

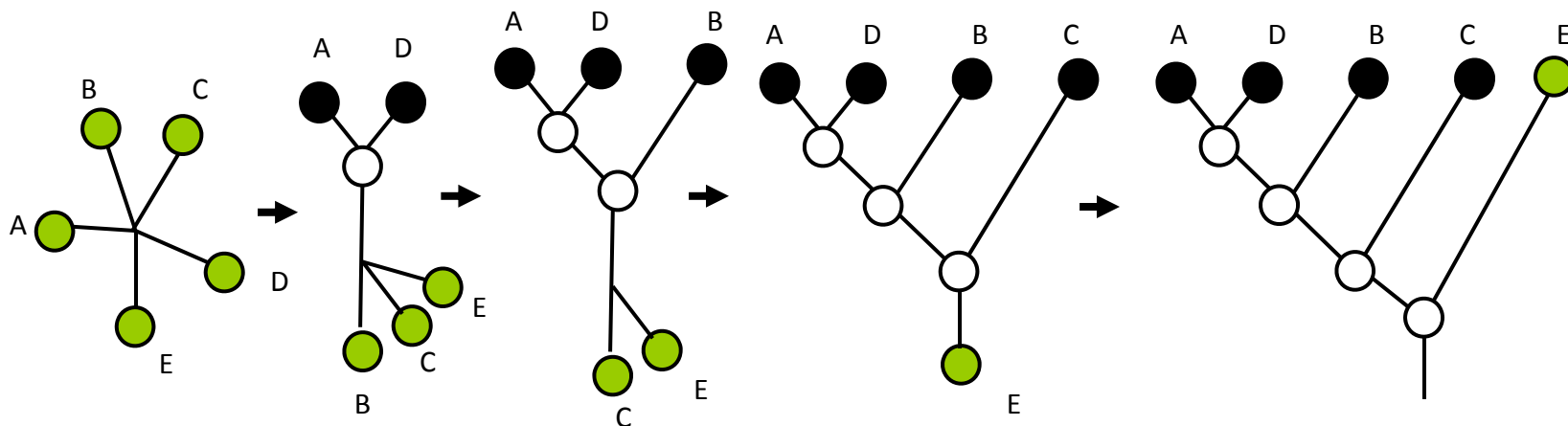
Distances Methods

getting distances from sequences

Sequence A	A	A	C	T
Sequence B	A	T	G	T
Sequence C	G	T	A	T
Sequence D	G	T	G	A

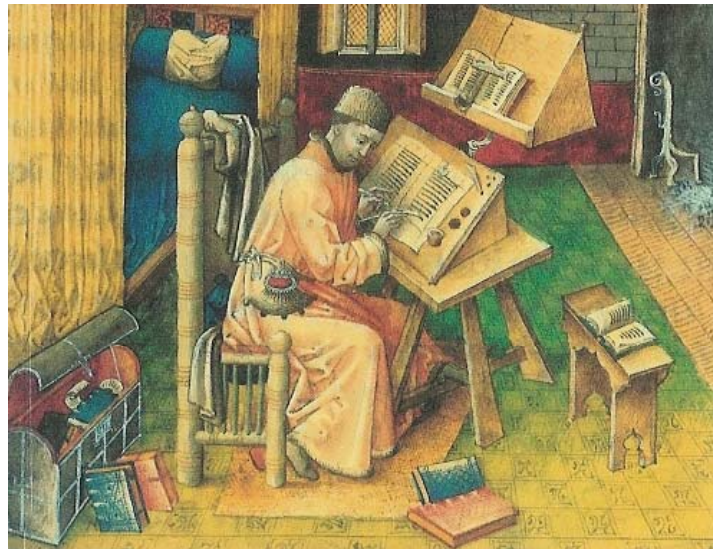
d_{ij}				
	A	B	C	D
A	-	-	-	-
B	2	-	-	-
C	3	2	-	-
D	4	2	2	-

principle of clustering using distances



The Drunk Monk

One unfortunate scribe "Philinus" was working late into the night during the 14th century. This scribe (possibly captured in the picture)* was diligent and hardworking, and always produced remarkably good copies (although he was rather partial to honey mead).



On that late evening he became confused after a few glasses of honey mead and could not remember which manuscript he was working from. He looked back at a section of the text he had written and compared it to a number of versions he had in his possession.

- A Buryed att caane thus seythe the Cronycle
- B Buryed at caane thus seythe the cronyclere
- C And buryed at cane thus seythe the Cronycle
- D Buryed at cane thus says so the Cronycle
- Philus Buried at cane thus seythe so the Cronycle

Multiple sequence alignment

A	Buryed	att	caane	thus	seythe	the	Cronycle	
B	Buryed	at	caane	thus	seythe	the	cronyclere	
C	And	buryed	at	cane	thus	seythe	the	Cronycle
D	Buryed	at	cane	thus	says	so	the	Cronycle
Philus	Buried	at	cane	thus	seythe	so	the	Cronycle

site patterns

- He realized he could recode the data. For example the above text recoded as:

-

- A 0111010
- B 0101011
- C 1100010
- D 0100100
- Philius 0000000

He wasn't sure, so he recoded a few more sections of text - but he got even more confused - and decided that he would need to build an evolutionary tree to work out which was the correct version of the manuscript.

A	0101010111010
B	0101010101011
C	1111111100010
D	0101010100100
Philis	0000000000000

To help him to work out the text he was copying from, create a distance matrix for this second site pattern matrix and build a UPGMA and NJ tree.

A

B

C

D

Ph

A B C D Ph

UPGMA

UPGMA works by progressively clustering the most similar species until all the species form a rooted tree.

1. Find the smallest number in the table
2. Form a new internal node between the joined species and calculate the average distance from this node to other unclustered species
3. Update the table with these distances

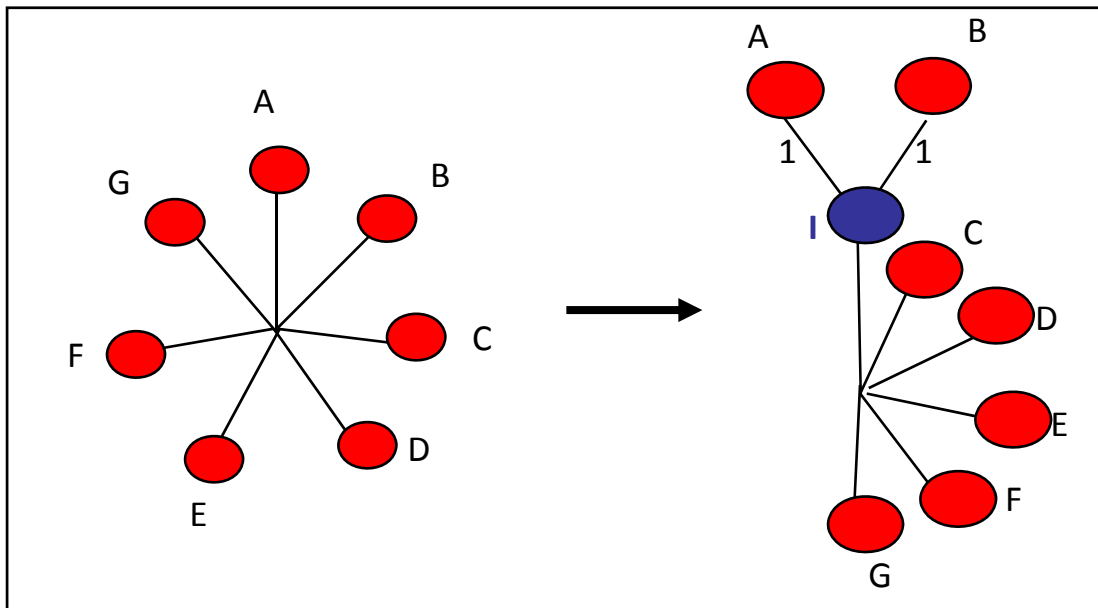
REPEAT until only 2 things are left to be joined.

Step 1 – Find the smallest entry in the distance matrix

d_{ij}	A	B	C	D	E	F
A	-					
B	2	-				
C	4	4	-			
D	4	4	2	-		
E	7	7	7	7	-	
F	5	5	5	5	6	-
G	8	8	8	8	9	5

Step 2 - Cluster taxa A and B, form a new internal node I

Calculate the lengths of the new edges $d(A,I)=d(B,I)=1/2 d(A,B)=1$



Step 3 – Update the distance matrix

$$d(C,I) = \frac{1}{2}(d(A,C) + d(B,C))$$

$$= 4$$

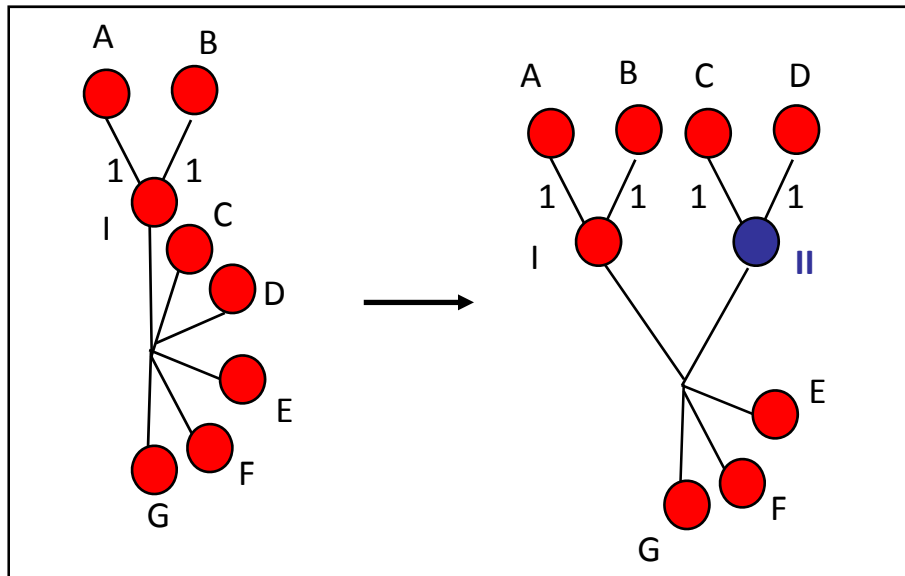
etc...

Step 1 – Find the smallest number in the distance matrix

d_{ij}	(A+B)	C	D	E	F
(A+B)	-				
C	4	-			
D	4	2	-		
E	7	7	7	-	
F	5	5	5	6	-
G	8	8	8	9	5

Step 2 - Cluster taxa C and D, form a new internal node II

Calculate the lengths of the new branches $d(C,II)=d(D,II)=1/2 d(C,D)=1$



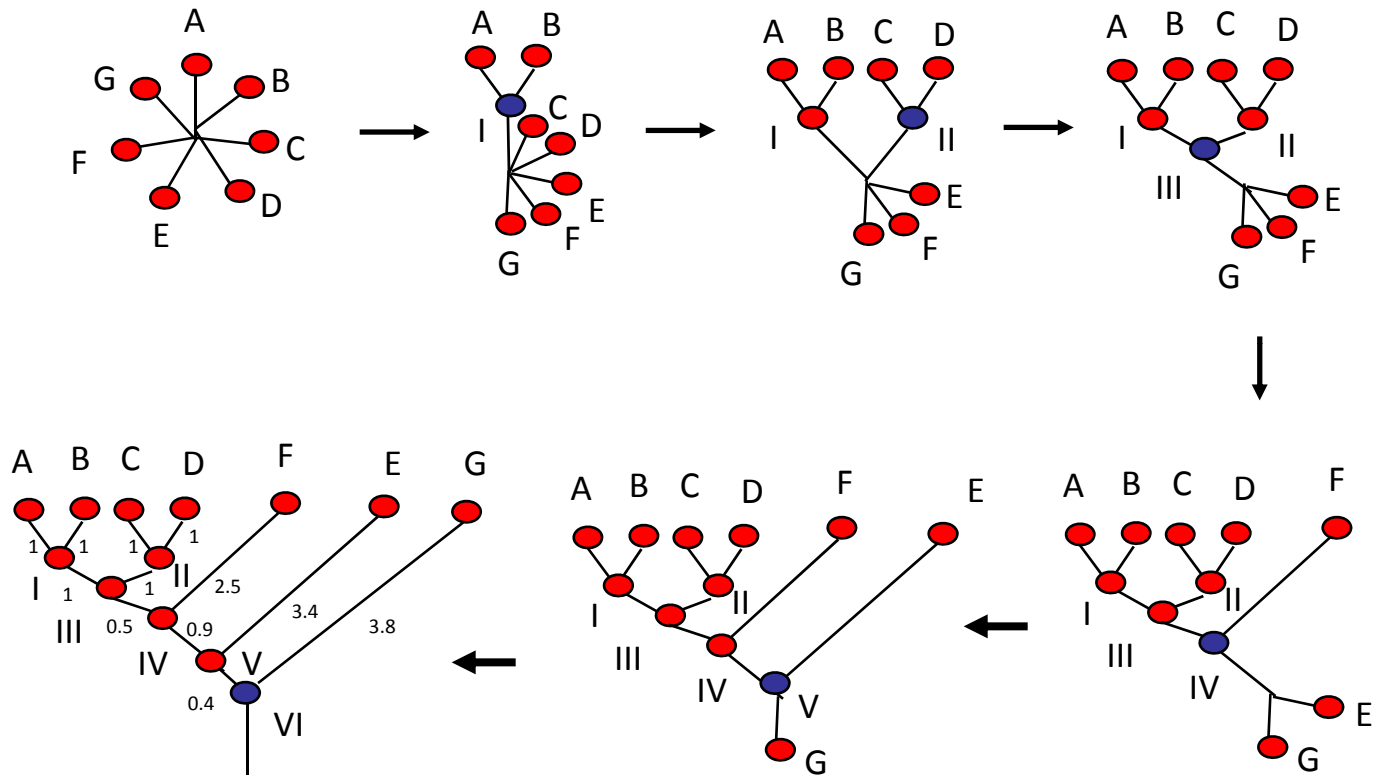
Step 3 – Update the distance matrix

$$d(I,II) = \frac{1}{4}(d(A,C) + d(A,D) + d(B,C) + d(B,D)) = 4$$

$$d(E,II) = \frac{1}{2}(d(E,C) + d(E,D)) = 7$$

etc...

And so on...



Neighbor-joining (NJ)

NJ works by progressively clustering taxa until all the taxa form an unrooted tree.

1. Rather than using the distance matrix directly to determine which taxa should be clustered at each stage, NJ uses the S matrix where

$$S(i,j) = (N-2)d(i,j) - R(i) - R(j)$$

N is the number of taxa.

R(i) is the sum of the ith row in the distance matrix.

R(j) is the sum of the jth row in the distance matrix.

2. Find the smallest number in the S matrix $S(x,y)$.

Neighbor-joining (NJ) cont.

3. Form a new internal node (z) that is a parent to x and y and calculate the edge lengths from z to x and z to y .

$$d(x,z) = 1/(2(N-2))[(N-2)d(x,y) + R(x) - R(y)]$$

$$d(y,z) = d(x,y) - d(x,z)$$

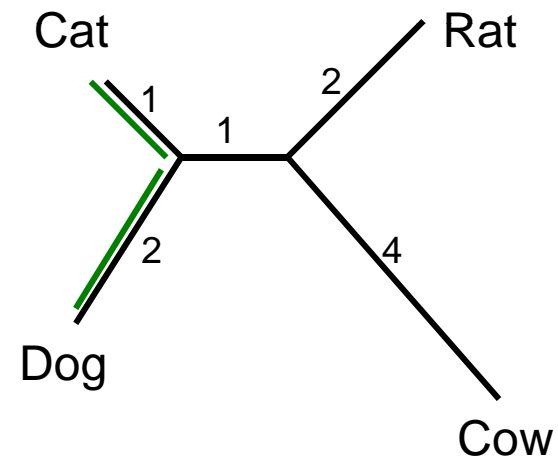
4. Update the distance matrix

$$d(w,z) = \frac{1}{2} (d(x,w) + d(y,w) - d(x,y))$$

REPEAT until only two things are left to be joined.

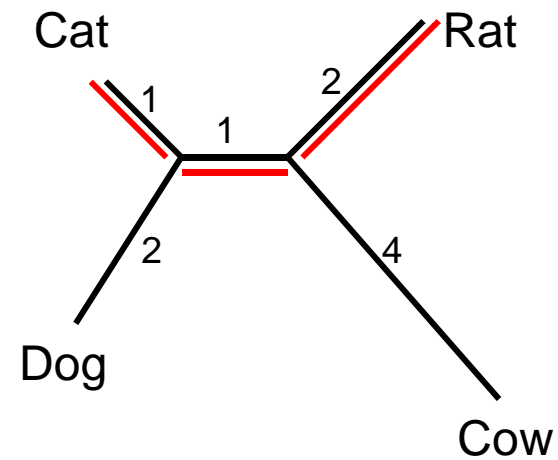
Perfectly “tree-like” distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



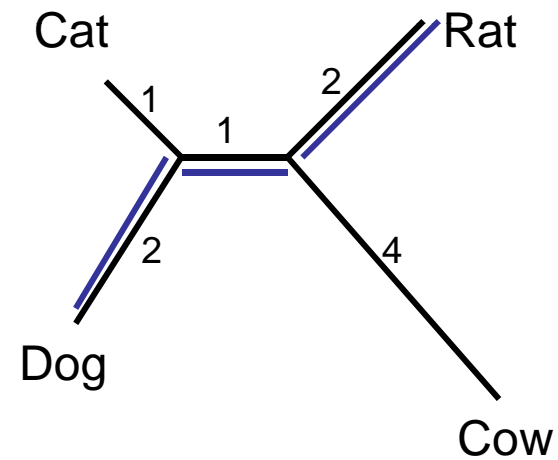
Perfectly “tree-like” distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



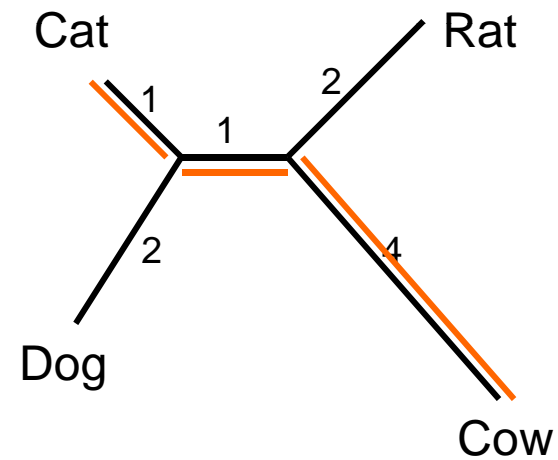
Perfectly “tree-like” distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



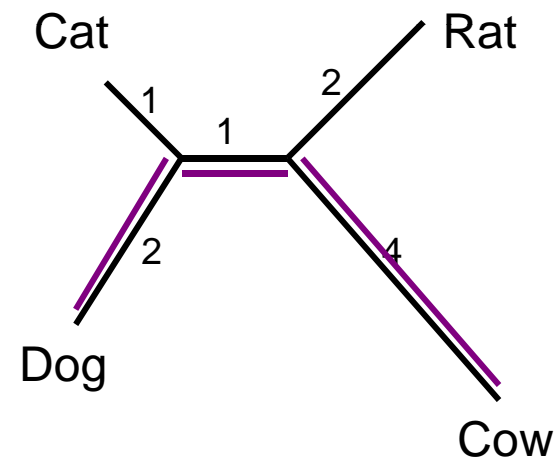
Perfectly “tree-like” distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



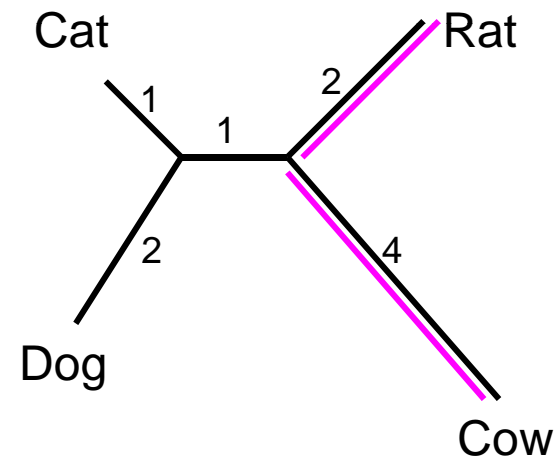
Perfectly “tree-like” distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



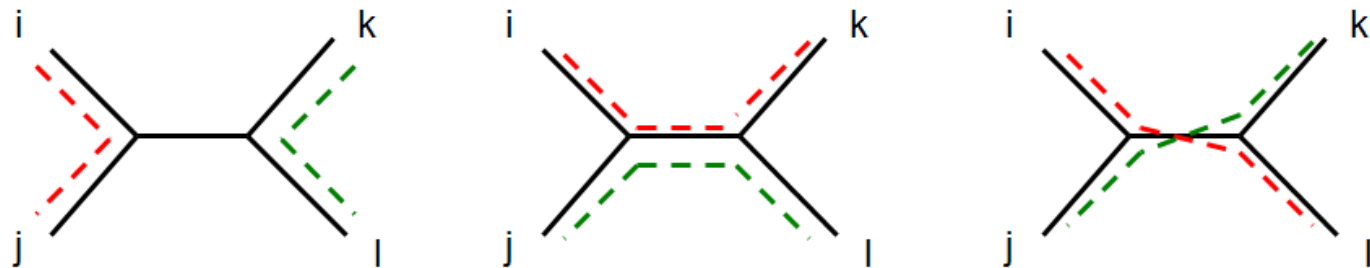
Perfectly “tree-like” distances

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



The 4-Point Condition

- Distances that fit exactly on a tree can be characterised by this condition for any quartet i, j, k, l of taxa sampled from the tree

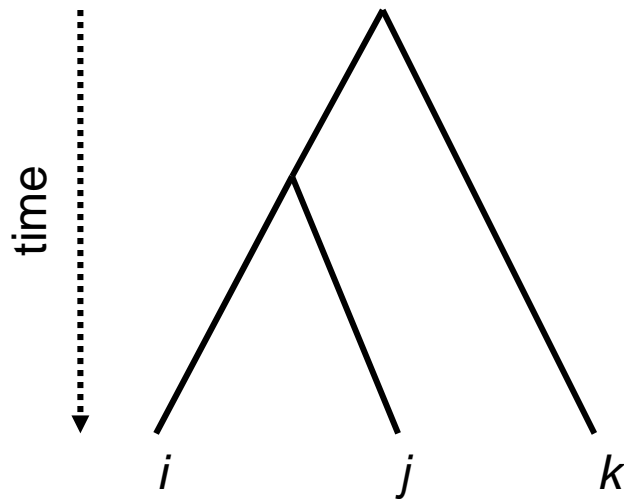


$$d(i,j)+d(k,l) < d(i,k)+d(j,l) = d(i,l)+d(j,k)$$

- Distances with this property are called **additive**, because the length of branches along the paths along the tree **add up** to the values in the distance matrix.

Clock-like distances

- An even stricter condition on distances is that they fit on a clock-like tree.
- Distances with this property are called **ultrametric**.



$$d(i,k) = d(j,k) > d(i,j)$$

UPGMA vs NJ

- The order in which taxa are progressively clustered differs between UPGMA and NJ
- NJ makes more precise estimates of branch lengths than does UPGMA
- UPGMA builds a rooted ultrametric tree, Neighbor Joining builds an unrooted tree (and requires that the distances only be additive)