

Bioinformatics I & II

Wayne Delpont & Sergei Kosakovsky Pond

Practical 3: Neighbor-Joining Algorithm

Consider the following distance matrix estimated from an alignment of sequences A-F.

	A	B	C	D	E	F
A						
B	5					
C	4	7				
D	7	10	7			
E	6	9	6	5		
F	8	11	8	9	8	

Draw a phylogeny using the Neighbor-Joining algorithm.

Step 1

Compute the net divergence r_i for each sequence i given by

$$r_i = \frac{1}{L-2} \sum_{k \in L} d(i, k) \quad (1)$$

where L is the number of leaves, and $d(i, k)$ the observed distance between i and k from the distance matrix.

Step 2

Create a corrected distance matrix, D_c , using the r_i estimates from Step 1, as

$$D_c(i, j) = d(i, j) - (r_i + r_j) \quad (2)$$

Step 3

Define a new node, U , that groups two sequences (A-F) which are maximally similar to each other, and maximally different from the remaining sequences (i.e. choose the minimum value from the modified distance matrix, D_c).

Step 4

Compute the branch lengths between the new node, U , and each of the children of that parent (i.e. the two sequences you joined in Step 3 above). For sequence i the branch length (B_{iU}) is

$$B_{iU} = \frac{1}{2}(d(i, j) + r_i - r_j) \quad (3)$$

Step 5

Compute the new distances from node U , joining leaves i and j , to each of the remaining leaves, k , as

$$d(U, i) = \frac{1}{2}(d(i, k) + d(j, k) - d(i, j)) \quad (4)$$

Repeat steps 1 through 5 until the last leaf has been added to the tree.