	30	0



- New average distance between AB and DE is:
  - AB to DE = (95 + 85) / 2 = 90

UPGMA – Example 2

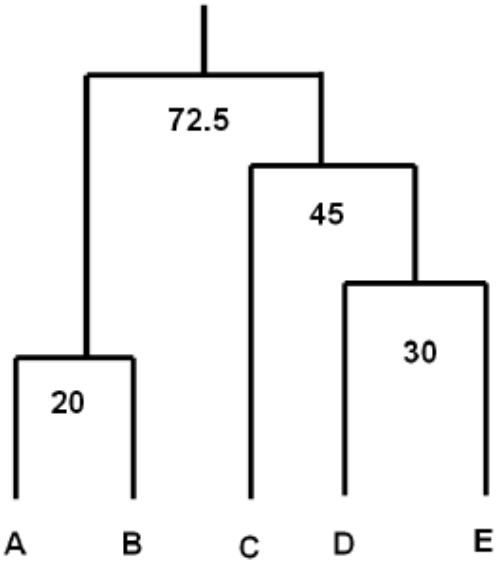
	AB	C	DE
AB	0		
C	55	0	
DE	90	45	0



- New Average distance between CDE and AB is:
  - CDE to AB = (90 + 55) / 2 = 72.5

UPGMA – Example 2

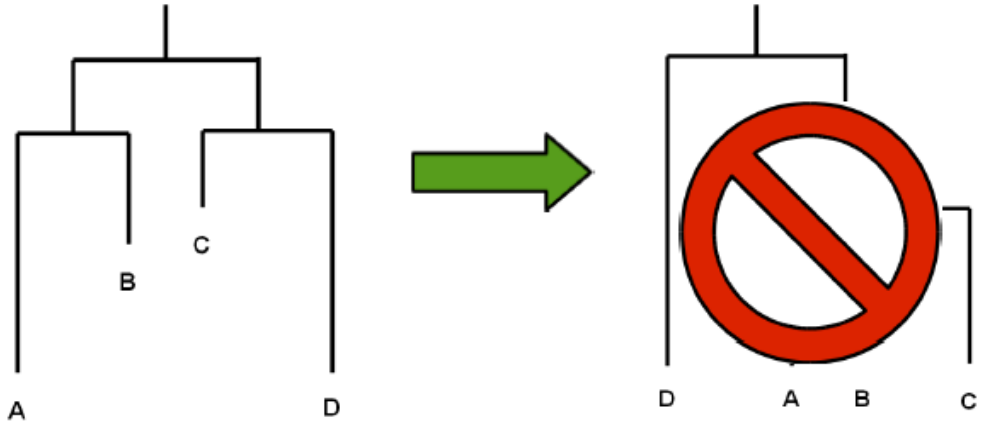
	AB	CDE
AB	0	
CDE	72.5	0



- There are only two clusters. so this completes the calculation!

Limitations of UPGMA

- Assume molecular clock (assuming the evolutionary rate is approximately constant)
- Clustering works only if the data is ultrametric
- Doesn't work the following case:



Neighbour Joining Method

~~UPGMA~~

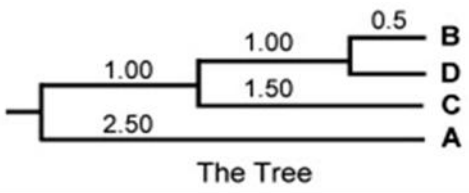
NJ method

- ~~Assumes a 'molecular clock'~~
  - ~~Rate of change is constant in sister species~~
  - ~~Hence branch lengths are equal~~
- Starts from a distance matrix
- Pairs up the most closely related taxa
- Treat this pair as new taxa

Therefore, taxa is not equidistant from the root

NJ corrects for unequal evolutionary rates between sequences using a conversion step.

	A	B	C	D
A	0			
B	3	0		
C	5	4	0	
D	7	1	2	0



The Tree

UPGMA

Neighbour Joining Method

	A	B	C	D	E
A	0				
B	4	0			
C	5	7	0		
D	2	6	3	0	
E	3	8	4	6	0

	A	B	C	D	E
A	0				
B	- 9	0			
C	-6	-7.7	0		
D	-8.3	-8	-9	0	
E	-8.7	-7.3	-11.3	-6.7	0

3.1. Join C & E with a common ancestor U

1. Calculate the net divergence r(i) for each OTU from all other OTUs

r(A) = 4 + 5 + 2 + 3 = 14 r(B) = 4 + 7 + 6 + 8 = 25
r(C) = 5 + 7 + 3 + 4 = 19 r(D) = 2 + 6 + 3 + 6 = 17
r(E) = 3 + 8 + 4 + 6 = 21

2. We calculate a new distance matrix using for each pair of OTUs

M(ij) = d(ij) - [r(i) + r(j)] / (N-2) N-2 = 5-2 = 3
M(AB) = d(AB) - [(r(A) + r(B)] / (N-2) = 4 – [14 + 25] / 3 = - 9
M(AC) = d(AC) - [(r(A) + r(C)] / (N-2) = 5 – [14 + 19] / 3 = - 6
M(AD) = d(AD) - [(r(A) + r(D)] / (N-2) = 2 – [14 + 17] / 3 = - 8.3
M(AE) = d(AE) - [(r(A) + r(E)] / (N-2) = 3 – [14 + 21] / 3 = - 8.7
M(BC) = d(BC) - [(r(B) + r(C)] / (N-2) = 7 – [25 + 19] / 3 = -7.7
M(BD) = d(BD) - [(r(B) + r(D)] / (N-2) = 6 – [25 + 17] / 3 = - 8
M(BE) = d(BE) - [(r(B) + r(E)] / (N-2) = 8 – [25 + 17] / 3 = -7.3
M(CD) = d(CD) - [(r(C) + r(D)] / (N-2) = 3 – [25 + 17] / 3 = -9
M(CE) = d(CE) - [(r(C) + r(E)] / (N-2) = 4 – [25 + 17] / 3 = - 11.3
M(DE) = d(DE) - [(r(D) + r(E)] / (N-2) = 6 – [25 + 17] / 3 = - 6.7

Neighbour Joining Method

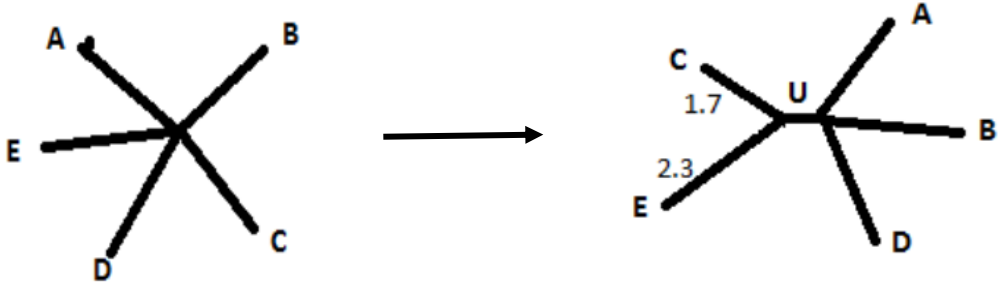
	A	B	C	D	E
A	0				
B	4	0			
C	5	7	0		
D	2	6	3	0	
E	3	8	4	6	0

r(A) = 4 + 5 + 2 + 3 = 14 r(B) = 4 + 7 + 6 + 8 = 25
r(C) = 5 + 7 + 3 + 4 = 19 r(D) = 2 + 6 + 3 + 6 = 17
r(E) = 3 + 8 + 4 + 6 = 21

3.2. Calculate the branch length from the internal node U to the external OTUs A and E with following formula.

S(CU) = d(CE) / 2 + [r(C) -r(E)] / 2(N-2)
S(CU) = 4/2 + [19 – 21] / 2 *3 = 2 + { [-2] /6 } = 1.7
S(EU) = d(CE) - S(CU) = 4 – 1.7 = 2.3

4. Change tree topology



Neighbour Joining Method

	U	A	B	D
U	0			
A	2	0		
B	5.5	4	0	
D	2.5	2	6	0

	U	A	B	D
U	0			
A	-7	0		
B	-7.25	-7.75	0	
D	-7.75	-7.25	-7	0

3.1. Join A & B with a common ancestor U1

d(AU) = [d(AC) + d(AE) - d(CE)] / 2 = [5 + 3 – 4] / 2 = 2
d(BU) = [d(BC) + d(BE) - d(CE)] / 2 = [7 + 8 – 4] /2 = 5.5
d(DU) = [d(DC) + d(DE) - d(CE)] / 2 = [3 + 6 – 4] /2 = 2.5

1. Calculate the net divergence r(i) for each OUT from all other OTUs

r(U) = 2 + 5.5 + 2.5 =10 r(A) = 2 + 4 + 2 = 8
r(B) = 5.5 + 4 + 6 = 15.5 r(D) = 2.5 + 2 + 6 = 10.5

2. We calculate a new distance matrix using for each pair of OUTs

M(ij) = d(ij) - [r(i) + r(j)] / (N-2) N-2 = 4-2 = 2
M(UA) = 2 – [10 + 8] /2 = -7
M(UB) = 5.5 – [10 + 15.5] /2 = -7.25
M(UD) = 2.5 – [10 + 10.5] /2 = -7.75
M(AB) = 4 – [8 + 15.5] /2 = -7.75
M(AD) = 2 – [8 + 10.5] / 2 = -7.25
M(BD) = 6 – [15.5 + 10.5] /2 = - 7

Neighbour Joining Method

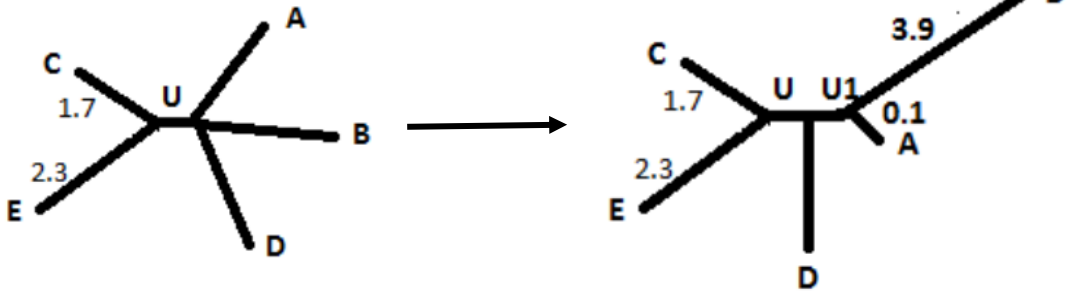
	U	A	B	D
U	0			
A	2	0		
B	5.5	4	0	
D	2.5	2	6	0

r(U) = 2 + 5.5 + 2.5 =10 r(A) = 2 + 4 + 2 = 8
r(B) = 5.5 + 4 + 6 = 15.5 r(D) = 2.5 + 2 + 6 = 10.5

3.2. Calculate the branch length from the internal node U to the external OTUs A and E with following formula.

S(AU1) = d(AB) / 2 + [r(A) -r(B)] / 2(N-2)
S(AU1)= 4 /2 + { [8 – 15.5] / 2 * 2} = 2 + {[-7.5] /4} = 0.1
S(BU1) = d(AB) - d(AU1) = 4– 0.1 = 3.9

4. Change tree topology



Neighbour Joining Method

	U	U1	D
U	0		
U1	1.7	0	
D	2.5	2	0

	U	U1	D
U	0		
U1	-6.2	0	
D	-6.2	-6.2	0

3.1. Join U & D with a common ancestor U2

d(UU1) = [d(UA) + d(UB) - d(AB)] / 2 = [2 + 5.5 – 4] / 2 = 1.75
d(DU) = [d(DC) + d(DE) - d(CE)] / 2 = [3 + 6 – 4] /2 = 2.5
d(DU1) = [d(DA) + d(DB) - d(AB)] / 2 = [2 + 6 – 4] /2 = 2

1. Calculate the net divergence r(i) for each OUT from all other OTUs

r(U) = 1.7 + 2.5 = 4.2 **r(U1) = 1.7 + 2 = 3.7**
r(D) = 2.5 + 2 = 4.5

2. We calculate a new distance matrix using for each pair of OUTs

M(ij) = d(ij) - [r(i) + r(j)] / (N-2) **N-2 = 3-2 = 1**
M(UU1) = 1.7 – [4.2 + 3.7] /1 = -6.2
M(UD) = 2.5 – [4.2 + 4.5] /1 = -6.2
M(U1D) = 2 – [3.7 + 4.5] /1 = -6.2

Neighbour Joining Method

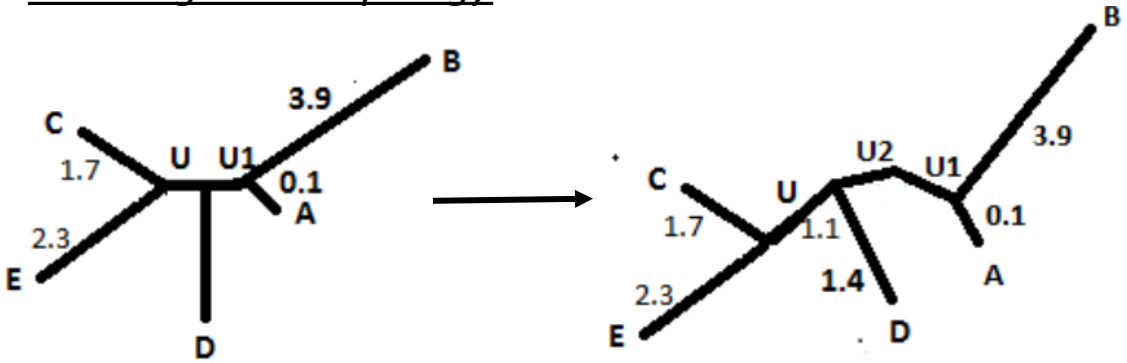
	U	U1	D
U	0		
U1	1.7	0	
D	2.5	2	0

r(U) = 1.7 + 2.5 = 4.2 **r(U1) = 1.7 + 2 = 3.7**
r(D) = 2.5 + 2 = 4.5

3.2. Calculate the branch length from the internal node U to the external OTUs A and E with following formula.

S(DU2) = d(UD) / 2 + [r(D) -r(U)] / 2(N-2)
S(DU2) = 2.5 /2 + { [4.5 – 4.2] / 2 * 1} = 1.25 + { 0 .15} = 1.4
S(UU2) = d(DU) - S(DU2) = 2.5 – 1.4 = 1.1

4. Change tree topology

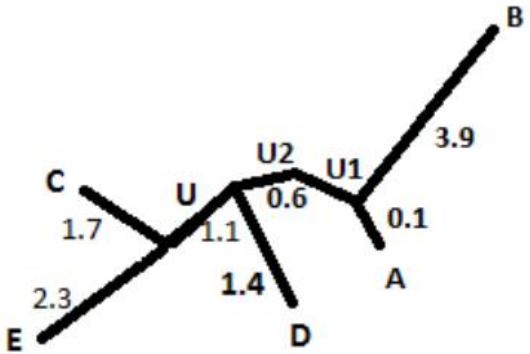


Neighbour Joining Method

	U1	U2
U1	0	
U2	0.6	0

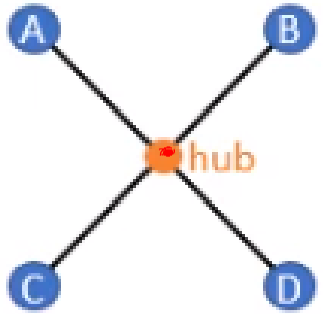
d(U1U2) = [d(U1D) + d(U1U) - d(DU)] / 2 = [2 + 1.7 – 2.5] / 2 = 0.6

Stop at this stage because matrix has reduce to 2 X 2



Neighbour Joining Method

	A	B	C	D
A	-	17	21	27
B	-	-	12	18
C	-	-	-	14
D	-	-	-	-



Neighbour Joining Method

Round 1 Step 1

Compute $\mathbf{r'_i}$ for each terminal node

	A	B	C	D
A	-	17	21	27
B	-	-	12	18
C	-	-	-	14
D	-	-	-	-

	A	B	C	D	r'_i
A	-	17	21	27	
B	-	-	12	18	
C	-	-	-	14	
D	-	-	-	-	

d	A	B	C	D	r'_i
A	-	17	21	27	32.5
B	-	-	12	18	23.5
C	-	-	-	14	23.5
D	-	-	-	-	29.5

$$r_i' = \Sigma_j Di,j/n-2$$

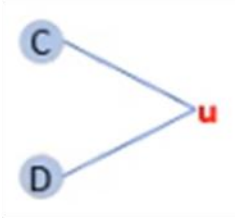
Neighbour Joining Method

Round 1 Step 2

Compute $d'_{i,j}$ for each terminal node

d	A	B	C	D	r'_i
A	-	17	21	27	32.5
B	-	-	12	18	23.5
C	-	-	-	14	23.5
D	-	-	-	-	29.5

d'	A	B	C	D
A	-	-39	-35	-35
B	-	-	-35	-35
C	-	-	-	-39
D	-	-	-	-



$$d'_{i,j} = d_{i,j} - r'_i - r'_j$$

Neighbour Joining Method

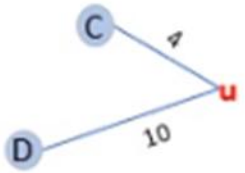
Round 1 Step 3

Calculate branch lengths

d	A	B	C	D	r'_i
A	-	17	21	27	32.5
B	-	-	12	18	23.5
C	-	-	-	14	23.5
D	-	-	-	-	29.5

$$V_i = 0.5\,(d_{i,j}) + 0 \cdot 5(r'_i - r'_j)$$

$$V_j = 0.5\,(d_{i,j}) + 0 \cdot 5(r'_j - r'_i)$$



Neighbour Joining Method

Round 2

	A	B	C	D
A	-	17	21	27
B	-	-	12	18
C	-	-	-	14
D	-	-	-	-

$$d'_{i,j,k} = (d_{i,k} + d_{j,k} - d_{i,j})/2$$

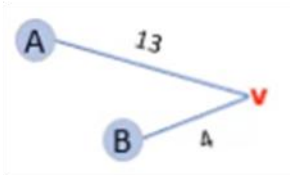
	A	B	CD	r'_i
A	-	17	17	34
B	-	-	8	25
CD	-	-	-	25

Neighbour Joining Method

Round 2

	A	B	CD	r
A	-	17	17	34
B	-	-	8	25
CD	-	-	-	25

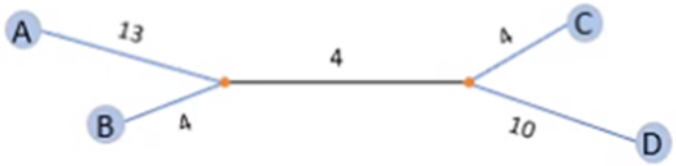
	A	B	CD
A	-	-42	-42
B	-	-	-42
CD	-	-	-



Neighbour Joining Method

Round 3

	AB	CD
AB	-	4
CD	-	-



	A	B	C	D
A	-	17	21	27
B	-	-	12	18
C	-	-	-	14
D	-	-	-	-