BIO 285/CSCI 285/MATH 285 Bioinformatics Programming Lecture 3 Phylogenetic Tree - Neighbor Joining Algorithm

Instructor: Lei Qian

Fisk University

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
- Disadvantage:
  - sequences reduced to one number

When UPGMA fails …

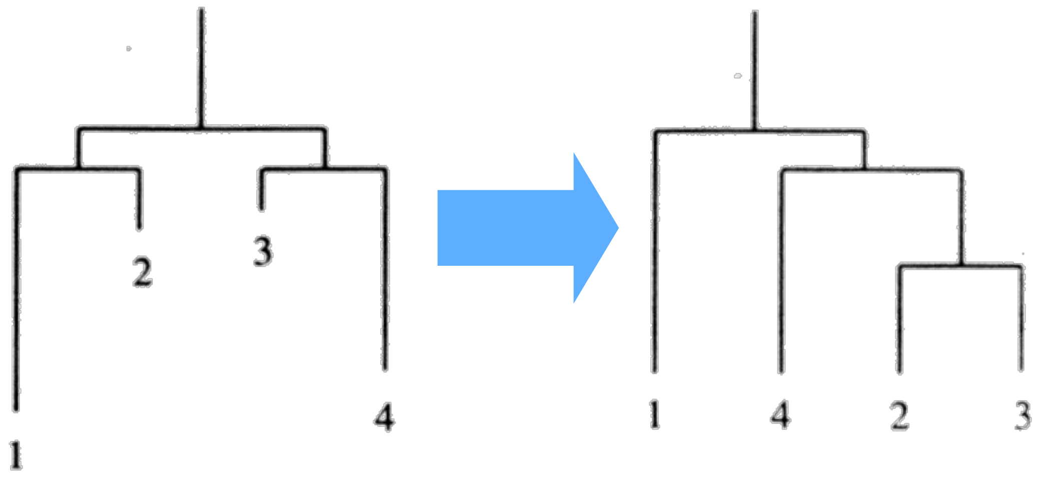


Figure 7.5 *A tree (left) that is reconstructed incorrectly by UPGMA (right).*

Neighbor Joining Algorithm

- unlike UPGMA
  - doesn’t make molecular clock assumption
  - produces unrooted trees

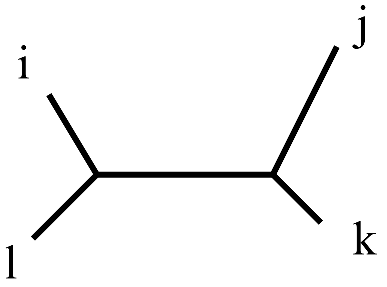
- does assume *additivity*: distance between pair of leaves is sum of lengths of edges connecting them

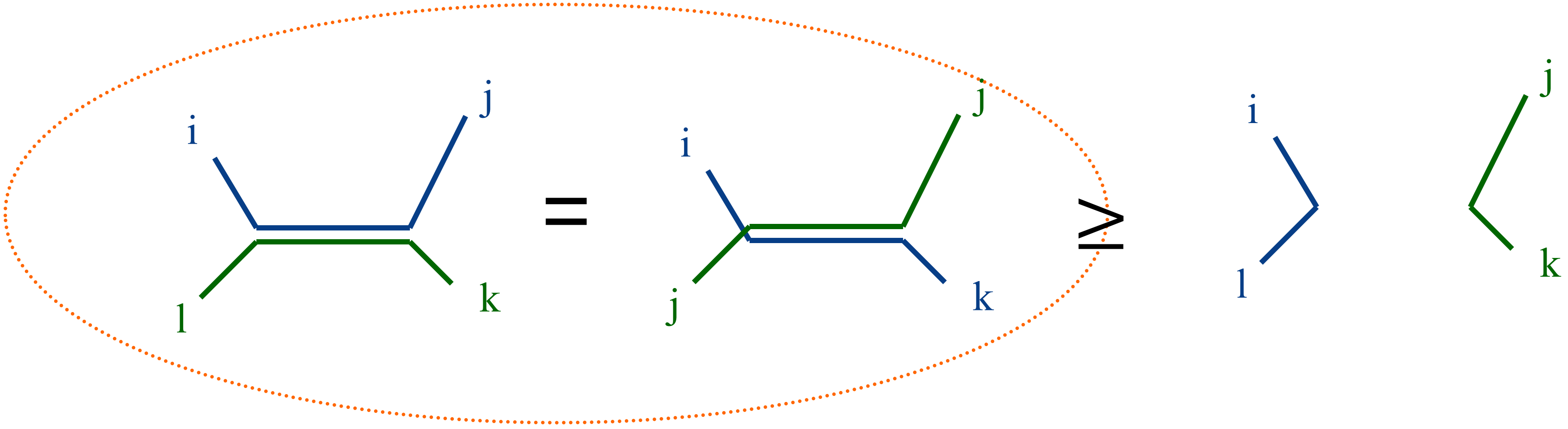
- like UPGMA, constructs a tree by iteratively joining subtrees

- two key differences
  - how pair of subtrees to be merged is selected on each iteration
  - how distances are updated after each merge

Testing for Additivity

- for every set of four leaves, *i*, *j*, *k*, and *l*, two of the distances $d_{ij}+d_{kl}$, $d_{ik}+d_{jl}$ and $d_{il}+d_{jk}$ must be equal and not less than the third





Compensating for long edges

Introduce “correction terms”

$$u_i = \frac{\Sigma_{i \neq k} D_{ik}}{n-2}$$

Average dist. to other taxa

“Corrected” distances:

$$\widehat{D}_{ij} = D_{ij} - u_i - u_j$$

Distances are reduced for pairs that are far away from all other species: They may be close to each other.

Neighbor joining

Repeat the following until only two leaves remain:

1. Build a Q-matrix such that each entry
   $$q_{ij} = (n-2)d_{ij} - \Sigma_{\mathrm{i}\neq k} d_{ik} - \Sigma_{\mathrm{j}\neq\mathrm{k}} d_{jk} \quad (= (n-2)\hat{d}_{ij})$$

   Select the (i, j) such that $q_{i,j}$ is minimum

2. Define a new leaf k whose distances to i and j are
   $$d_{ik} = \frac{1}{2}d_{ij} + \frac{1}{2}(u_i - u_j) = \frac{1}{2}d_{ij} + \frac{1}{2(n-2)}(\Sigma_{i\neq k} d_{ik} - \Sigma_{j\neq k} d_{jk})$$
   $$d_{jk} = \frac{1}{2}d_{ij} + \frac{1}{2}(u_j - u_i) = \frac{1}{2}d_{ij} - \frac{1}{2(n-2)}(\Sigma_{i\neq k} d_{ik} - \Sigma_{j\neq k} d_{jk})$$

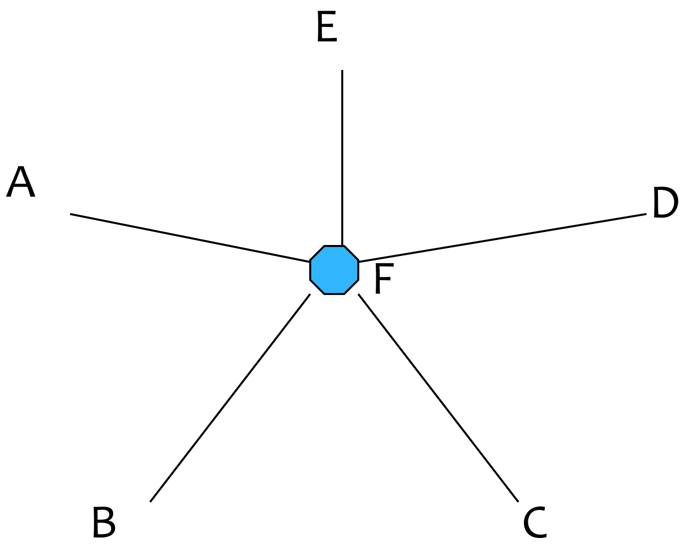
2. Compute the distance from k to every other leaf r
   $$d_{kr} = \frac{d_{ir} + d_{jr} - d_{ij}}{2}$$

3. Delete i and j.

Connect the 2 remaining leaves by a branch of length d_{ij}

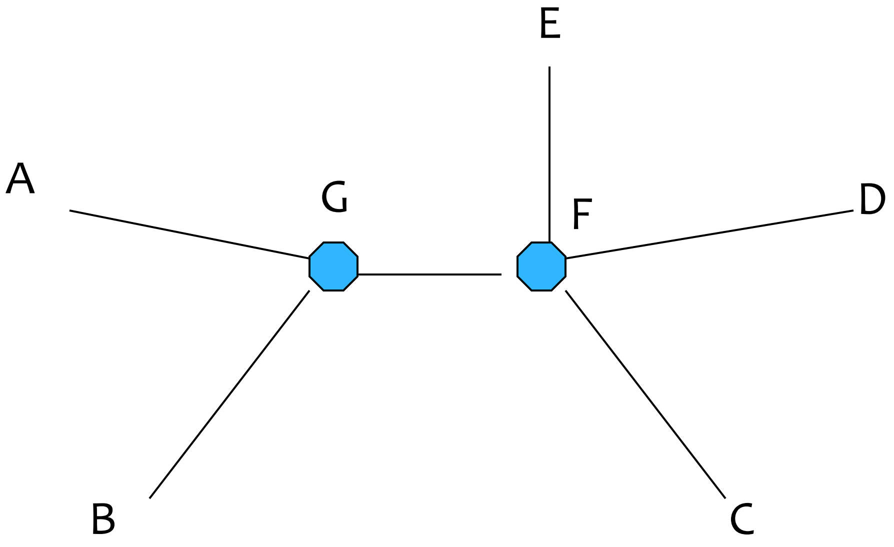
Neighbor-joining

- First start with a “star tree”



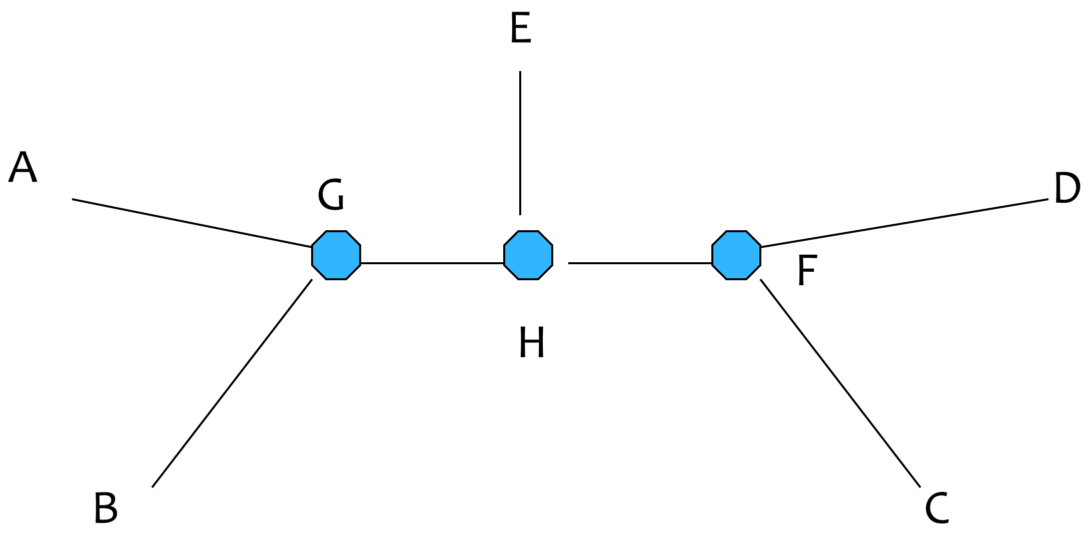
Neighbor-joining

- Combine the closest two nodes (from distance matrix)
  - In our case it is node A and B at distance 3
  - We can now apply the NJ algorithm to (G, E, D, C)



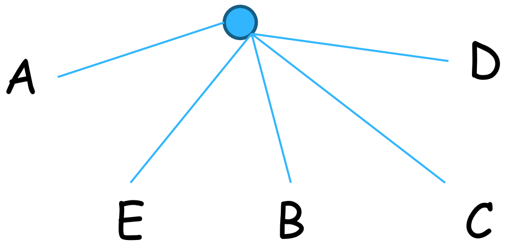
Neighbor-joining

- Repeat this until you have added n-2 nodes
  - N-2 will make it a binary tree, so we only have to include one more node.



Neighbor Joining

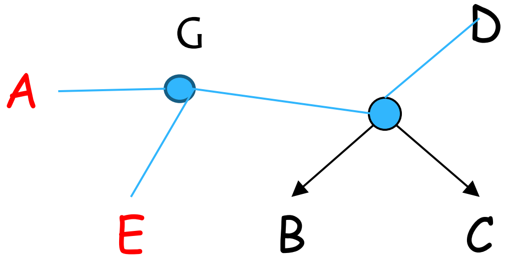
$$q_{ij} = (n-2)d_{ij} - \Sigma_{i\neq k} d_{ik} - \Sigma_{j\neq k} d_{jk}$$



d	A	B	C	D	E
A	0	10	12	8	7
B	10	0	4	4	14
C	12	4	0	6	16
D	8	4	6	0	12
E	7	14	16	12	0

$$q_{AB} = (5-2) * d_{AB} - s_A - s_B$$
$$= 3 * 10 - 37 - 32 = -39$$

$$s_A = 37, s_B = 32, s_C = 38, s_D = 30, s_E = 49$$



Q	A	B	C	D	E
A	0	-39	-39	-43	-65
B	-39	0	-58	-50	-39
C	-39	-58	0	-50	-39
D	-43	-50	-50	0	-43
E	-65	-39	-39	-43	0

Neighbor Joining

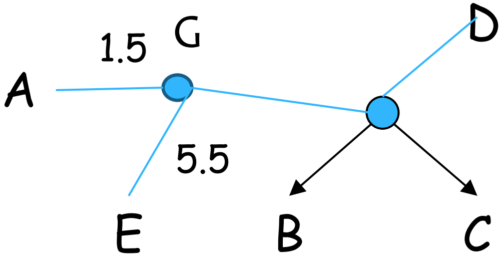
$$d_{ik} = \frac{1}{2}D_{ij} + \frac{1}{2}\left(u_i - u_j\right) = \frac{1}{2}D_{ij} + \frac{1}{2(n-2)}\left(\Sigma_{i\neq k}D_{ik} - \Sigma_{j\neq k}D_{jk}\right)$$

$$d_{jk} = \frac{1}{2}D_{ij} + \frac{1}{2}\left(u_j - u_i\right) = \frac{1}{2}D_{ij} - \frac{1}{2(n-2)}\left(\Sigma_{i\neq k}D_{ik} - \Sigma_{j\neq k}D_{jk}\right)$$

$$d_{AG} = \frac{1}{2}d_{AE} + \frac{1}{6}(s_A - s_E) = \frac{7}{2} + \frac{37-49}{6} = 1.5$$

$$d_{EG} = \frac{1}{2}d_{AE} - \frac{1}{6}(s_A - s_E) = \frac{7}{2} - \frac{37-49}{6} = 5.5$$

d	A	B	C	D	E
A	0	10	12	8	7
B	10	0	4	4	14
C	12	4	0	6	16
D	8	4	6	0	12
E	7	14	16	12	0



$$s_A = 37, s_B = 32, s_C = 38, s_D = 30, s_E = 49$$

$$d_{kr} = \frac{d_{ir} + d_{jr} - d_{ij}}{2}$$

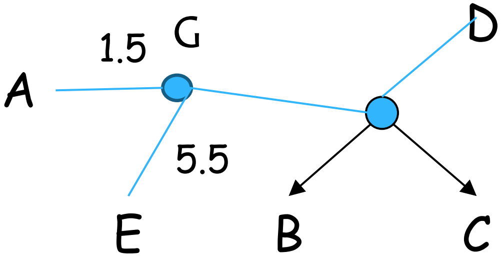
$$d_{BG} = (d_{BA} + d_{BE} - d_{AE})/2 = (10 + 14 - 7)/2 = 8.5$$

d	B	C	D	G
B	0	4	4	8.5
C	4	0	6	10.5
D	4	6	0	6.5
G	8.5	10.5	6.5	0

Update distance matrix

Neighbor Joining

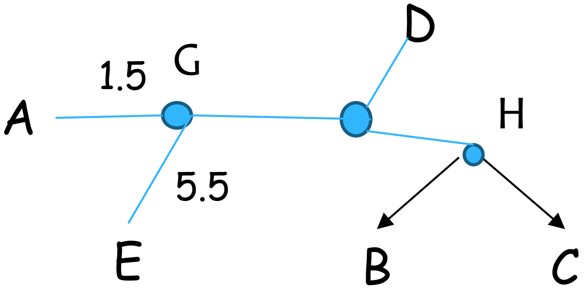
$q_{ij} = (n-2)d_{ij} - \Sigma_{\mathrm{i}\neq k} d_{ik} - \Sigma_{\mathrm{j}\neq \mathrm{k}} d_{jk}$



d	B	C	D	G
B	0	4	4	8.5
C	4	0	6	10.5
D	4	6	0	6.5
G	8.5	10.5	6.5	0

$s_B = 16.5, s_C = 20.5, s_D = 16.5, s_G = 25.5$

$$q_{BC} = 2 * d_{BC} - s_B - s_C$$
$$= 2 * 4 - 16.5 - 20.5 = -29$$



Q	B	C	D	G
B	0	-29	-25	-19
C	-29	0	-25	-25
D	-25	-25	0	-29
G	-19	-25	-29	0

Q- matrix

Neighbor Joining

$$d_{ik} = \frac{1}{2}D_{ij} + \frac{1}{2}(u_i - u_j) = \frac{1}{2}D_{ij} + \frac{1}{2(n-2)}(\Sigma_{i\neq k}D_{ik} - \Sigma_{j\neq k}D_{jk})$$

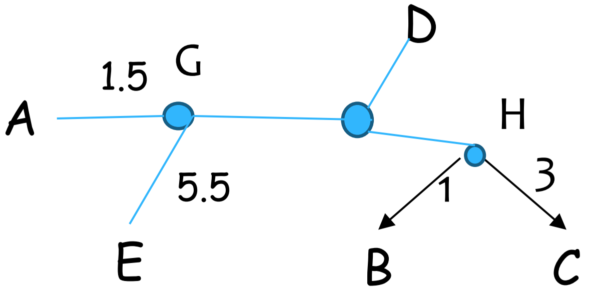
$$d_{jk} = \frac{1}{2}D_{ij} + \frac{1}{2}(u_j - u_i) = \frac{1}{2}D_{ij} - \frac{1}{2(n-2)}(\Sigma_{i\neq k}D_{ik} - \Sigma_{j\neq k}D_{jk})$$

$$d_{BH} = \frac{1}{2}d_{BC} + \frac{1}{4}(s_B - s_C) = \frac{4}{2} + \frac{16.5 - 20.5}{4} = 1$$

$$d_{CH} = \frac{1}{2}d_{BC} - \frac{1}{4}(s_B - s_C) = \frac{4}{2} - \frac{16.5 - 20.5}{4} = 3$$

d	B	C	D	G
B	0	4	4	8.5
C	4	0	6	10.5
D	4	6	0	6.5
G	8.5	10.5	6.5	0

$$s_B = 16.5, s_C = 20.5, s_D = 16.5, s_G = 25.5$$



$$d_{kr} = \frac{d_{ir} + d_{jr} - d_{ij}}{2}$$

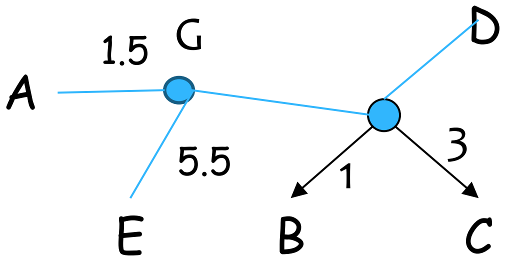
$$d_{DH} = (d_{DB}+d_{DC} - d_{BC})/2 = (4 + 6 - 4)/2 = 3$$

d	H	D	G
H	0	3	7.5
D	3	0	6.5
G	7.5	6.5	0

Update distance matrix

Neighbor Joining

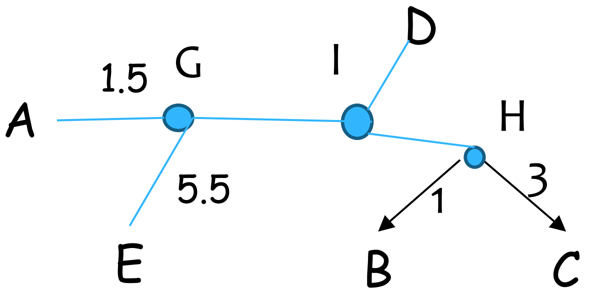
$q_{ij} = (n-2)d_{ij} - \Sigma_{\mathrm{i \neq k}} d_{ik} - \Sigma_{\mathrm{j \neq k}} d_{jk}$



d	H	D	G
H	0	3	7.5
D	3	0	6.5
G	7.5	6.5	0

$s_H = 10.5, s_D = 9.5, s_G = 14$

$$q_{HD} = 1 * d_{HD} - s_H - s_D$$
$$= 3 - 10.5 - 9.5 = -17$$



Q	H	D	G
H	0	-17	-17
D	-17	0	-17
G	-17	--17	0

Q- matrix

Neighbor Joining

$$d_{ik} = \frac{1}{2}D_{ij} + \frac{1}{2}(u_i - u_j) = \frac{1}{2}D_{ij} + \frac{1}{2(n-2)}(\Sigma_{i\neq k}D_{ik} - \Sigma_{j\neq k}D_{jk})$$

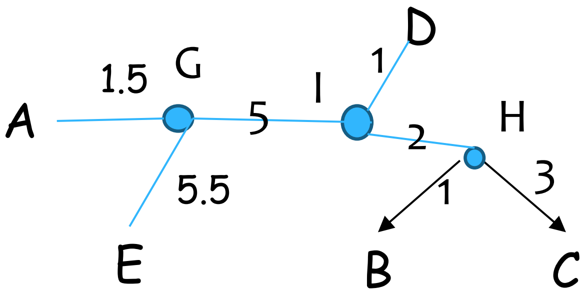
$$d_{jk} = \frac{1}{2}D_{ij} + \frac{1}{2}(u_j - u_i) = \frac{1}{2}D_{ij} - \frac{1}{2(n-2)}(\Sigma_{i\neq k}D_{ik} - \Sigma_{j\neq k}D_{jk})$$

$$d_{DI} = \frac{1}{2}d_{DH} + \frac{1}{2}(s_D - s_H) = \frac{3}{2} + \frac{9.5 - 10.5}{2} = 1$$

$$d_{HI} = \frac{1}{2}d_{DI} - \frac{1}{2}(s_D - s_H) = \frac{3}{2} - \frac{9.5 - 10.5}{4} = 2$$

d	H	D	G
H	0	3	7.5
D	3	0	6.5
G	7.5	6.5	0

$$s_H = 10.5, s_D = 9.5, s_G = 14$$



$$d_{kr} = \frac{d_{ir} + d_{jr} - d_{ij}}{2}$$

d	I	G
I	0	5
G	5	0

$$d_{GI} = (d_{GH} + d_{GD} - d_{HD})/2 = (7.5 + 6.5 - 3)/2 = 5$$

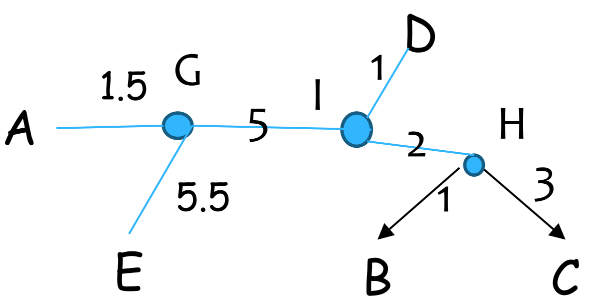
Update distance matrix

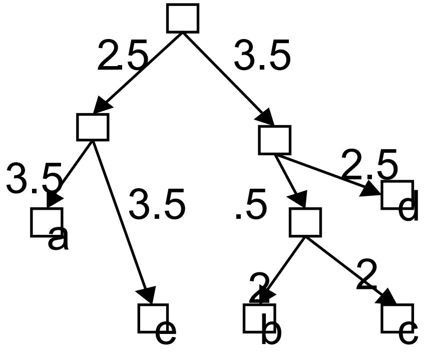
Neighbor Joining

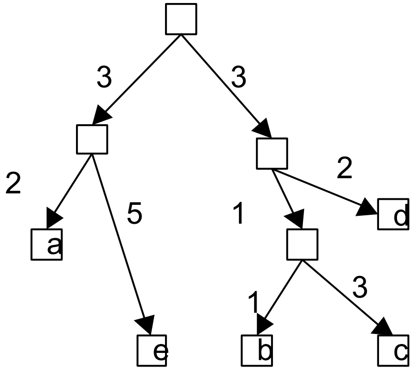
Original Distance

d	A	B	C	D	E
A	0	10	12	8	7
B	10	0	4	4	14
C	12	4	0	6	16
D	8	4	6	0	12
E	7	14	16	12	0

NJ







NJ Tree Distance

d	A	B	C	D	E
A	0	9.5	11.5	7.5	7
B	9.5	0	4	4	13.5
C	11.5	4	0	6	15.5
D	7.5	4	6	0	11.5
E	7	13.5	15.5	11.5	0

UPGMA Tree Distance

d	A	B	C	D	E
A	0	10	12	10	7
B	10	0	4	4	13
C	12	4	0	6	15
D	10	4	6	0	13
E	7	13	15	13	0

Neighbor Joining

UIUC TeachEnG Neighbor Joining Algorithm game

http://teacheng.illinois.edu/PhylogeneticTree