

# CSC 334: TOPICS IN COMPUTATIONAL BIOLOGY

---

“Algorithms for Genomic Data”

Fall 2015

Smith College

Instructor: Prof. Sara Sheehan

# Outline: 10/5

- Neighbor-Joining (NJ) paper

# NJ: Figure 1

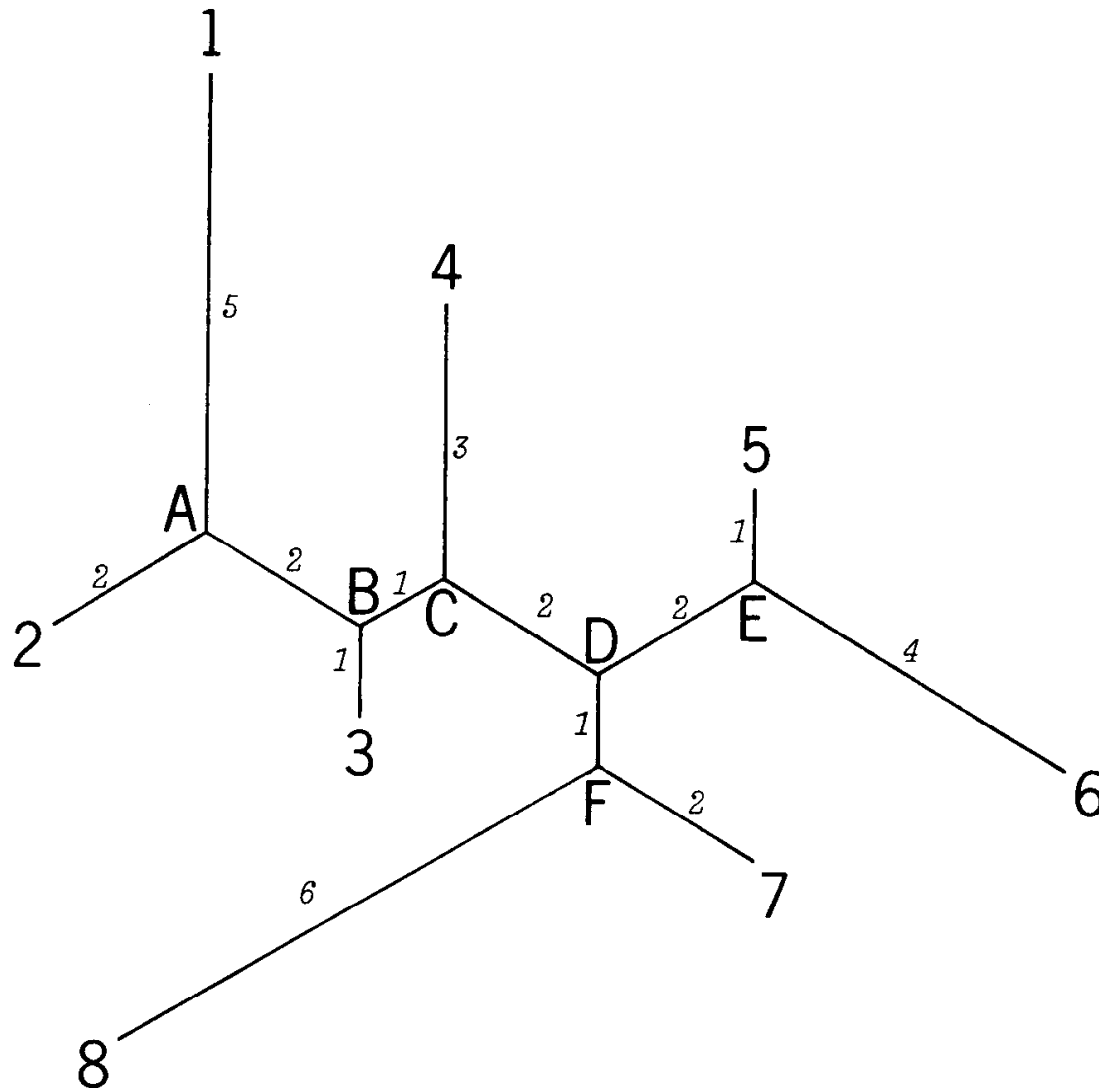


FIG. 1.—An unrooted tree of eight OTUs, 1–8. A–F are interior nodes, and italic numbers are branch lengths.

# NJ: Table 1

**Table 1**  
**Distance Matrix for the Tree in Figure 1**

OTU	OTU						
	1	2	3	4	5	6	7
2 ..	7						
3 ..	8	5					
4 ..	11	8	5				
5 ..	13	10	7	8			
6 ..	16	13	10	11	5		
7 ..	13	10	7	8	6	9	
8 ..	17	14	11	12	10	13	8

## NJ: Figure 2

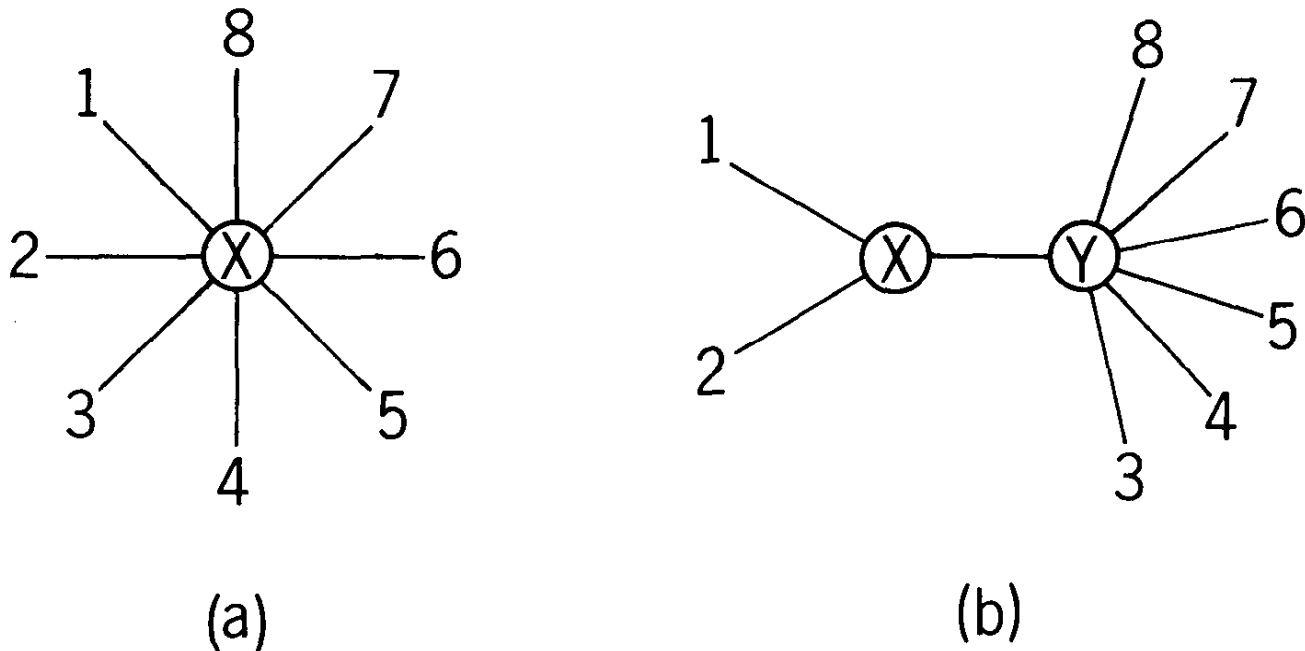


FIG. 2.—(a), A starlike tree with no hierarchical structure; and (b), a tree in which OTUs 1 and 2 are clustered.

# NJ: Table 2

**Table 2**

**$S_{ij}$  Matrices for Two Cycles of the NJ Method for the Data in Table 1**

A. Cycle 1: Neighbors = [1, 2]							
OTU	OTU						
	1	2	3	4	5	6	7
2 ..	36.67						
3 ..	38.33	38.33					
4 ..	39.00	39.00	38.67				
5 ..	40.33	40.33	40.00	39.67			
6 ..	40.33	40.33	40.00	39.67	37.00		
7 ..	40.17	40.17	39.83	39.50	38.83	38.83	
8 ..	40.17	40.17	39.83	39.50	38.83	38.83	37.67

# NJ: Table 2

B. Cycle 2: Neighbors = [5, 6]

OTU	OTU					
	1-2	3	4	5	6	7
3 ..	31.50					
4 ..	32.30	32.30				
5 ..	33.90	33.90	33.70			
6 ..	33.90	33.90	33.70	31.30		
7 ..	33.70	33.70	33.50	33.10	33.10	
8 ..	33.70	33.70	33.50	33.10	33.10	31.90

# NJ: Figure 3

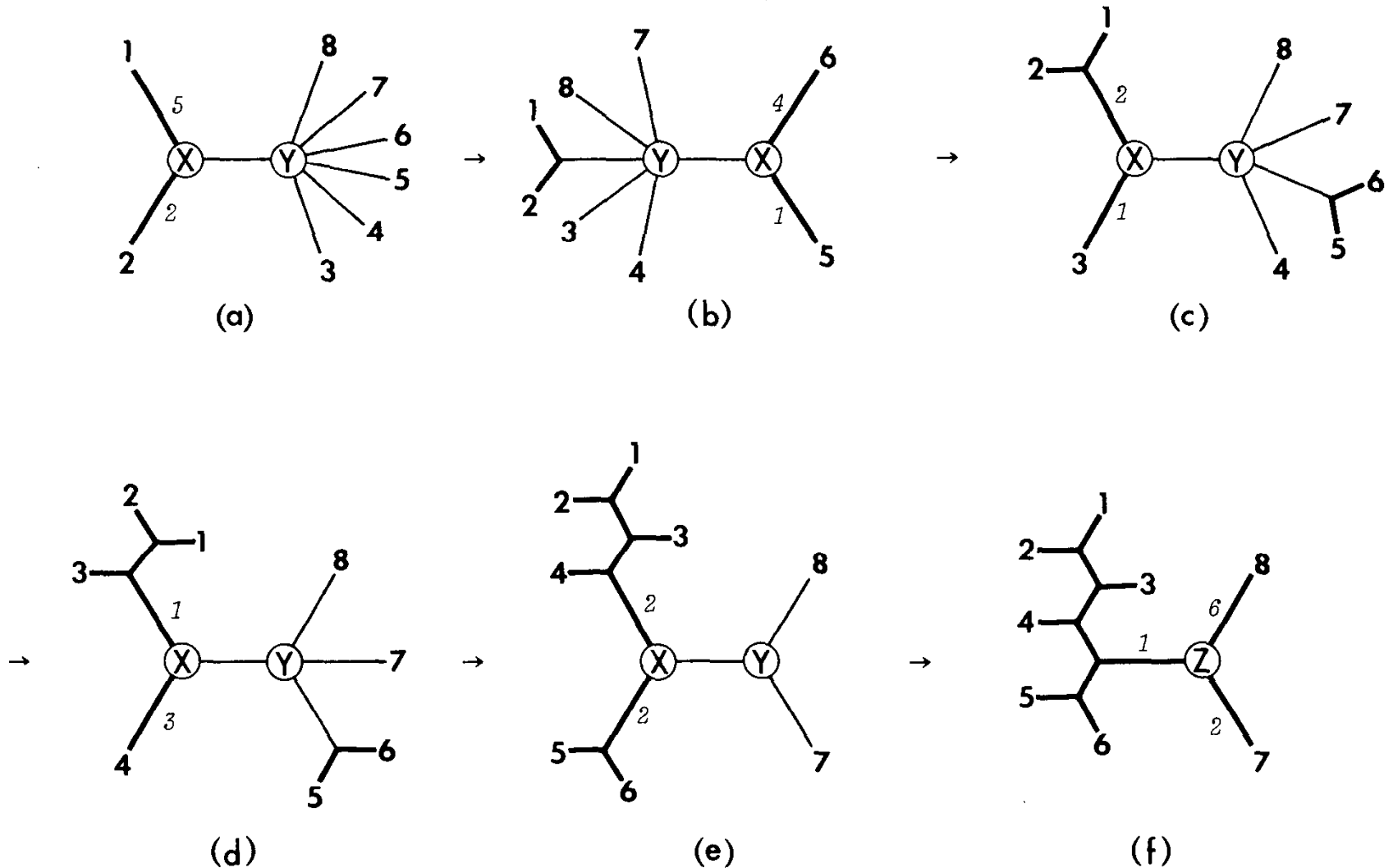


FIG. 3.—Application of the neighbor-joining method to the distance matrix of table 1. Italic numbers are branch lengths, and branches with thicker lines indicate that their lengths have been determined.



$$S_{12} = L_{XY} + (L_{1X} + L_{2X}) + \sum_{i=3}^N L_{iY}$$

$$= \frac{1}{2(N-2)} \sum_{k=3}^N (D_{1k} + D_{2k}) + \frac{1}{2} D_{12} + \frac{1}{N-2} \sum_{3 \leq i < j} D_{ij}.$$

**Table 1**  
**Distance Matrix for the Tree in Figure 1**

		OTU						
OTU		1	2	3	4	5	6	7
2	..	7						
3	..	8	5					
4	..	11	8	5				
5	..	13	10	7	8			
6	..	16	13	10	11	5		
7	..	13	10	7	8	6	9	
8	..	17	14	11	12	10	13	8

# NJ: Figure 6

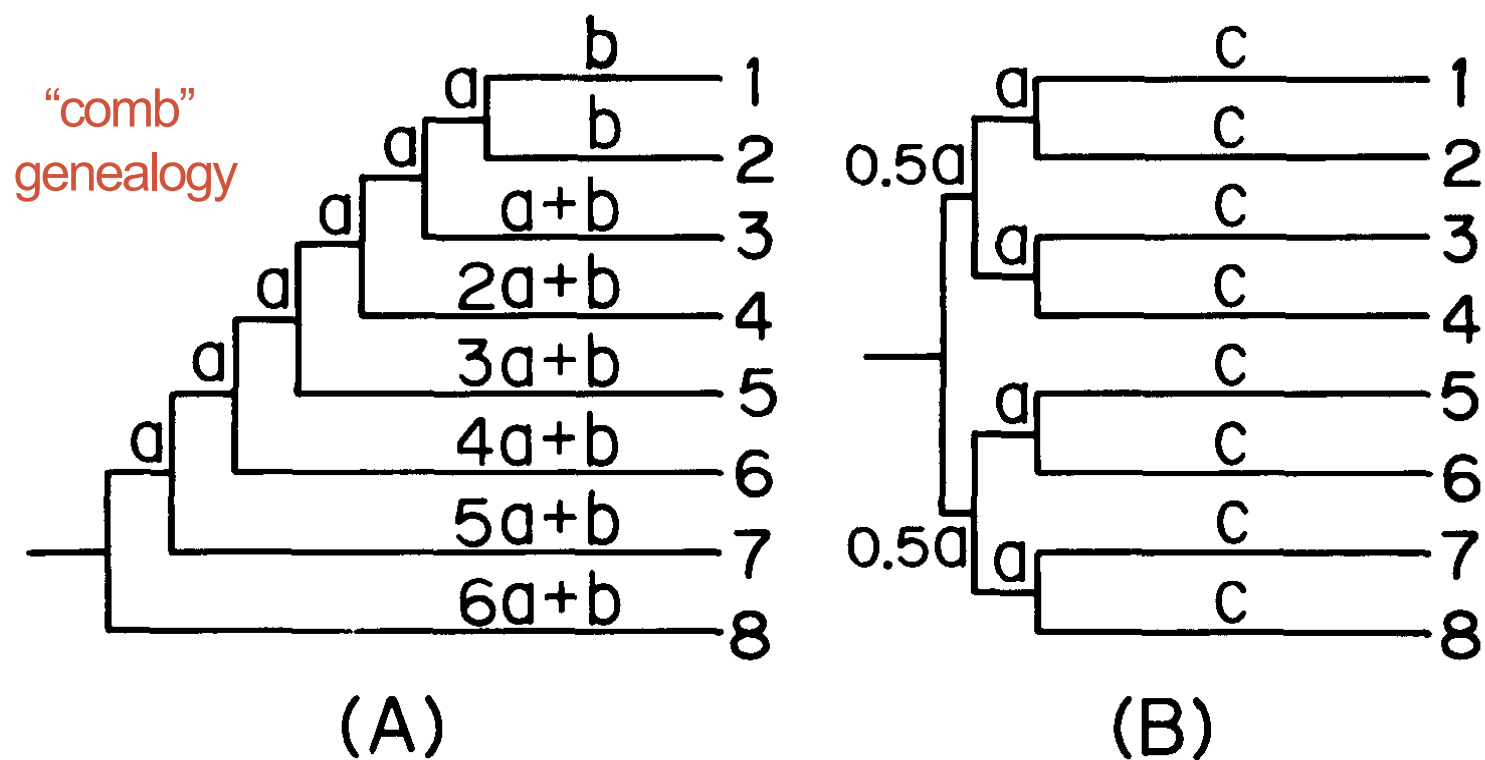


FIG. 6.—Model trees (A) and (B) under the assumption of constant rate of nucleotide substitution

Table 3

$P_c$  and  $d_T$  (in parentheses) for Six Tree-making Methods for the Case of  $a = 0.01$ ,  $b = 0.04$ , and  $c = 0.07$

METHOD	MODEL TREE A <sup>a</sup>			MODEL TREE B <sup>a</sup>		
	300	600	900	300	600	900
UPGMA:						
$p^b$ .....	14 (3.18)	36 (1.72)	58 (0.98)	14 (4.54)	36 (2.74)	51 (1.68)
$d^c$ .....	15 (3.18)	34 (1.74)	56 (1.04)	13 (4.56)	35 (2.70)	52 (1.60)
MF:						
$p$ .....	39 (1.76)	73 (0.58)	95 (0.10)	24 (2.86)	51 (1.30)	67 (0.76)
$d$ .....	38 (1.92)	72 (0.62)	95 (0.10)	19 (2.94)	48 (1.42)	64 (0.86)
DW:						
$p$ .....	42 (1.70)	75 (0.54)	96 (0.08)	26 (2.36)	55 (1.12)	79 (0.48)
$d$ .....	37 (1.74)	74 (0.58)	95 (0.10)	28 (2.36)	58 (1.06)	79 (0.46)
LI:						
$p$ .....	41 (1.58)	71 (0.70)	94 (0.12)	40 (2.04)	70 (0.78)	90 (0.22)
$d$ .....	36 (1.84)	66 (0.82)	89 (0.24)	39 (2.10)	70 (0.78)	90 (0.26)
ST:						
$p$ .....	48 (1.26)	75 (0.54)	97 (0.06)	45 (1.66)	75 (0.62)	91 (0.22)
$d$ .....	44 (1.48)	70 (0.62)	96 (0.08)	43 (1.62)	74 (0.64)	91 (0.22)
NJ:						
$p$ .....	48 (1.36)	76 (0.54)	97 (0.06)	46 (1.64)	76 (0.60)	91 (0.20)
$d$ .....	41 (1.60)	70 (0.62)	96 (0.08)	45 (1.62)	75 (0.60)	91 (0.20)

<sup>a</sup> As shown in fig. 6.

<sup>b</sup> Trees reconstructed from data on the proportion of different nucleotides between the sequences compared.

<sup>c</sup> Trees reconstructed from the Jukes-Cantor distance.

Table 4

$P_c$  and  $d_T$  (in parentheses) for Six Tree-making Methods for the Case of  $a = 0.02$ ,  $b = 0.13$ , and  $c = 0.19$

METHOD	MODEL TREE A <sup>a</sup>			MODEL TREE B <sup>a</sup>		
	300	600	900	300	600	900
UPGMA:						
$p$ .....	15 (3.24)	50 (1.32)	62 (0.82)	11 (4.62)	28 (2.94)	54 (1.48)
$d$ .....	15 (3.28)	49 (1.34)	61 (0.84)	13 (4.50)	30 (2.90)	57 (1.44)
MF:						
$p$ .....	34 (2.38)	65 (0.82)	79 (0.44)	10 (4.00)	25 (2.22)	43 (1.48)
$d$ .....	30 (2.70)	62 (1.02)	76 (0.54)	9 (4.12)	22 (2.28)	43 (1.48)
DW:						
$p$ .....	27 (2.40)	66 (0.96)	77 (0.54)	17 (3.54)	39 (1.92)	54 (1.10)
$d$ .....	27 (2.52)	62 (1.02)	70 (0.70)	18 (3.54)	36 (1.98)	53 (1.16)
LI:						
$p$ .....	23 (2.60)	44 (1.34)	67 (0.80)	25 (3.54)	50 (1.52)	81 (0.52)
$d$ .....	20 (2.82)	33 (1.78)	55 (1.12)	20 (3.70)	49 (1.54)	81 (0.50)
ST:						
$p$ .....	35 (2.06)	67 (0.74)	82 (0.38)	34 (2.40)	60 (1.08)	82 (0.38)
$d$ .....	26 (2.42)	61 (0.96)	78 (0.48)	31 (2.50)	58 (1.16)	83 (0.36)
NJ:						
$p$ .....	36 (2.14)	64 (0.88)	83 (0.34)	34 (2.32)	63 (0.96)	82 (0.36)
$d$ .....	26 (2.38)	58 (1.08)	78 (0.48)	33 (2.56)	61 (1.04)	83 (0.34)

NOTE.—Notations are as in table 3.

<sup>a</sup> As shown in fig. 6.

Table 5

$P_c$  and  $d_T$  (in parentheses) for Six Tree-making Methods for the Case of  $a = 0.03$ ,  $b = 0.34$ , and  $c = 0.42$

METHOD	MODEL TREE A <sup>a</sup>			MODEL TREE B <sup>a</sup>		
	500	1,000	2,000	500	1,000	2,000
UPGMA:						
$p$ .....	9 (3.78)	27 (2.10)	62 (0.86)	10 (5.20)	18 (3.76)	54 (1.32)
$d$ .....	9 (3.78)	27 (2.10)	62 (0.88)	11 (5.30)	18 (3.74)	55 (1.26)
MF:						
$p$ .....	15 (4.02)	41 (1.82)	62 (0.92)	3 (5.68)	17 (3.64)	28 (2.40)
$d$ .....	13 (4.42)	34 (2.14)	55 (1.14)	3 (5.72)	13 (3.80)	26 (2.48)
DW:						
$p$ .....	16 (3.78)	46 (1.54)	63 (0.82)	4 (5.42)	18 (3.28)	41 (1.72)
$d$ .....	15 (4.22)	40 (1.96)	58 (0.98)	5 (5.50)	18 (3.48)	35 (1.82)
LI:						
$p$ .....	3 (4.26)	37 (2.00)	53 (1.18)	15 (4.48)	28 (2.98)	70 (0.90)
$d$ .....	3 (4.84)	25 (2.60)	39 (1.66)	12 (4.72)	27 (3.06)	66 (1.02)
ST:						
$p$ .....	10 (3.56)	44 (1.62)	68 (0.76)	13 (4.00)	36 (2.34)	74 (0.62)
$d$ .....	6 (4.06)	40 (1.82)	56 (1.04)	10 (4.32)	34 (2.34)	71 (0.72)
NJ:						
$p$ .....	11 (3.70)	44 (1.68)	67 (0.80)	13 (4.46)	34 (2.38)	75 (0.62)
$d$ .....	5 (4.24)	38 (2.00)	57 (1.06)	14 (4.44)	32 (2.42)	73 (0.72)

NOTE.—Notations are as in table 3.

<sup>a</sup> As shown in fig. 6.

## NJ: Figure 7

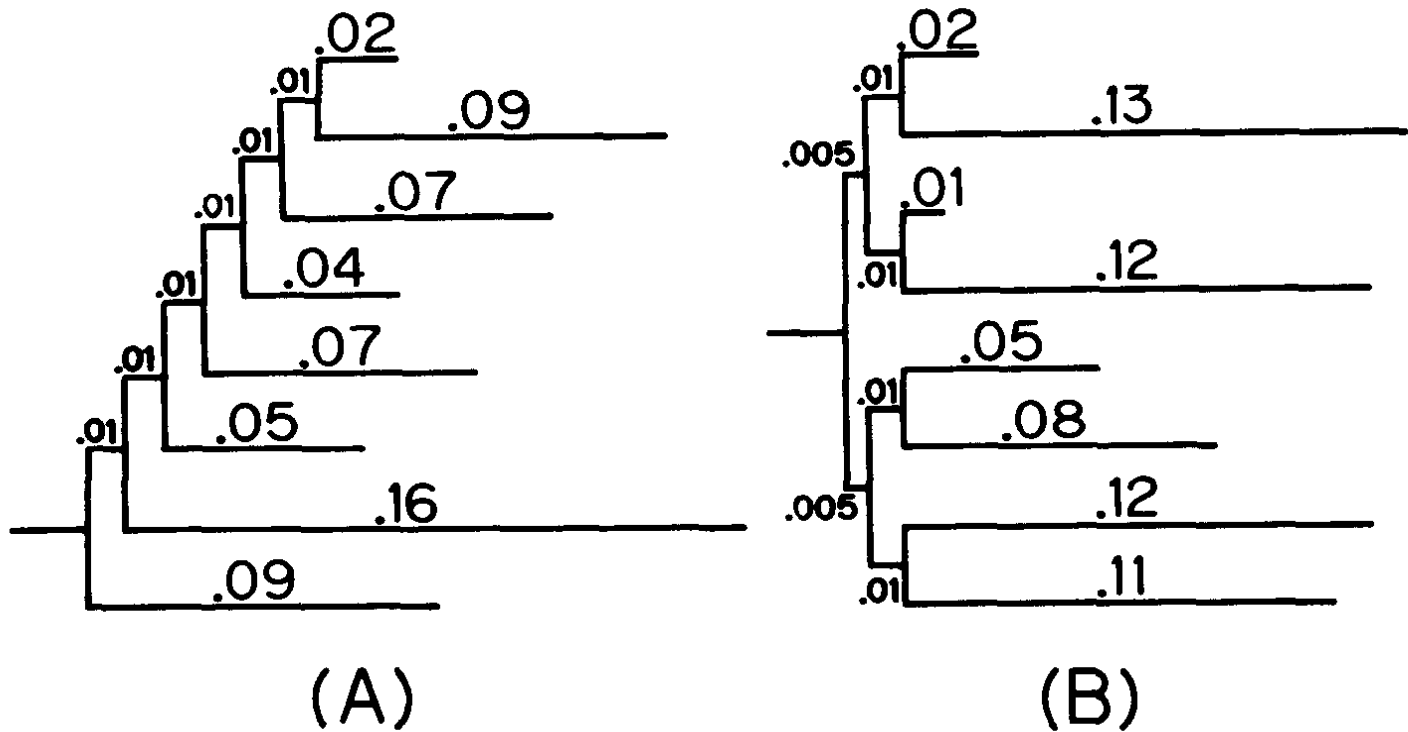


FIG. 7.—Model trees (A) and (B) under the assumption of varying rate of nucleotide substitution

**Table 6**

**$P_c$  and  $d_T$  (in parentheses) for Six Tree-making Methods for the Case of Varying Rate of Nucleotide Substitution**

Method	Model Tree A <sup>a</sup>	Model Tree B <sup>a</sup>
UPGMA: $p$ . . . . .	0 (8.06)	0 (9.74)
MF: $p$ . . . . .	77 (0.50)	57 (1.46)
DW: $p$ . . . . .	69 (0.72)	59 (1.26)
LI: $p$ . . . . .	46 (1.30)	45 (1.68)
ST: $p$ . . . . .	77 (0.50)	69 (0.82)
NJ: $p$ . . . . .	75 (0.56)	72 (0.78)

NOTE.—Notations are as in table 3.

<sup>a</sup> As shown in fig. 7.