Phylogenetic Tree Construction- Neighbor Joining Method

Manu Madhavan

Lecture 9

Recap

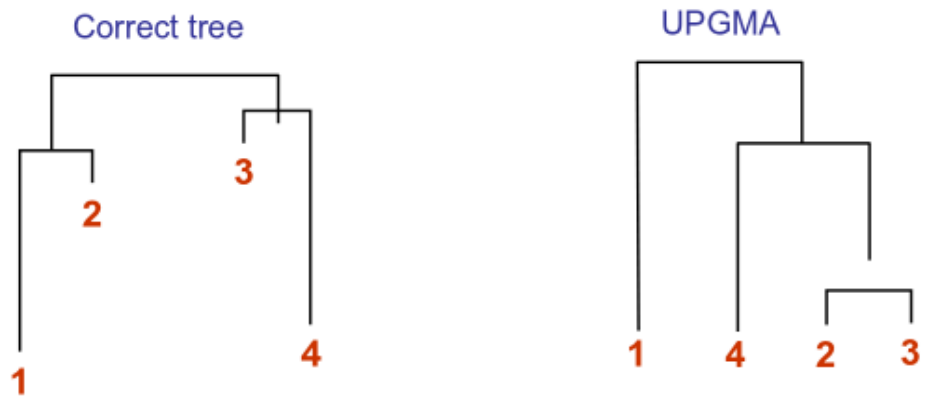
- Phylogenetic Trees - importance
- Methods- Distance based and Character based
- UPGMA

Outline

- Distance Based - Neighbor Joining

Limitations of UPGMA

UPGMA suffers from the molecular clock assumption that the mutation rate over time is constant for all species. However, this is not true as certain species such as rat and mice evolve much faster than others.



Neighbor Joining

- Pairing species in such a way that a tree is created with the smallest possible branch lengths overall.
- On any unrooted tree, pairs of species that are separated from each other by just one internal node are said to be **neighbors**
- Starts with a star-like tree with all species coming off a single central node regardless of their number
- Neighbors are then sequentially found that minimize the total length of the branches on the tree

Neighbor Joining

- Pair two closest clusters according to their distance, $d(C_i, C_j)$ (Criterion#1)
- Would be good to choose the pair that is also far away from the clusters (Criterion#2) measured by $U(C_i) = \sum_j d(C_i, C_j)$
- In NJ method, two clusters merge based on the Q-score, measured as $Q(i, j) = (r - 2)d(C_i, C_j) - U(C_i) - U(C_j)$, where r is the number of clusters

Neighbor Joining

Input: Distance matrix d

1. Compute Q matrix, as $Q(i,j) = (r-2)d(C_i, C_j) - U(C_i) - U(C_j)$
2. Join the closest pairs from Q matrix and add a new node u
3. For newly joined Neighbors (say a and b), compute the branch length as follows:
   $\delta(a,u) = \frac{1}{2}d(a,b) + \frac{1}{2(n-2)}(U(a) - U(b))$
   $\delta(b,u) = d(a,b) - \delta(a,u)$
4. Compute the distance between other nodes from the new node as follows:
   $\delta(u,k) = \frac{1}{2}(d(a,k) + d(b,k) - d(a,b))$
5. Repeat from step-1, replacing the pair of joined neighbors with the new node and using the distances calculated in the previous step.

Example

	a	b	c	d	e
a	0	5	9	9	8
b	5	0	10	10	9
c	9	10	0	8	7
d	9	10	8	0	3
e	8	9	7	3	0

- $U(C_i) = \sum_j d(C_i, C_j)$
- $Q(i,j) = (r-2)d(C_i, C_j) - U(C_i) - U(C_j)$
- $\delta(a,u) = \frac{1}{2}d(a,b) + \frac{1}{2(n-2)}(U(a) - U(b))$
- $\delta(b,u) = d(a,b) - \delta(a,u)$
- $\delta(u,k) = \frac{1}{2}(d(a,k) + d(b,k) - d(a,b))$

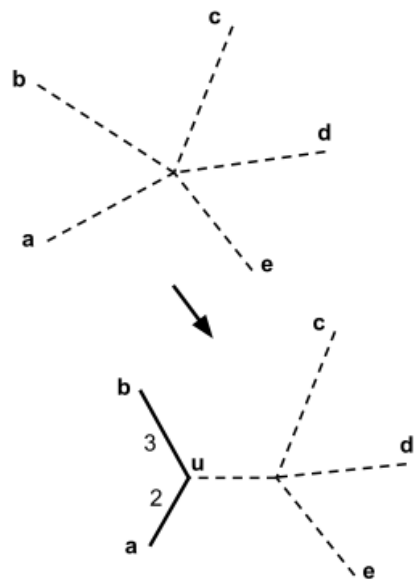
Example: Step-1

	a	b	c	d	e
a		−50	−38	−34	−34
b	−50		−38	−34	−34
c	−38	−38		−40	−40
d	−34	−34	−40		−48
e	−34	−34	−40	−48	

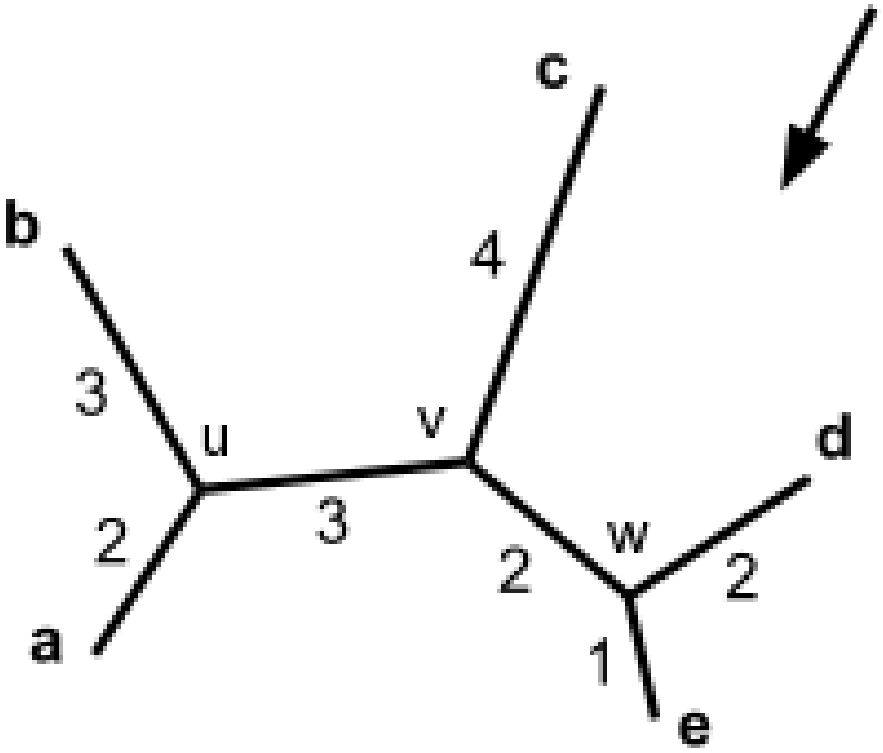
- $U(C_i) = \sum_j d(C_i, C_j)$
- $Q(i,j) = (r-2)d(C_i, C_j) - U(C_i) - U(C_j)$
- $\delta(a,u) = \frac{1}{2}d(a,b) + \frac{1}{2(n-2)}(U(a) - U(b))$
- $\delta(b,u) = d(a,b) - \delta(a,u)$
- $\delta(u,k) = \frac{1}{2}(d(a,k) + d(b,k) - d(a,b))$

New Distance matrix

	u	c	d	e
u	0	7	7	6
c	7	0	8	7
d	7	8	0	3
e	6	7	3	0



Example-Final stage



Character Based Methods

- In character-based methods, the goal is to first create a valid algorithm for scoring the probability that a given tree would produce th observed sequences at its leaves, then to search through the space of possible trees for a tree that maximizes that probability.
- The concept of parsimony is at the very heart of all character-based methods of phylogenetic reconstruction
- In a biological sense, it is used to describe the process of attaching preference to one evolutionary pathway over another on the basis of which pathway requires the invocation of the smallest number of mutational events
- Maximum parsimony algorithm

Next

- Character based methods- Maximum Parsimony
- **Quiz-1**- Introduction, Sequence Alignment