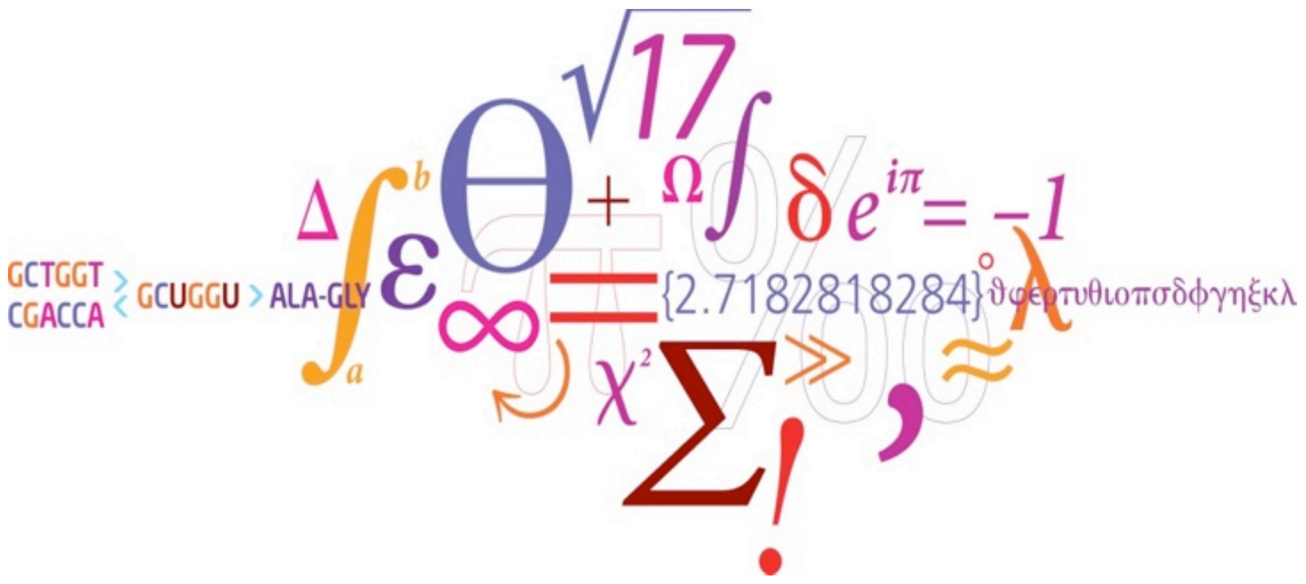
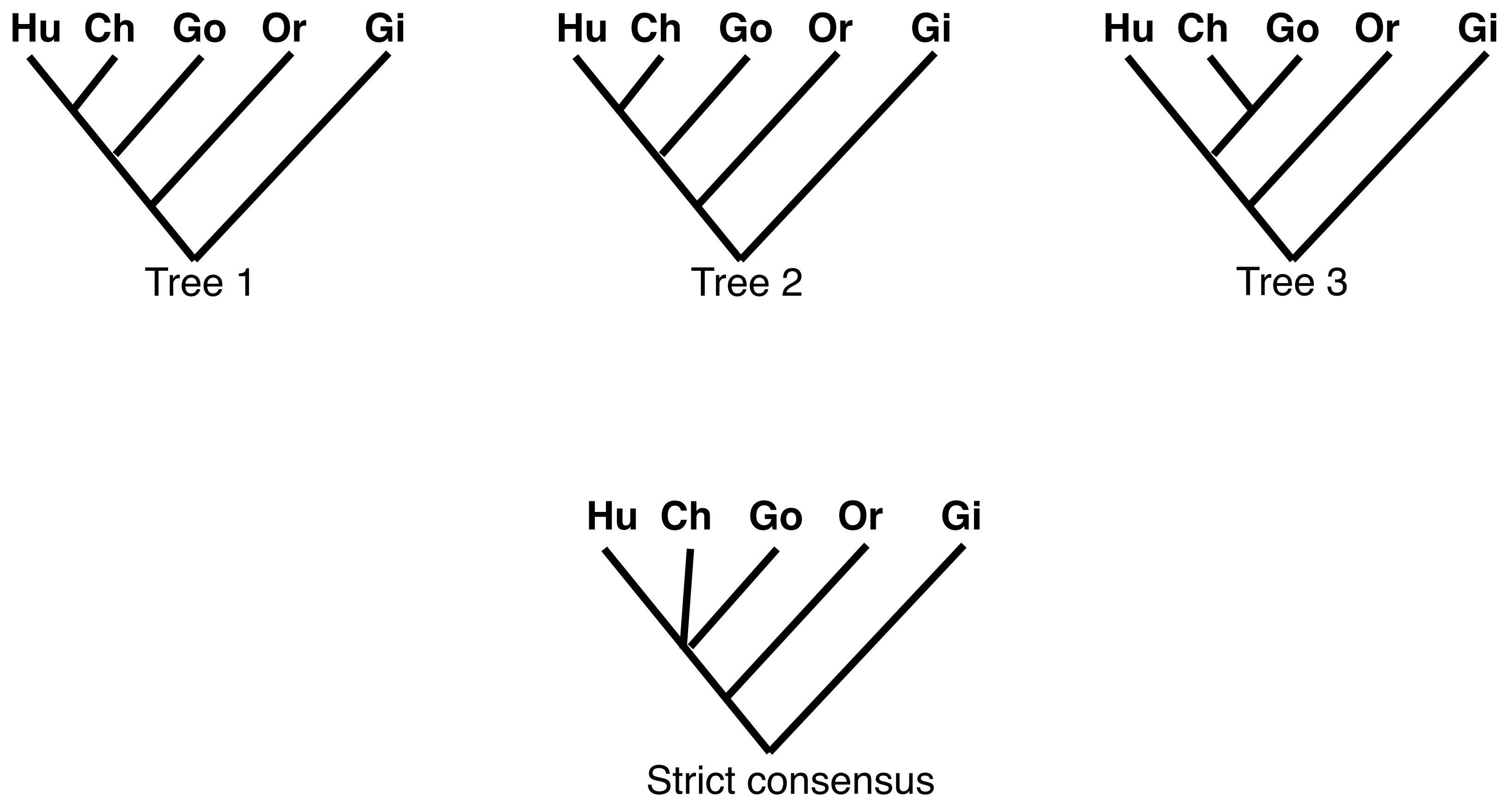
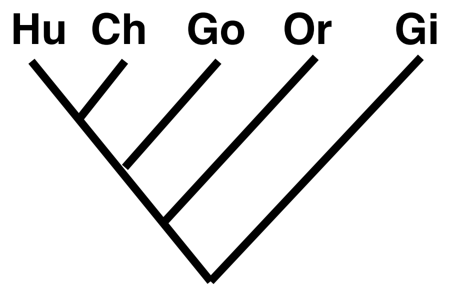
Consensus Trees



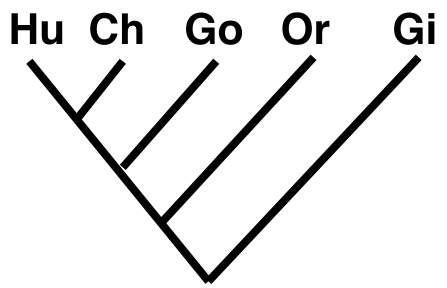
Strict Consensus Tree



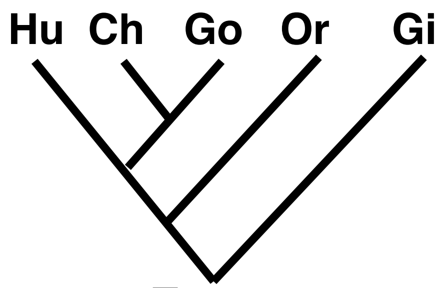
Majority Rule Consensus Tree



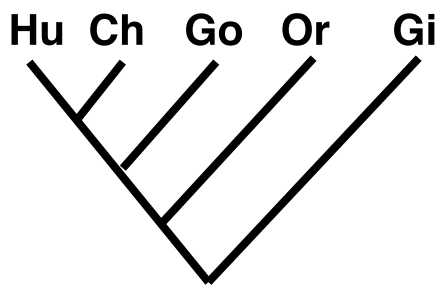
Tree 1



Tree 2

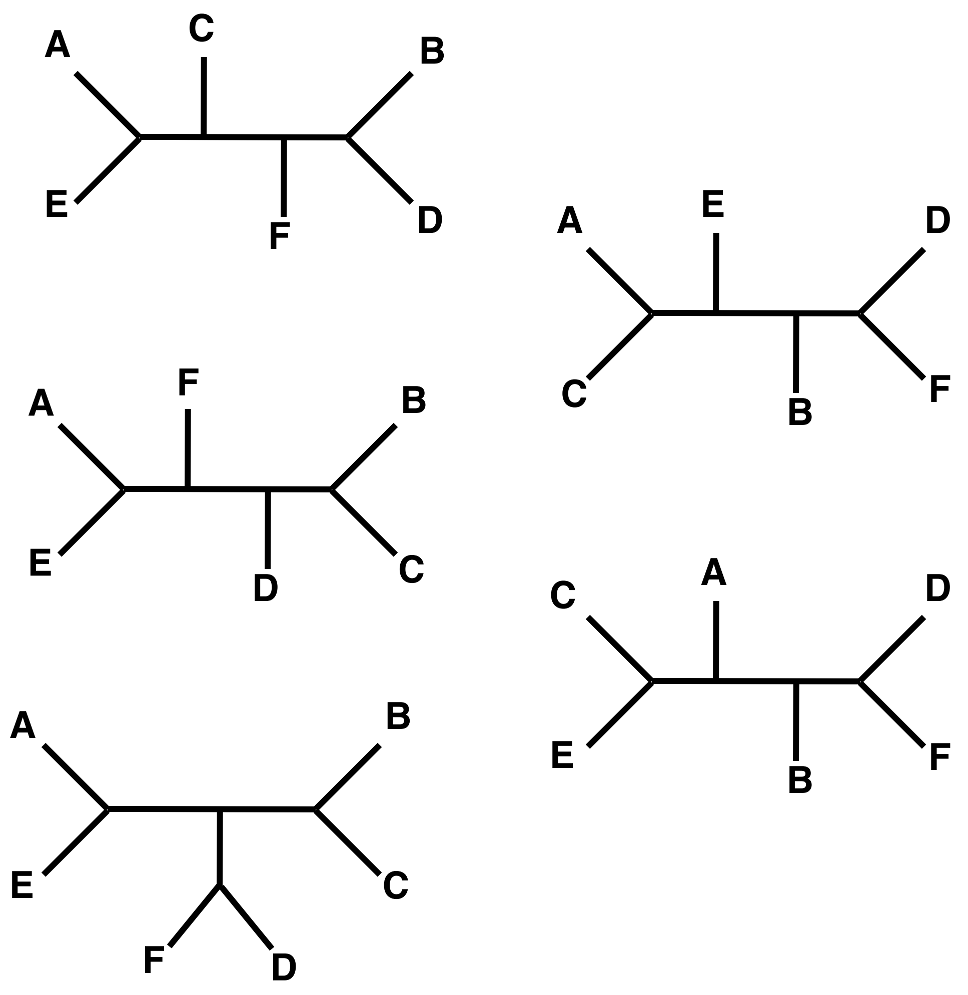


Tree 3

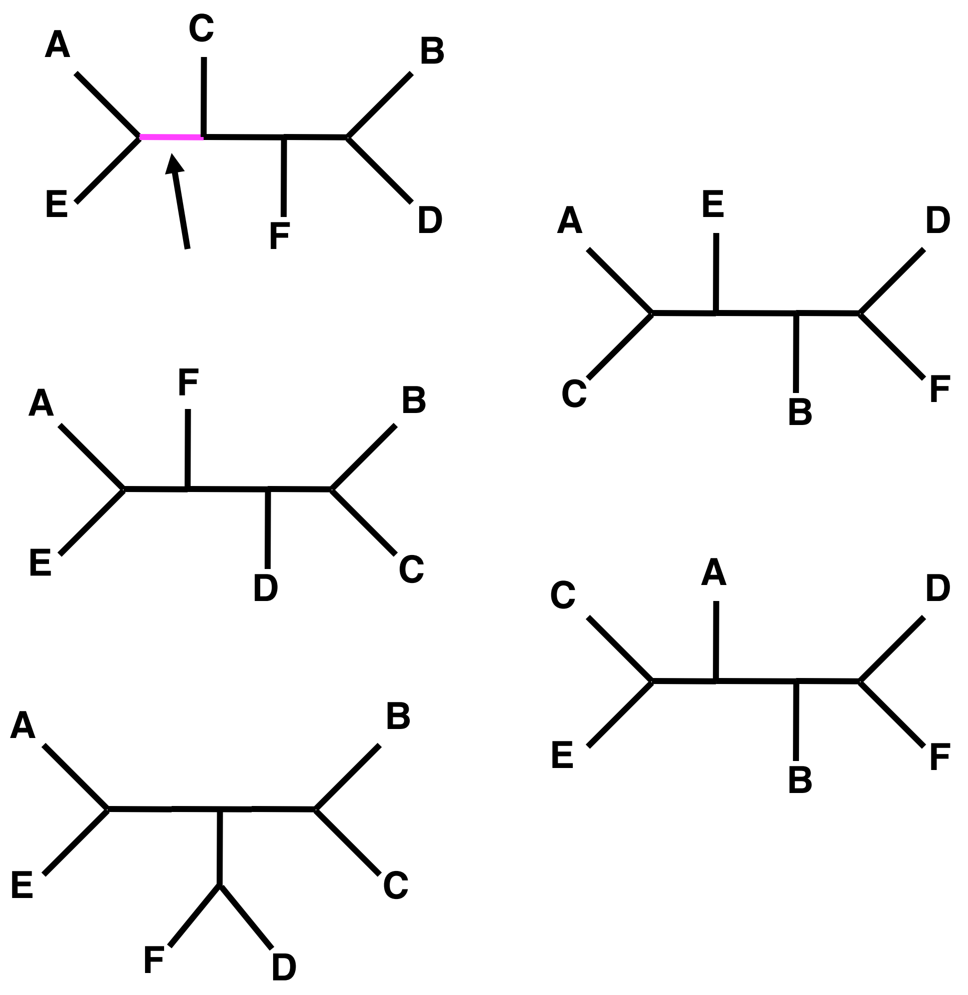


Majority rule consensus

Majority Rule Consensus Tree

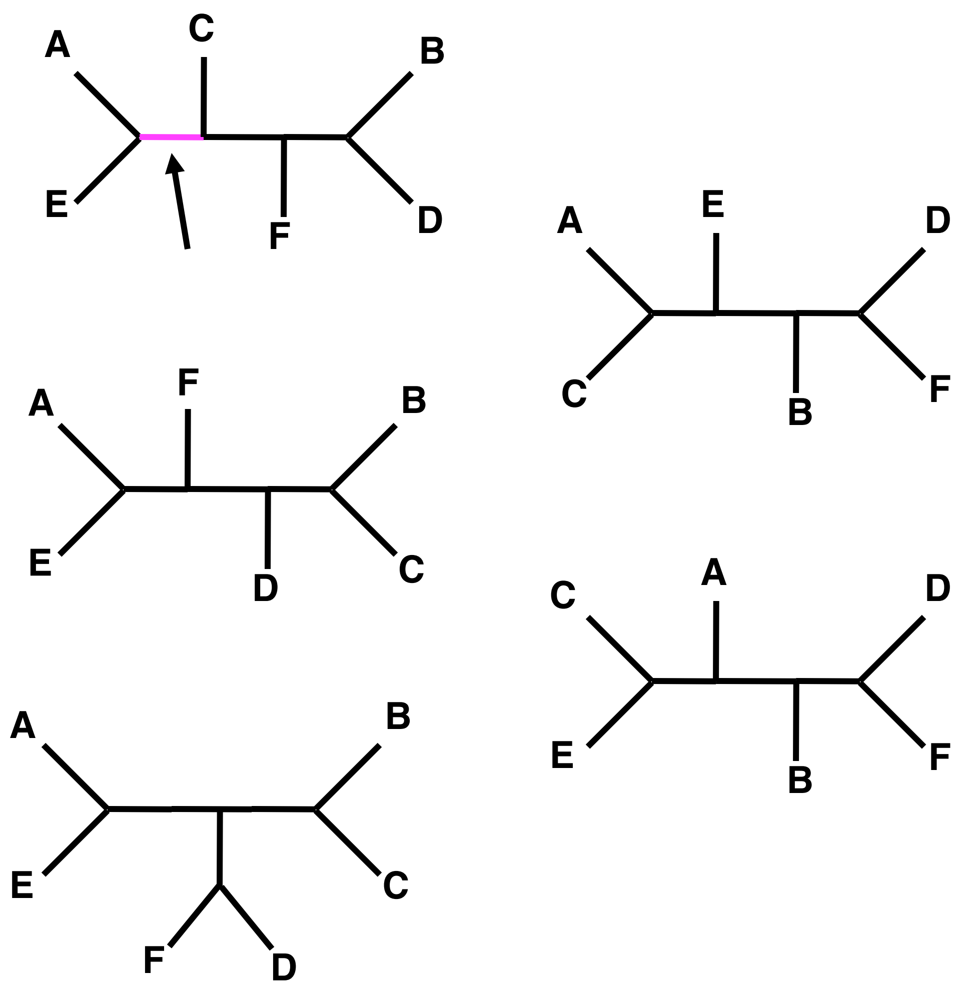


Majority Rule Consensus Tree



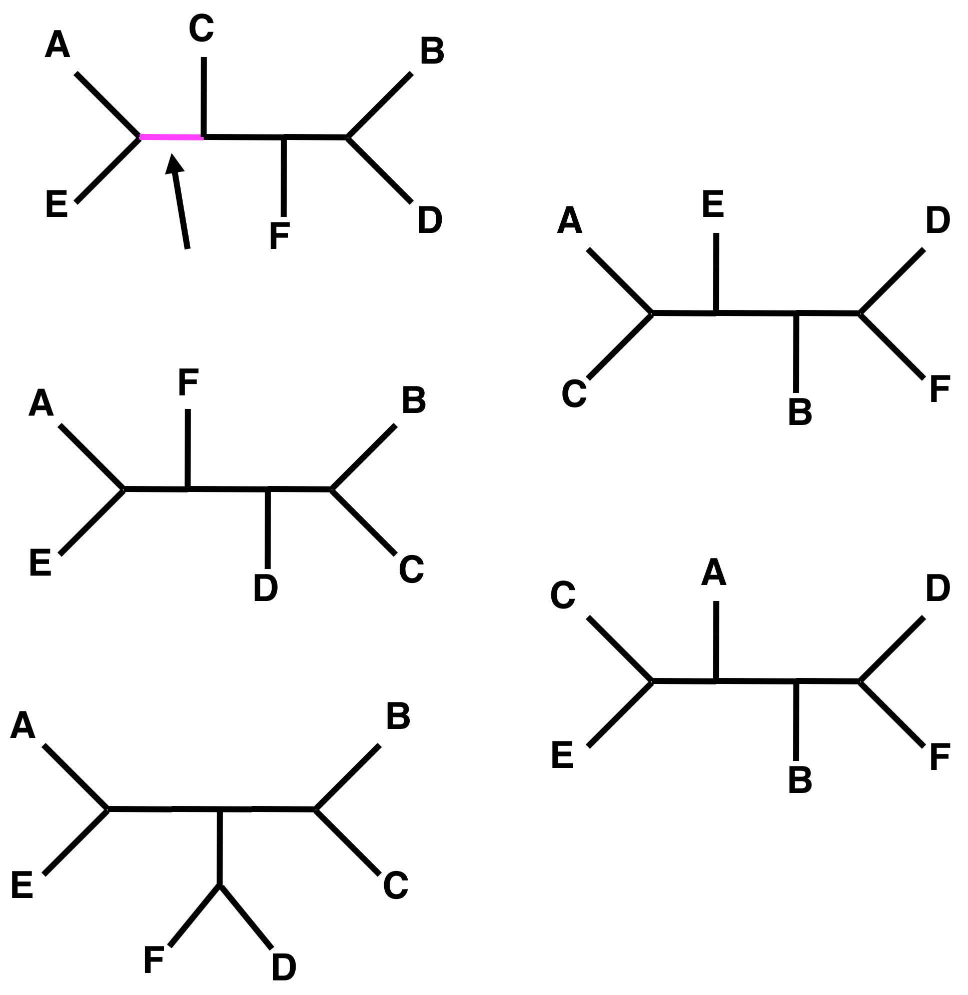
A	B	C	D	E	F	Count

Majority Rule Consensus Tree



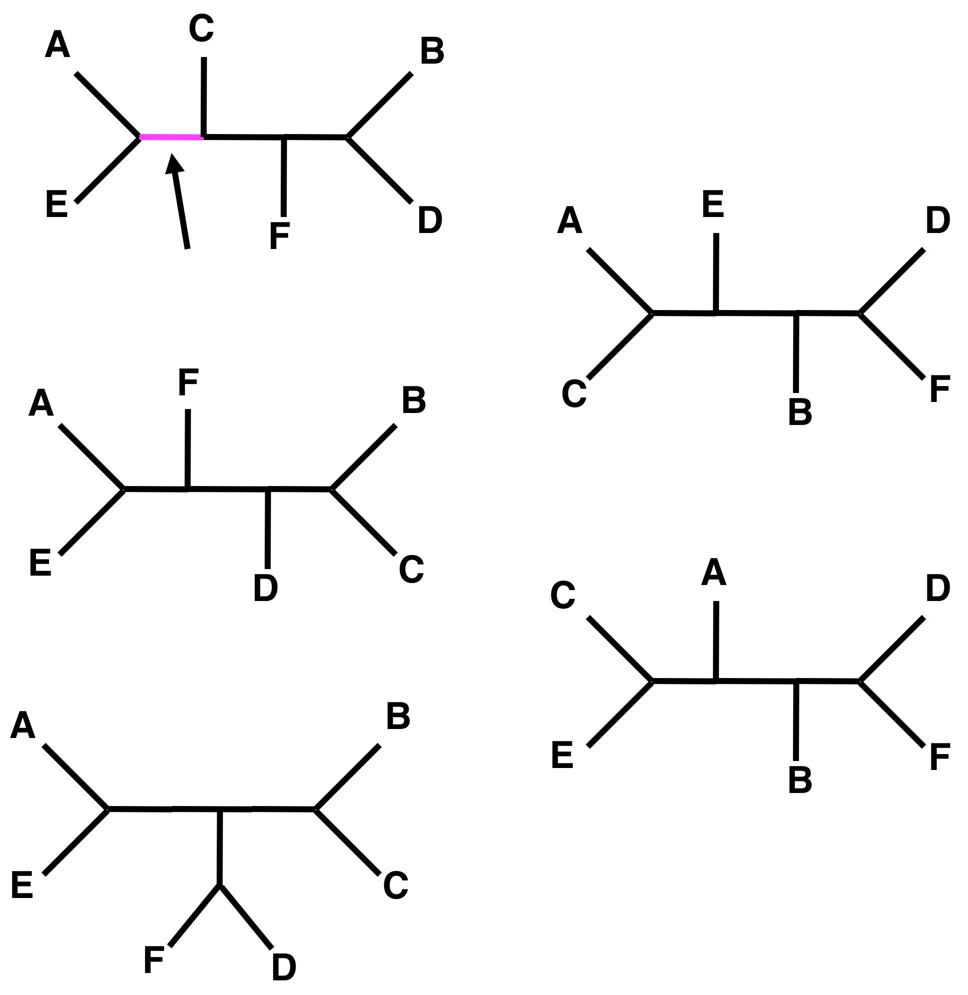
A	B	C	D	E	F	Count

Majority Rule Consensus Tree



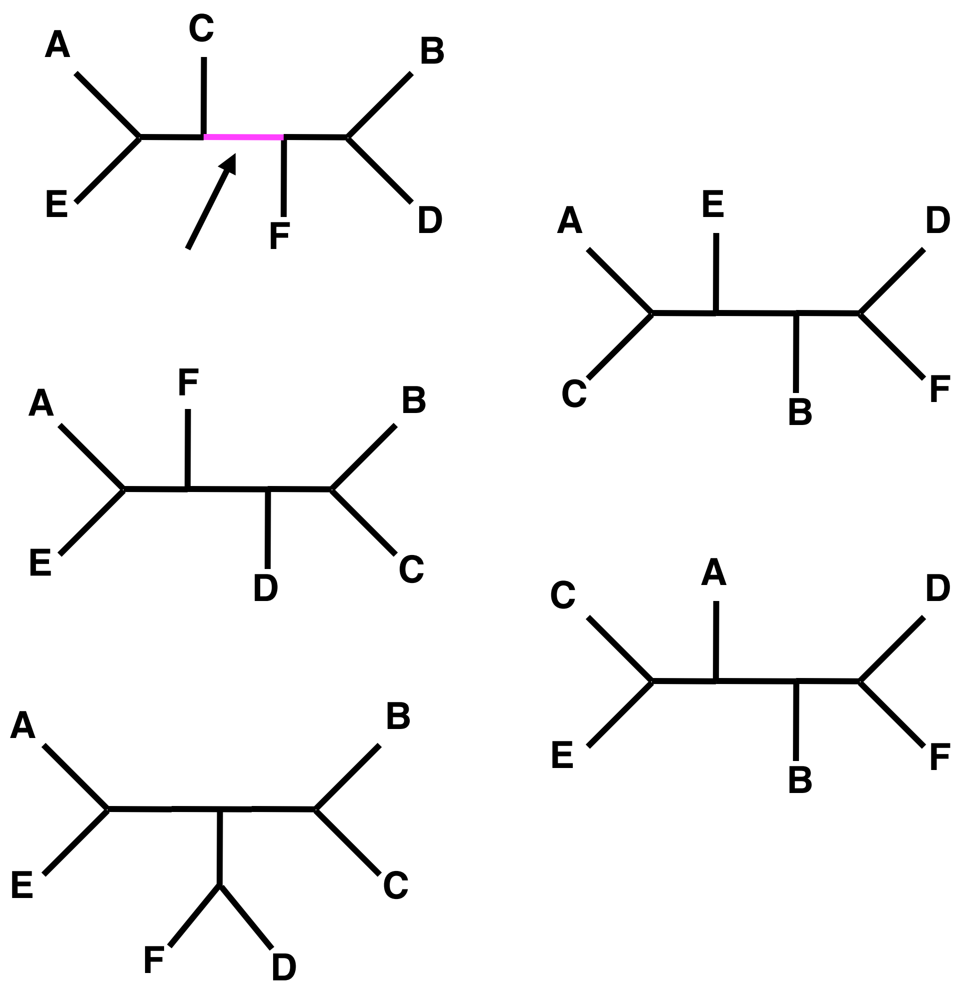
A	B	C	D	E	F	Count
*	-	-	-	*	-	

Majority Rule Consensus Tree



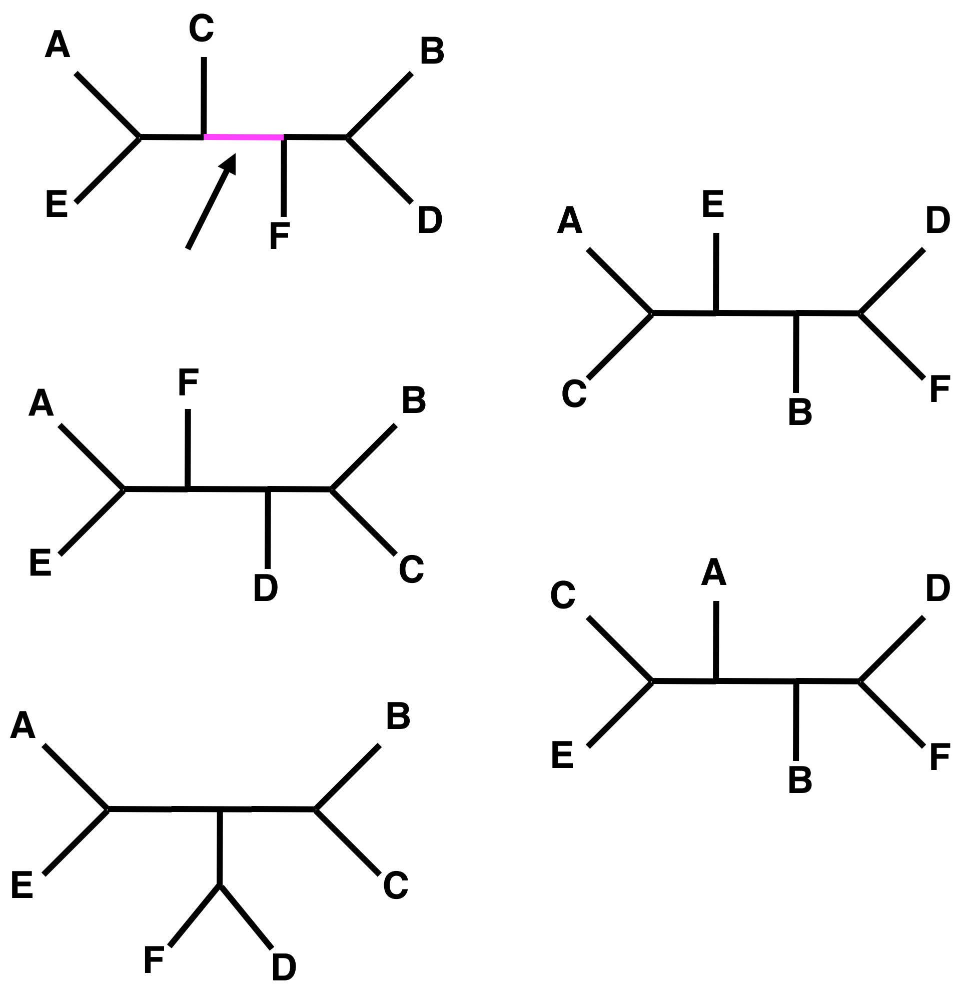
A	B	C	D	E	F	Count
*	-	-	-	*	-	I

Majority Rule Consensus Tree



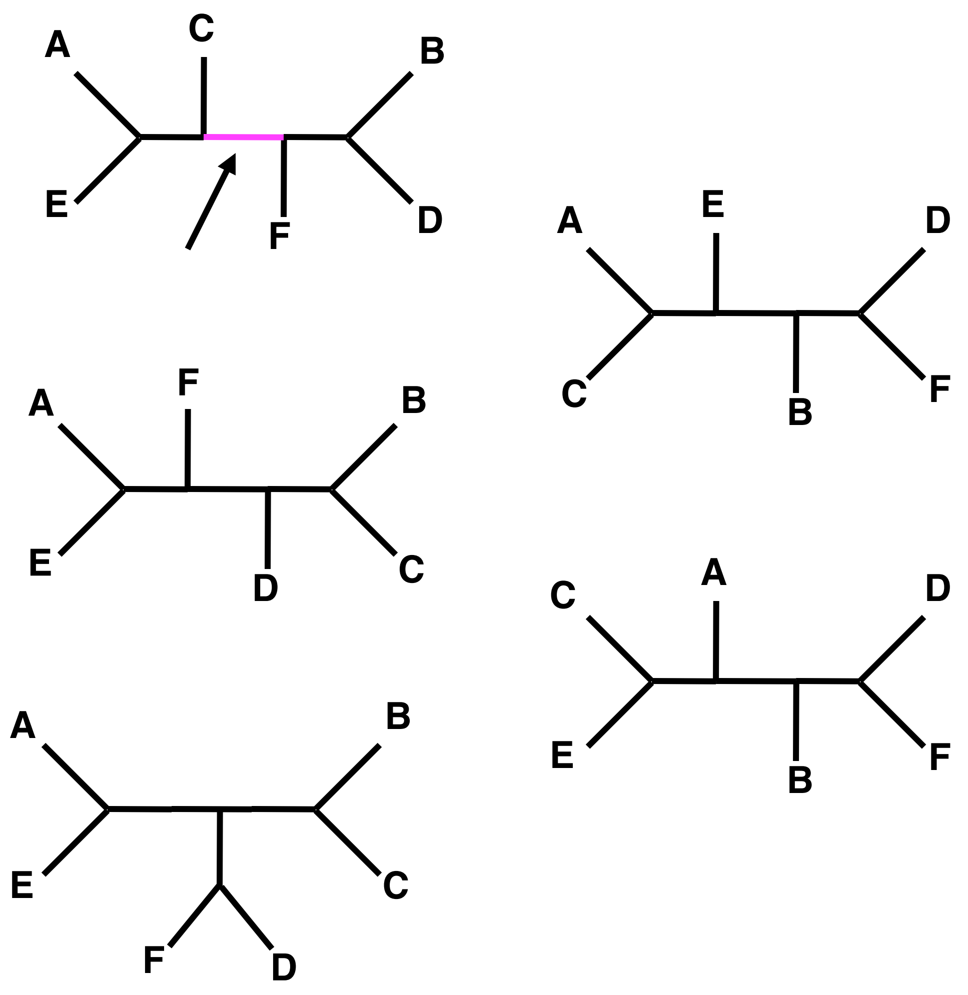
A	B	C	D	E	F	Count
*	-	-	-	*	-	I

Majority Rule Consensus Tree



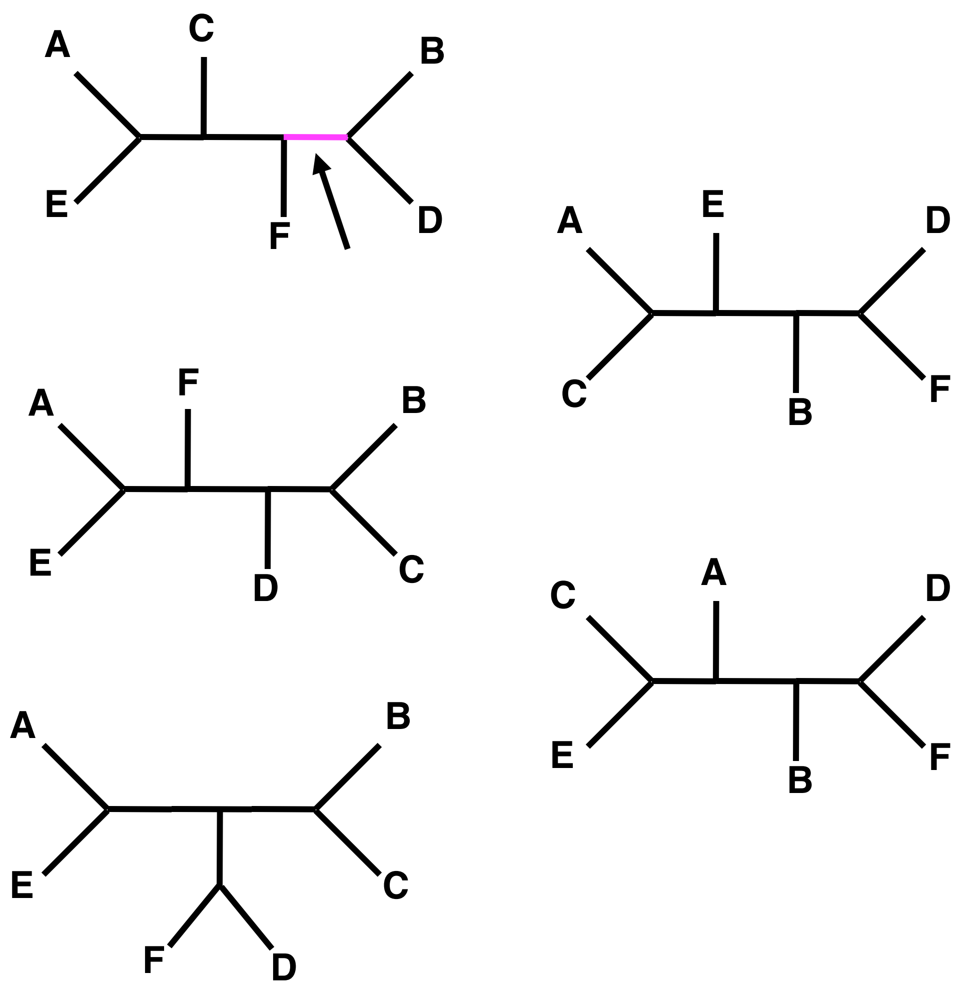
A	B	C	D	E	F	Count
*	-	-	-	*	-	I
*	-	*	-	*	-	

Majority Rule Consensus Tree



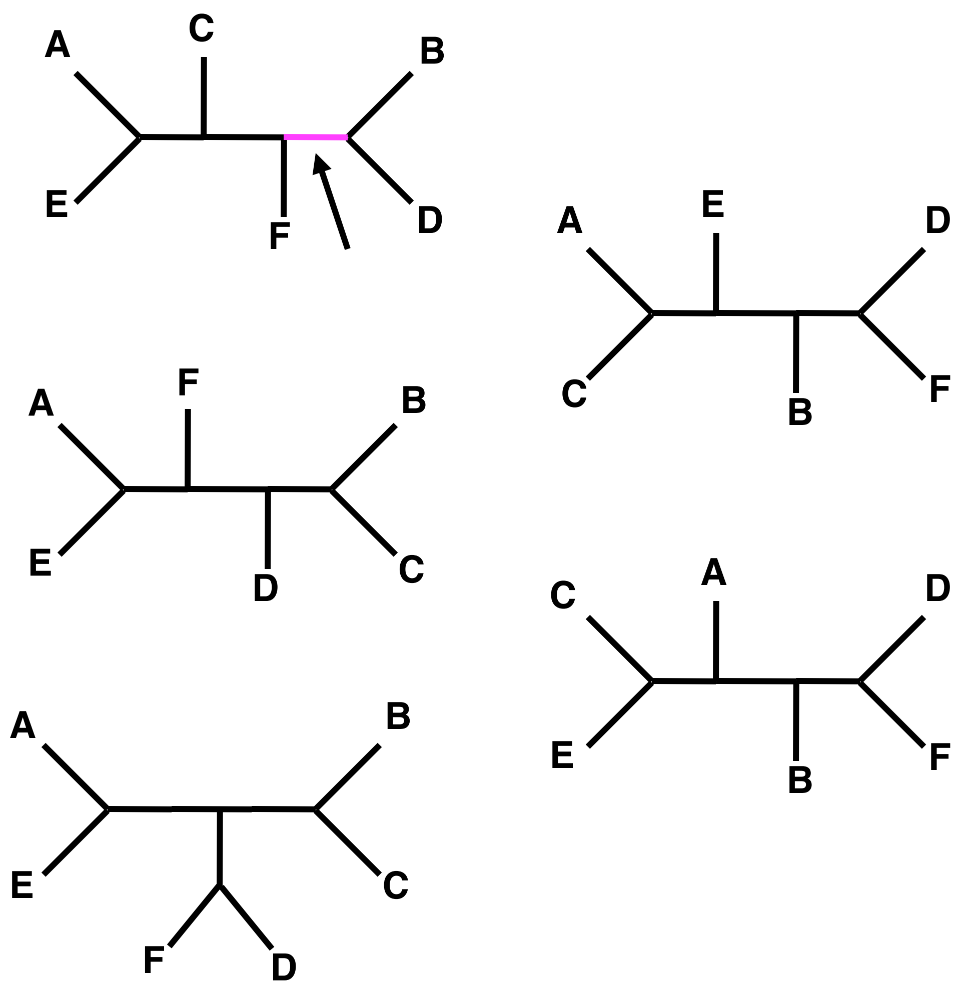
A	B	C	D	E	F	Count
*	-	-	-	*	-	I
*	-	*	-	*	-	I

Majority Rule Consensus Tree



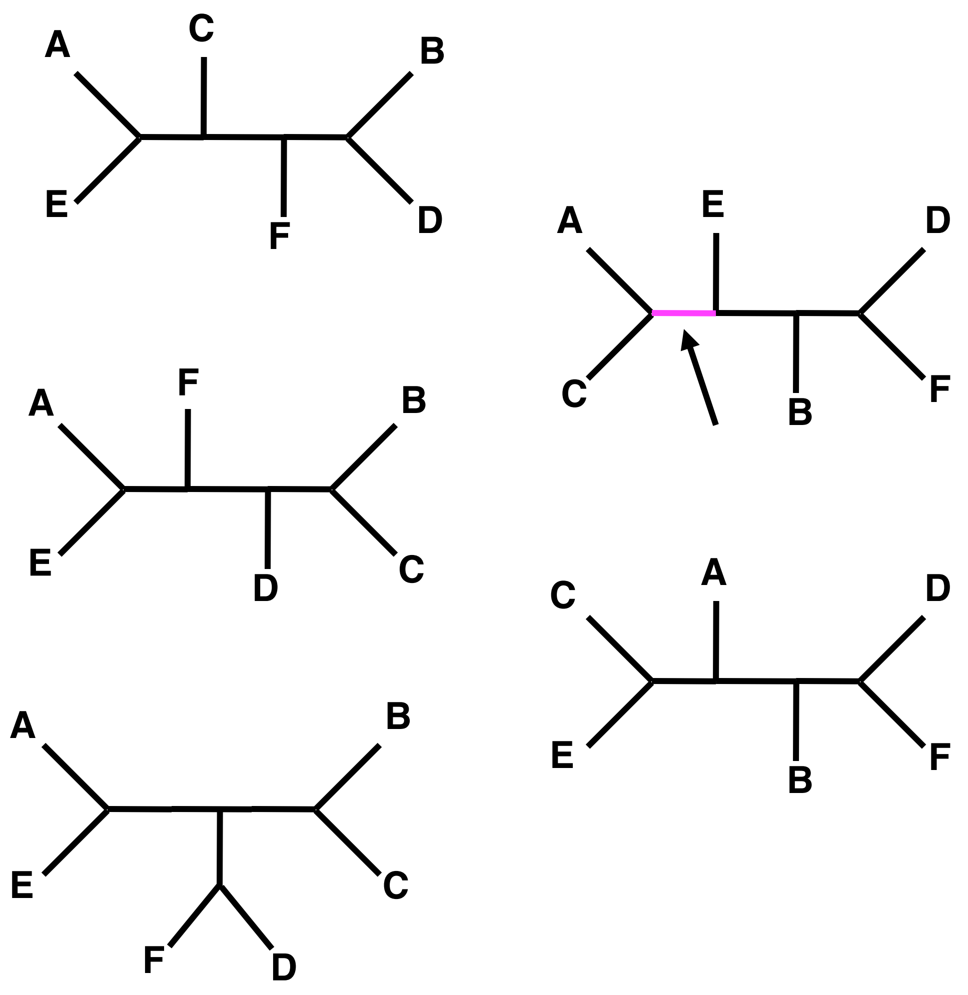
A	B	C	D	E	F	Count
*	-	-	-	*	-	I
*	-	*	-	*	-	I

Majority Rule Consensus Tree



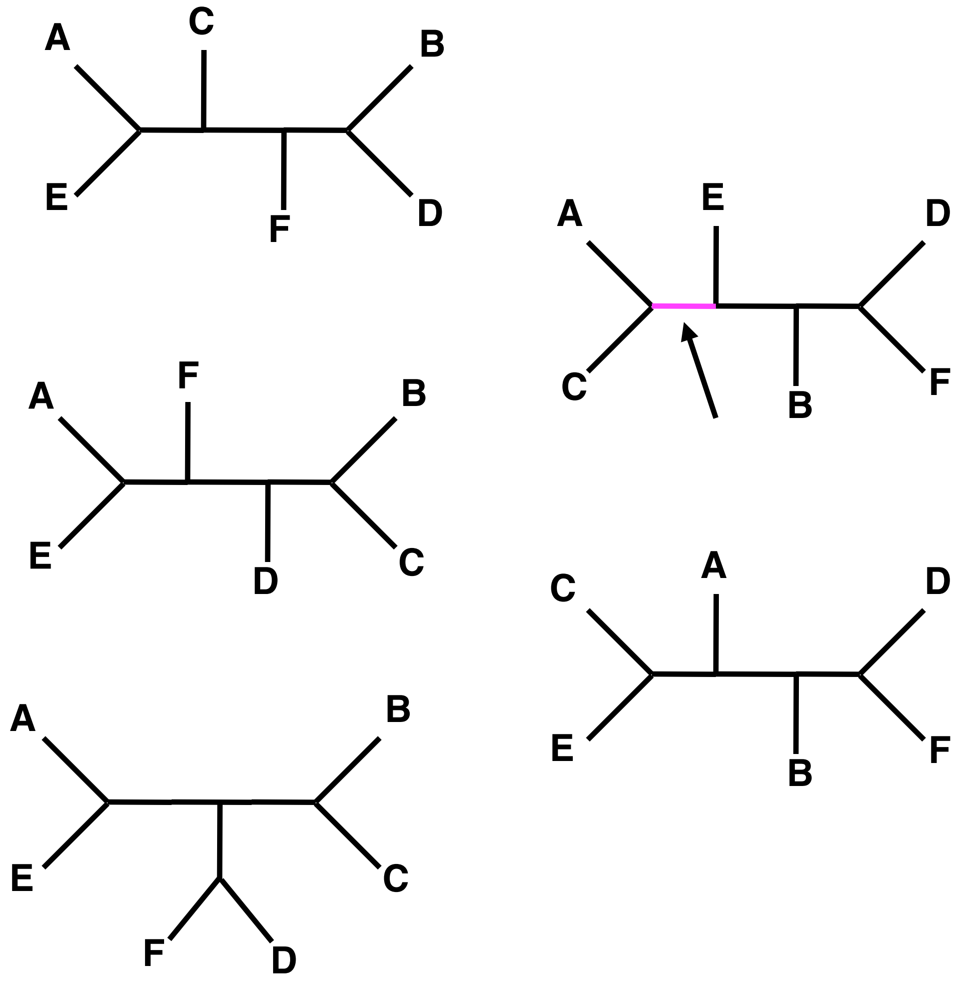
A	B	C	D	E	F	Count
*	-	-	-	*	-	I
*	-	*	-	*	-	I
-	*	-	*	-	-	I

Majority Rule Consensus Tree



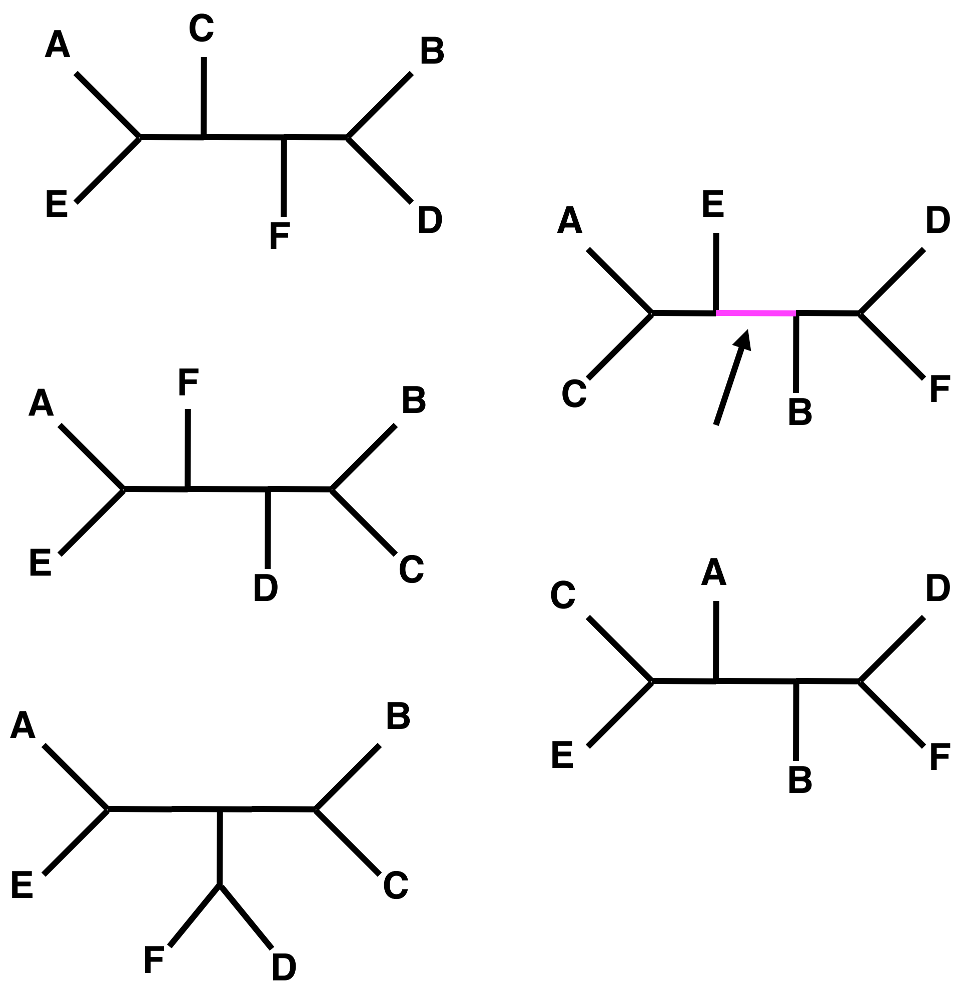
A	B	C	D	E	F	Count
*	-	-	-	*	-	I
*	-	*	-	*	-	I
-	*	-	*	-	-	I

Majority Rule Consensus Tree



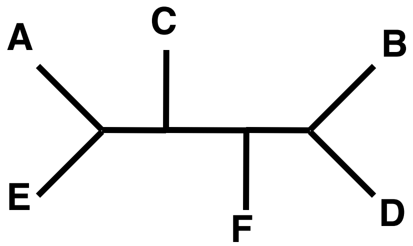
A	B	C	D	E	F	Count
*	-	-	-	*	-	I
*	-	*	-	*	-	I
-	*	-	*	-	-	I
*	-	*	-	-	-	I

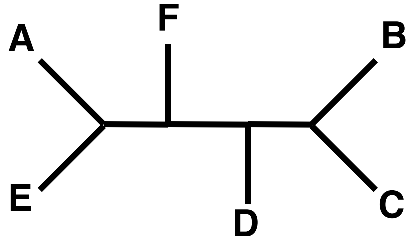
Majority Rule Consensus Tree

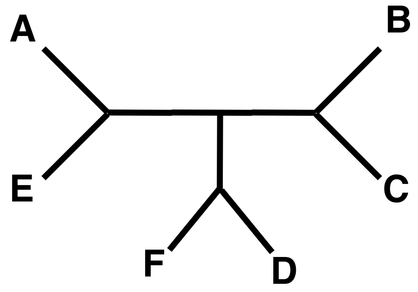


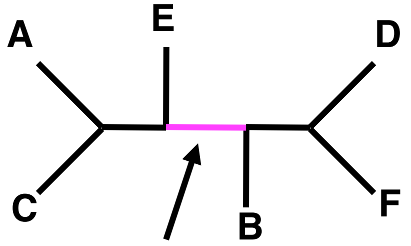
A	B	C	D	E	F	Count
*	–	–	–	*	–	I
*	–	*	–	*	–	I
–	*	–	*	–	–	I
*	–	*	–	–	–	I

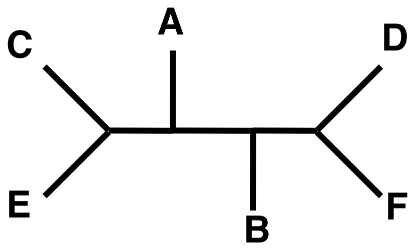
Majority Rule Consensus Tree





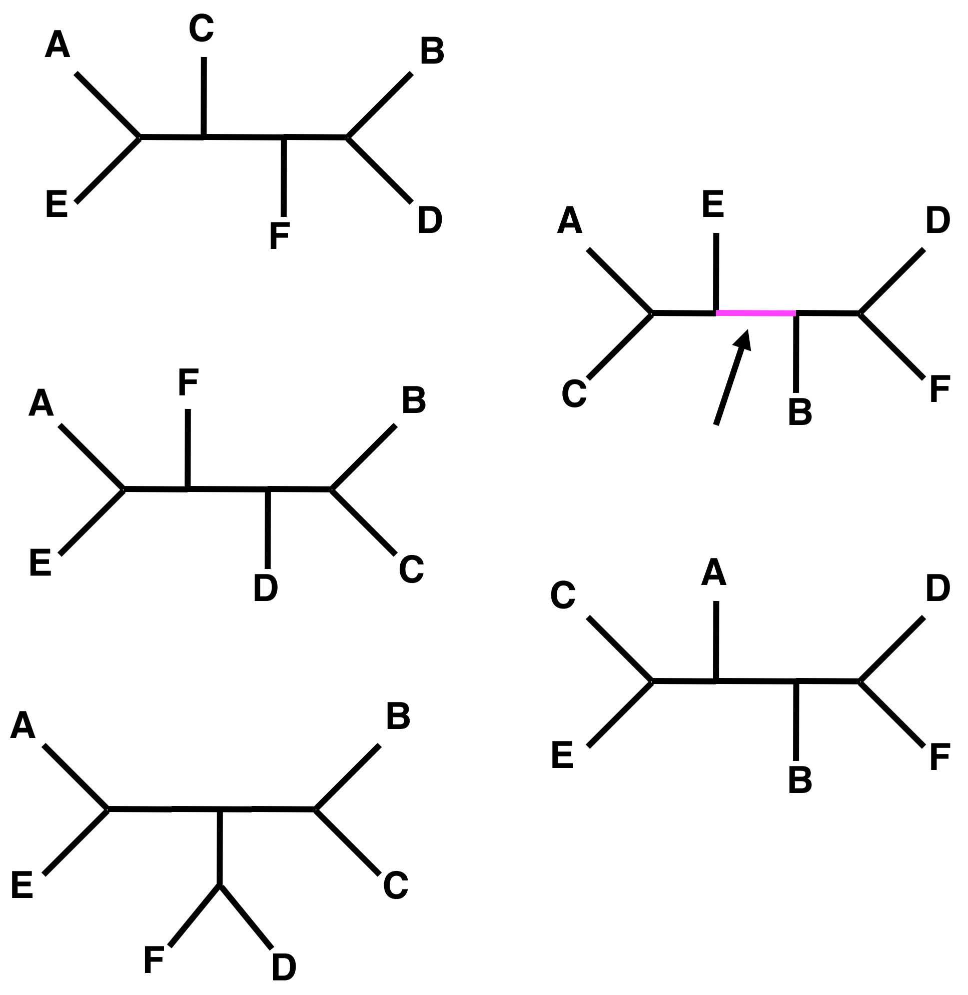






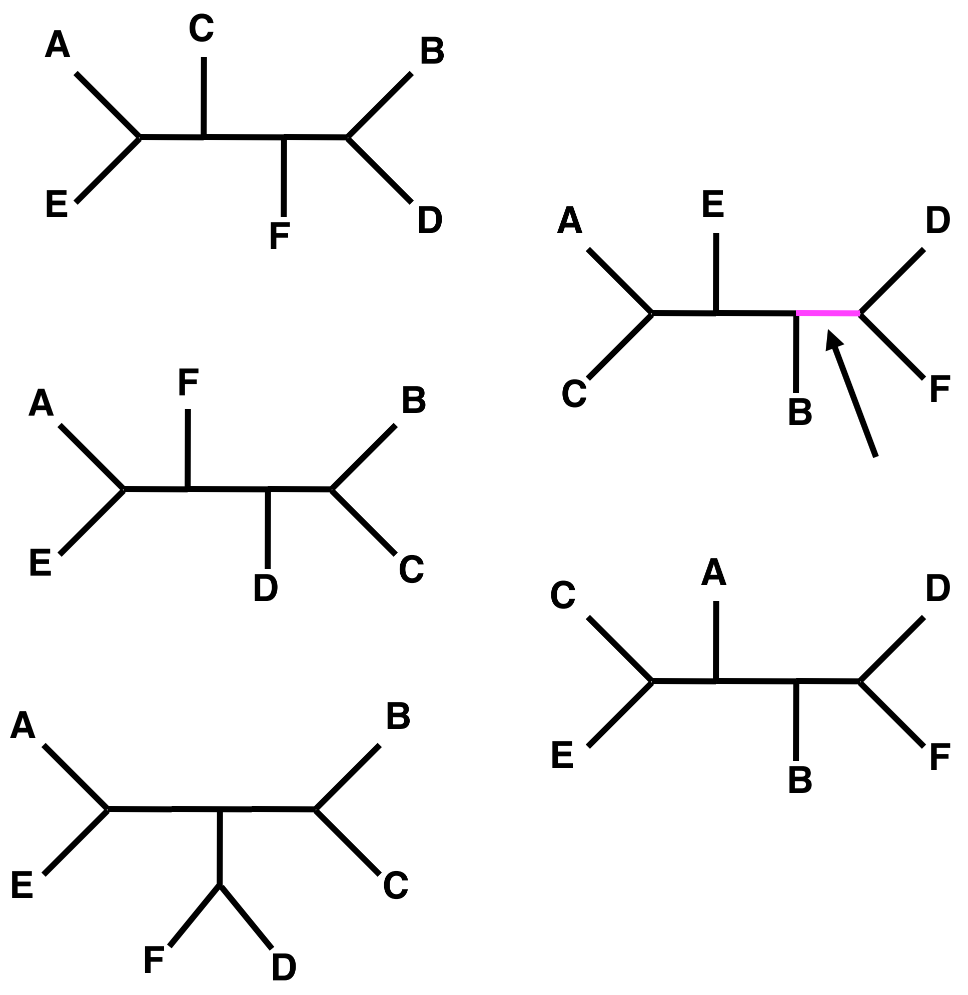
A	B	C	D	E	F	Count
*	-	-	-	*	-	I
*	-	*	-	*	-	I
-	*	-	*	-	-	I
*	-	*	-	-	-	I

Majority Rule Consensus Tree



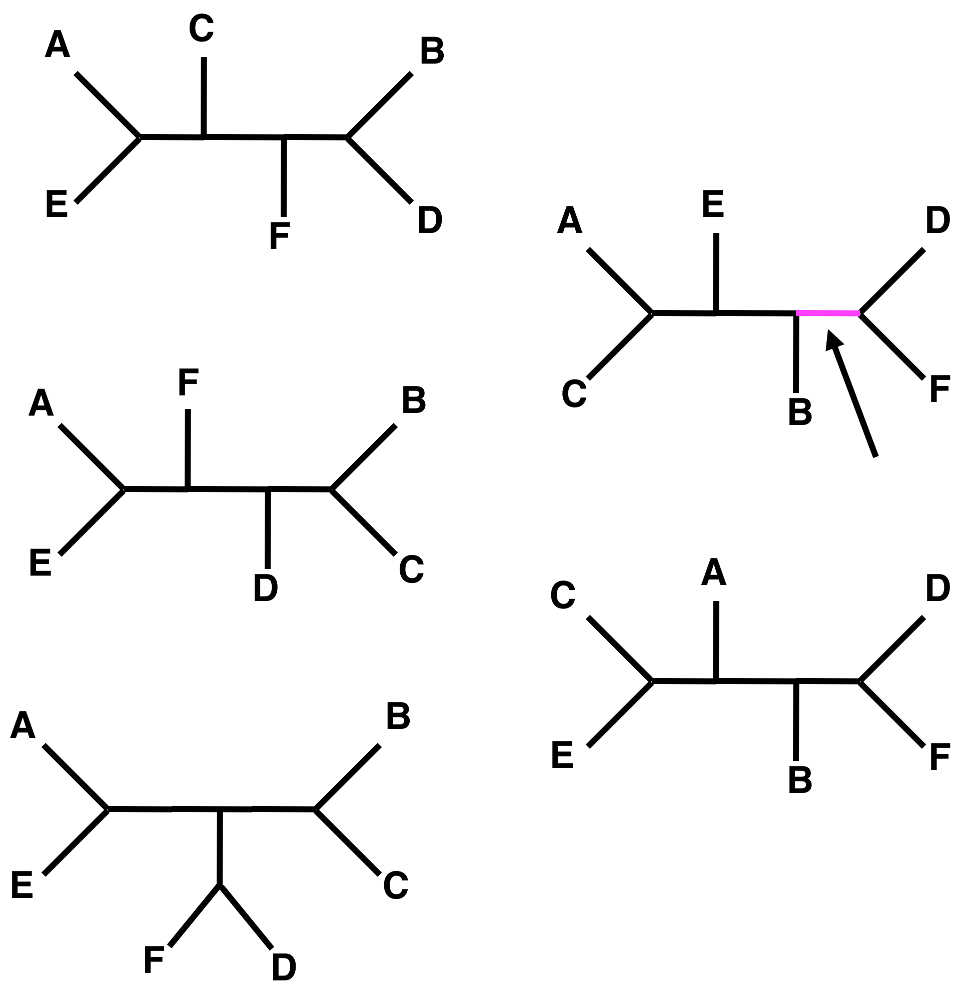
A	B	C	D	E	F	Count
*	-	-	-	*	-	I
*	-	*	-	*	-	II
-	*	-	*	-	-	I
*	-	*	-	-	-	I

Majority Rule Consensus Tree



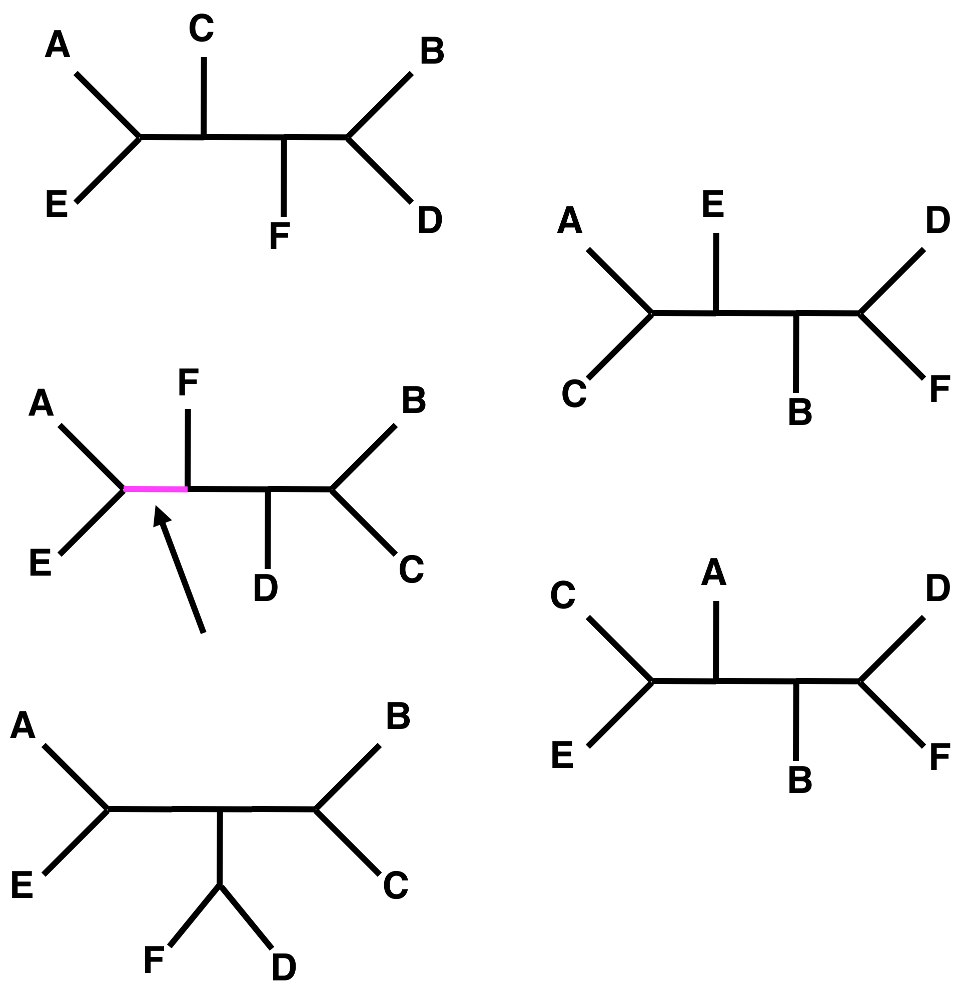
A	B	C	D	E	F	Count
*	-	-	-	*	-	I
*	-	*	-	*	-	II
-	*	-	*	-	-	I
*	-	*	-	-	-	I

Majority Rule Consensus Tree



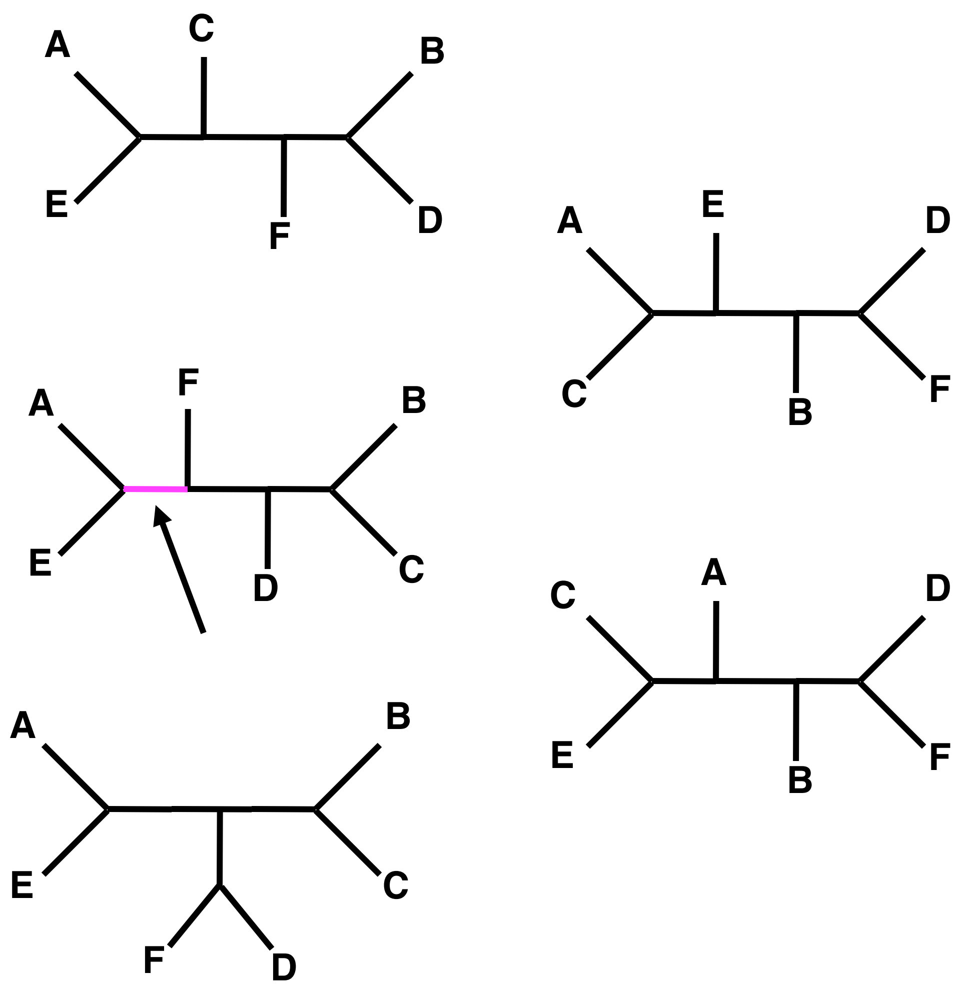
A	B	C	D	E	F	Count
*	-	-	-	*	-	I
*	-	*	-	*	-	II
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	I

Majority Rule Consensus Tree



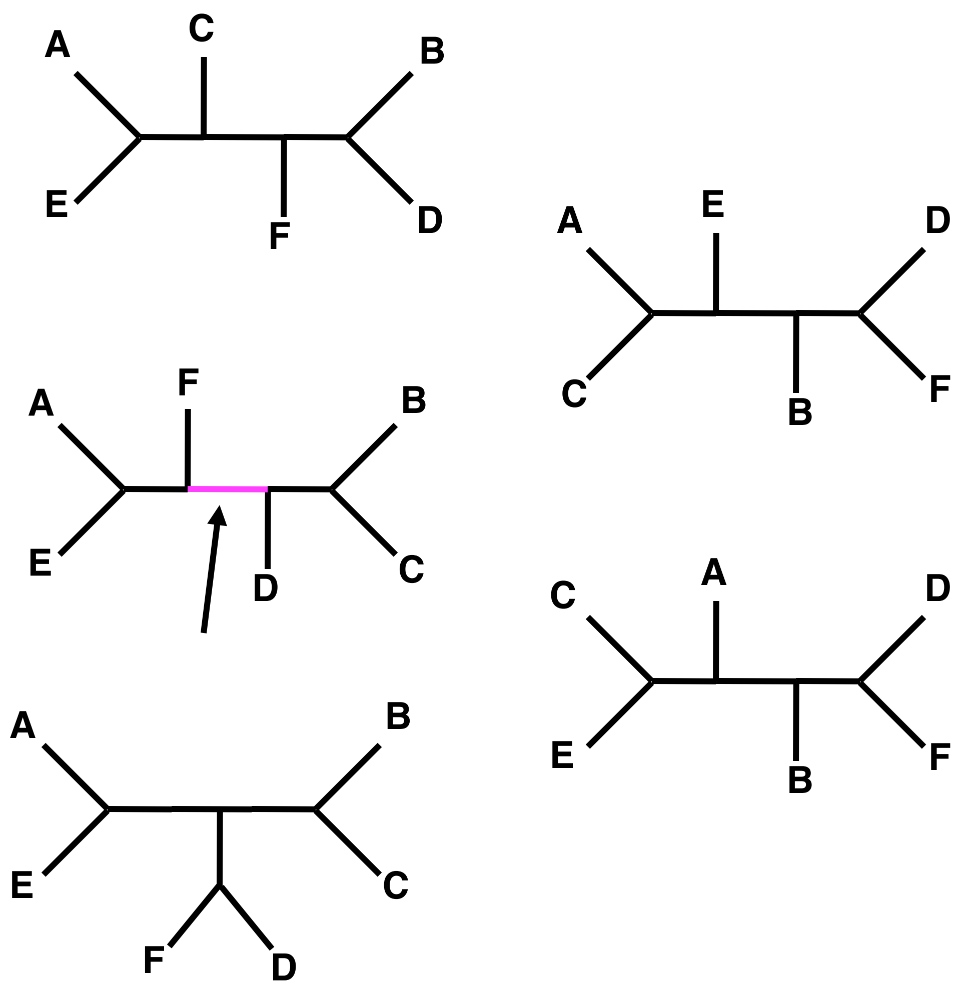
A	B	C	D	E	F	Count
*	-	-	-	*	-	I
*	-	*	-	*	-	II
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	I

Majority Rule Consensus Tree



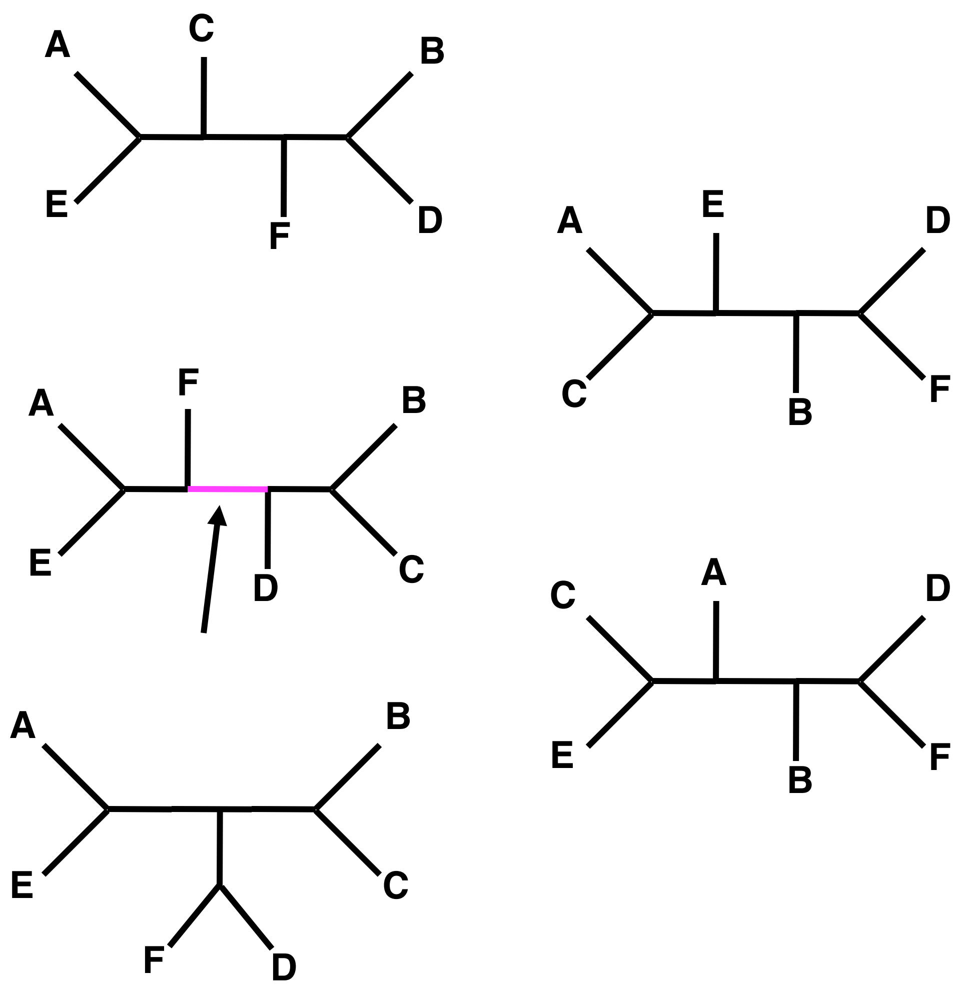
A	B	C	D	E	F	Count
*	-	-	-	*	-	II
*	-	*	-	*	-	II
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	I

Majority Rule Consensus Tree



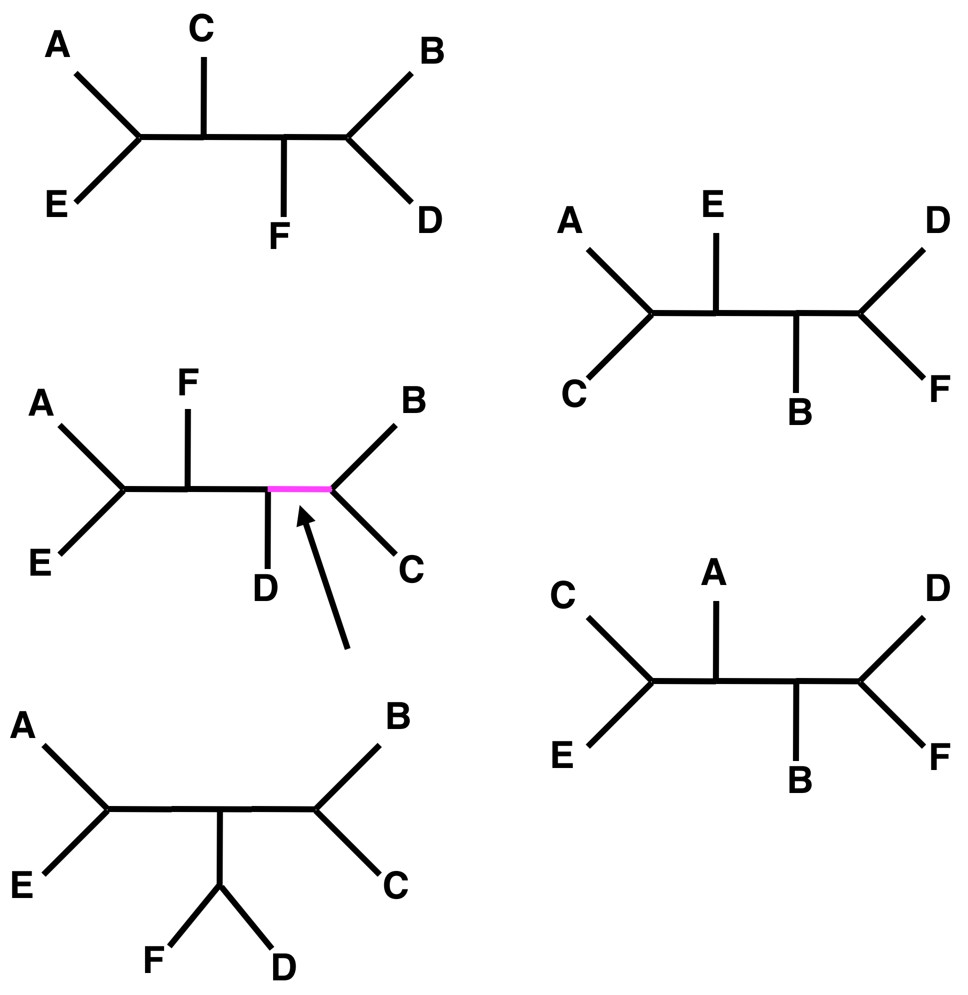
A	B	C	D	E	F	Count
*	-	-	-	*	-	II
*	-	*	-	*	-	II
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	I

Majority Rule Consensus Tree



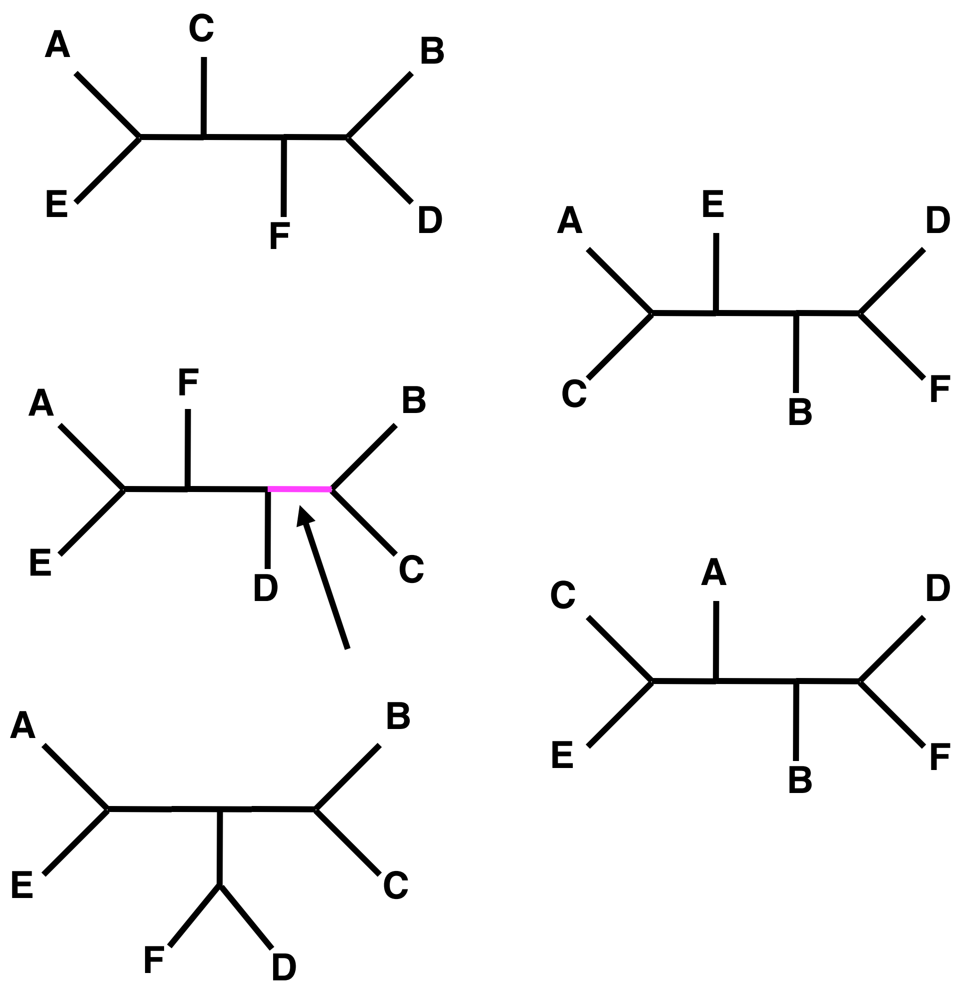
A	B	C	D	E	F	Count
*	-	-	-	*	-	II
*	-	*	-	*	-	II
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	I
*	-	-	-	*	*	I

Majority Rule Consensus Tree



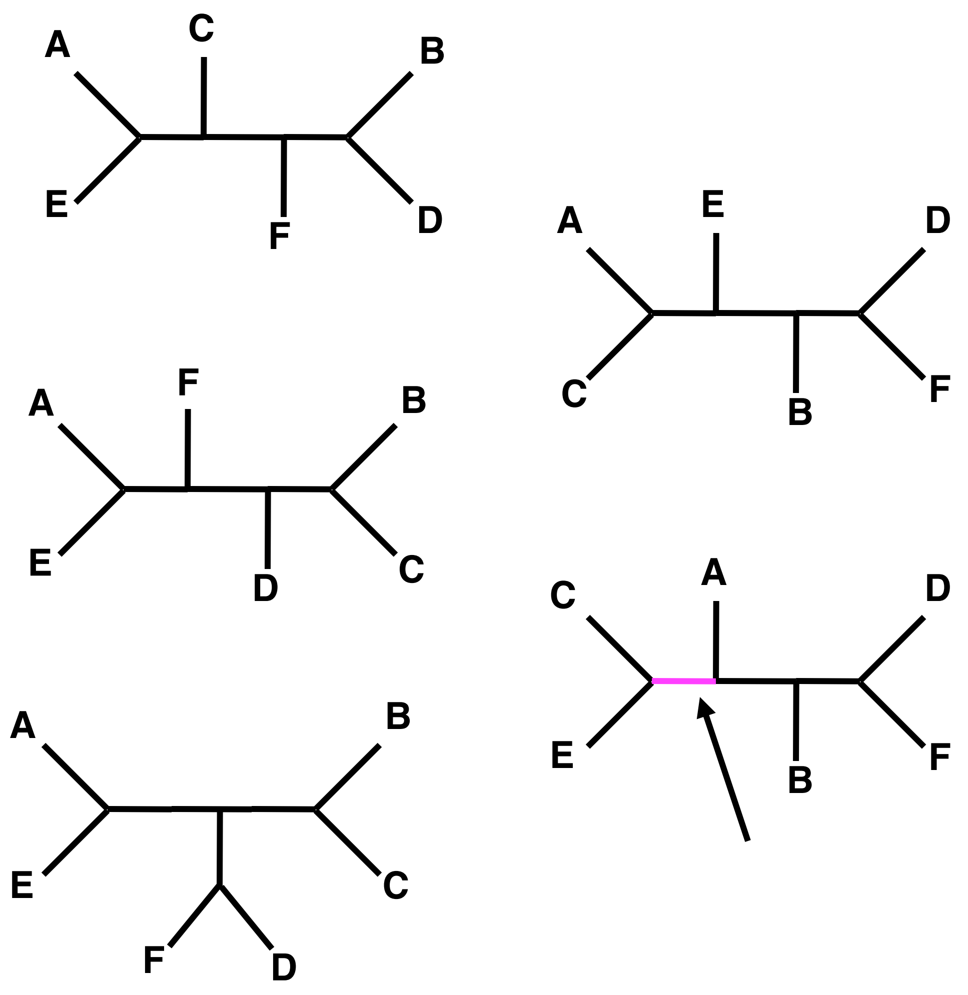
A	B	C	D	E	F	Count
*	–	–	–	*	–	II
*	–	*	–	*	–	II
–	*	–	*	–	–	I
*	–	*	–	–	–	I
–	–	–	*	–	*	I
*	–	–	–	*	*	I

Majority Rule Consensus Tree



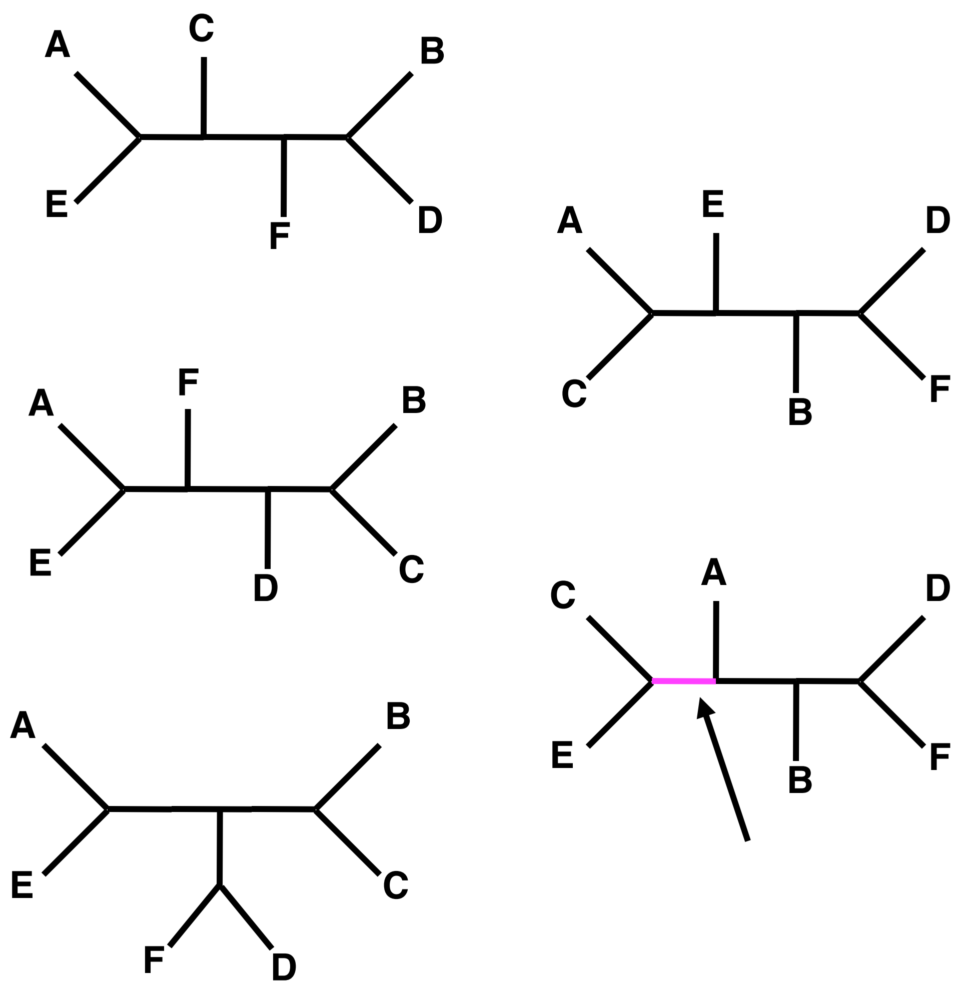
A	B	C	D	E	F	Count
*	-	-	-	*	-	II
*	-	*	-	*	-	II
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	I
*	-	-	-	*	*	I
-	*	*	-	-	-	I

Majority Rule Consensus Tree



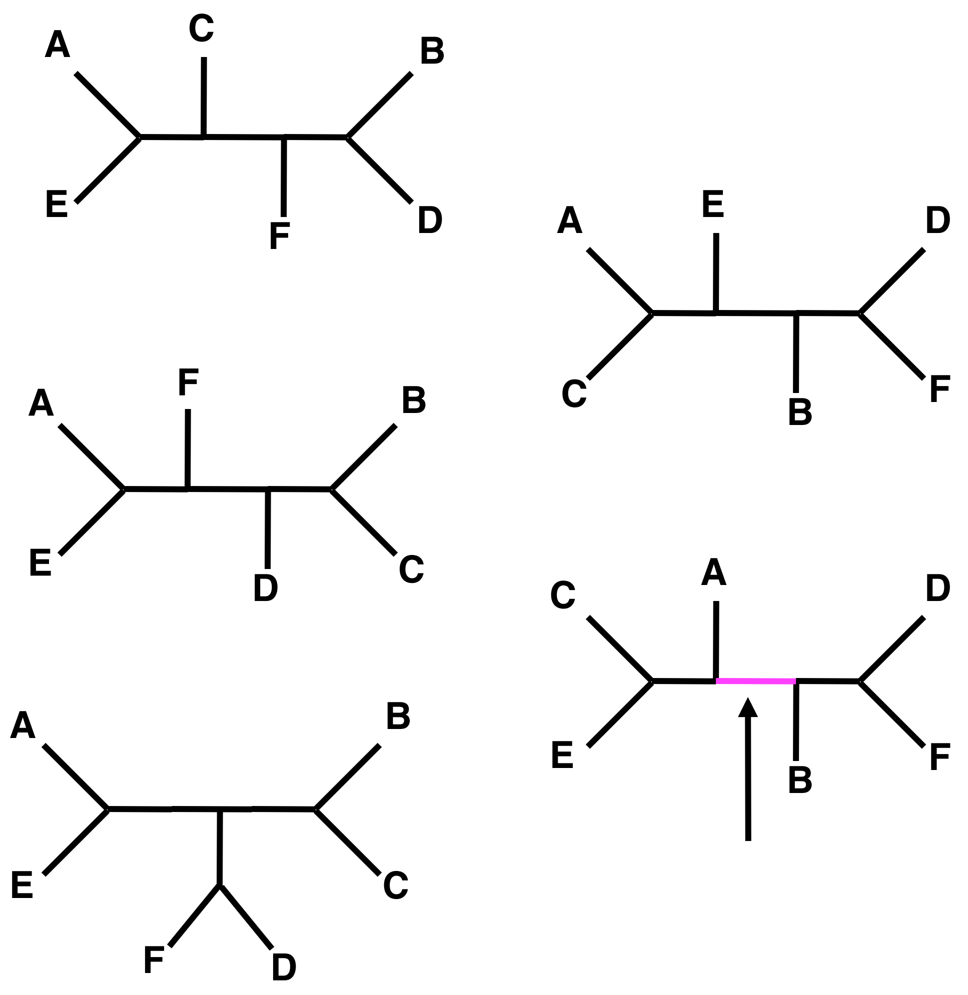
A	B	C	D	E	F	Count
*	-	-	-	*	-	II
*	-	*	-	*	-	II
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	I
*	-	-	-	*	*	I
-	*	*	-	-	-	I

Majority Rule Consensus Tree



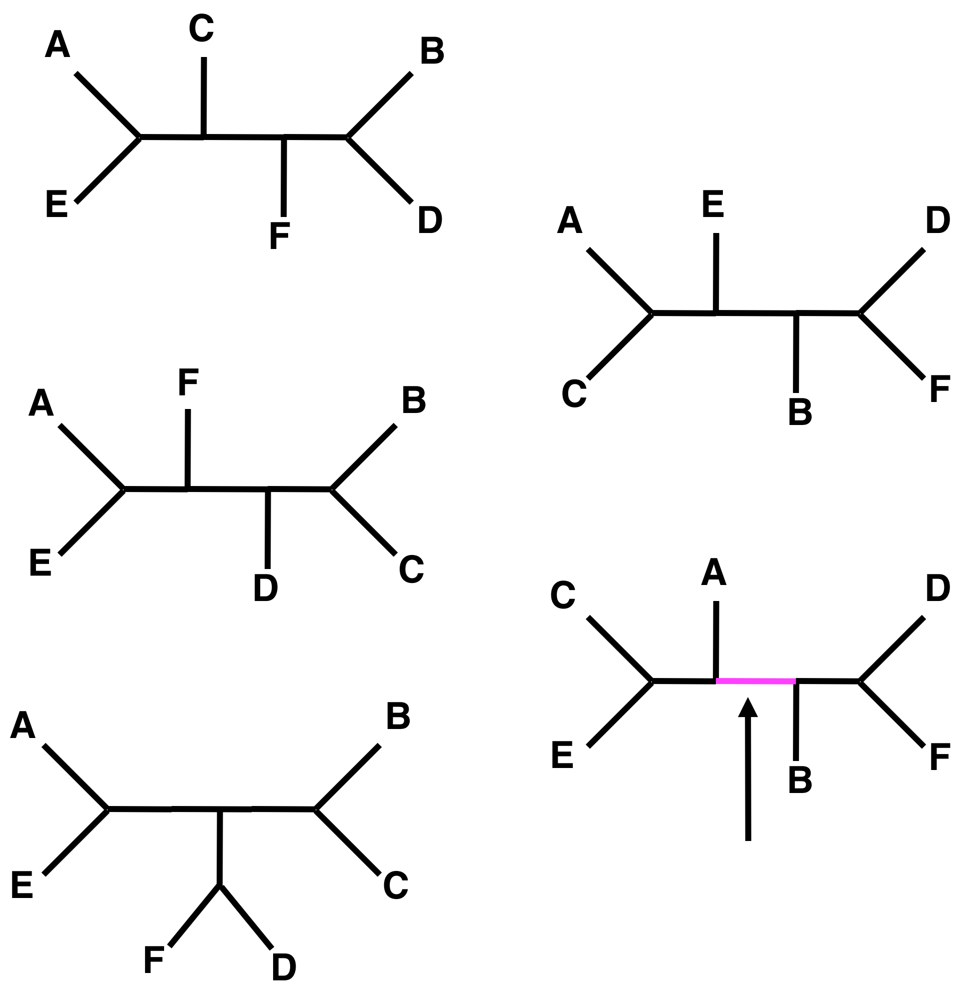
A	B	C	D	E	F	Count
*	-	-	-	*	-	II
*	-	*	-	*	-	II
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	I
*	-	-	-	*	*	I
-	*	*	-	-	-	I
-	-	*	-	*	-	I

Majority Rule Consensus Tree



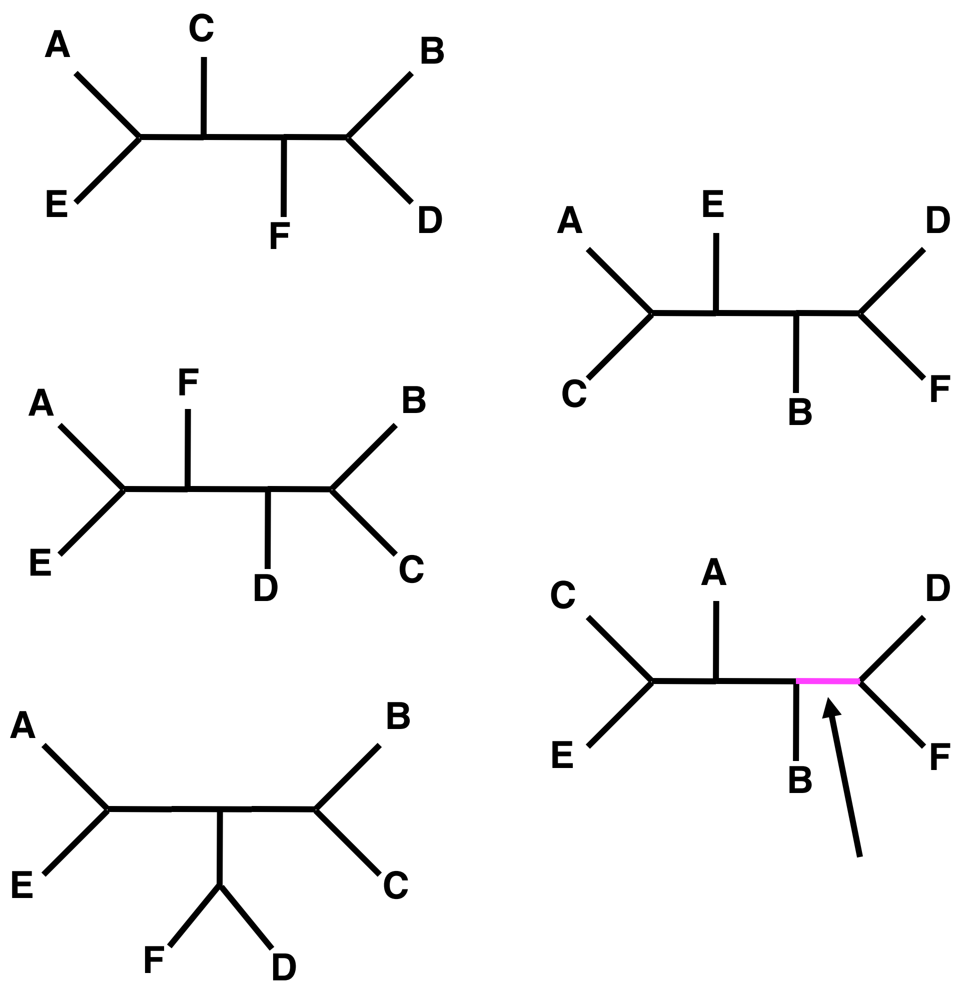
A	B	C	D	E	F	Count
*	-	-	-	*	-	II
*	-	*	-	*	-	II
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	I
*	-	-	-	*	*	I
-	*	*	-	-	-	I
-	-	*	-	*	-	I

Majority Rule Consensus Tree



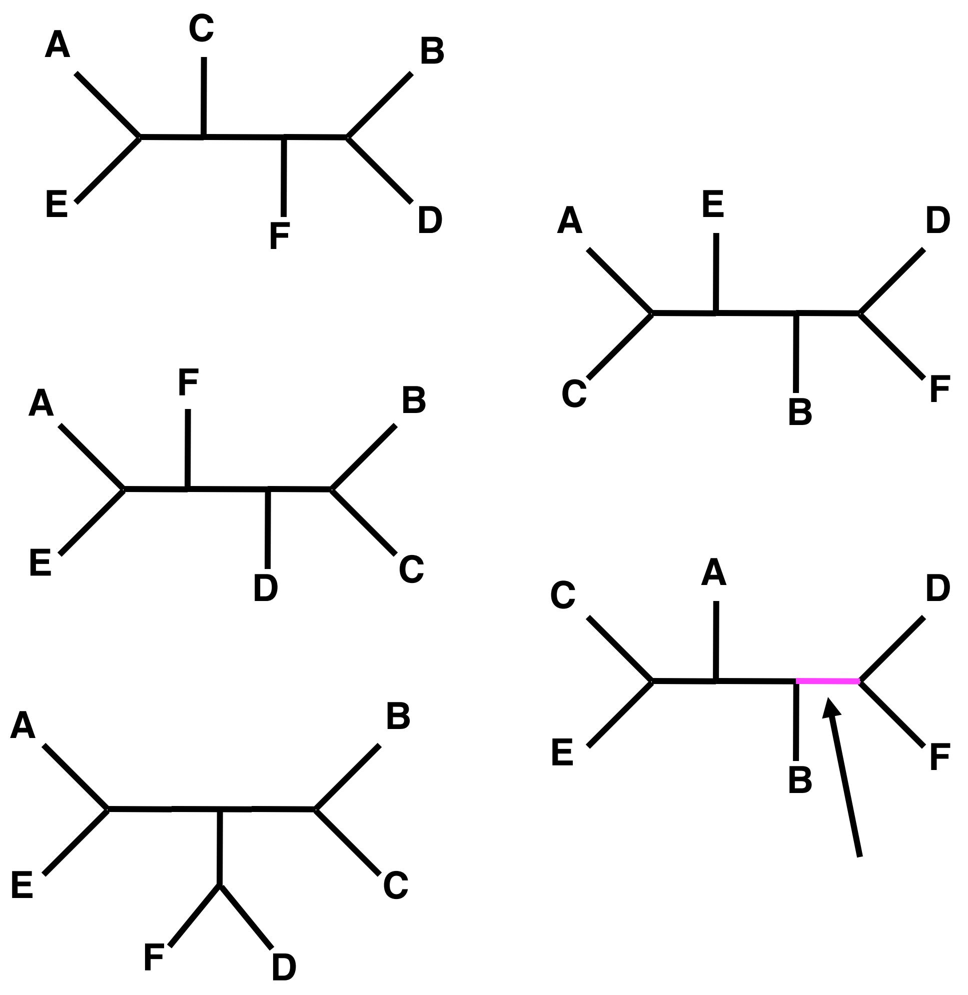
A	B	C	D	E	F	Count
*	-	-	-	*	-	II
*	-	*	-	*	-	III
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	I
*	-	-	-	*	*	I
-	*	*	-	-	-	I
-	-	*	-	*	-	I

Majority Rule Consensus Tree



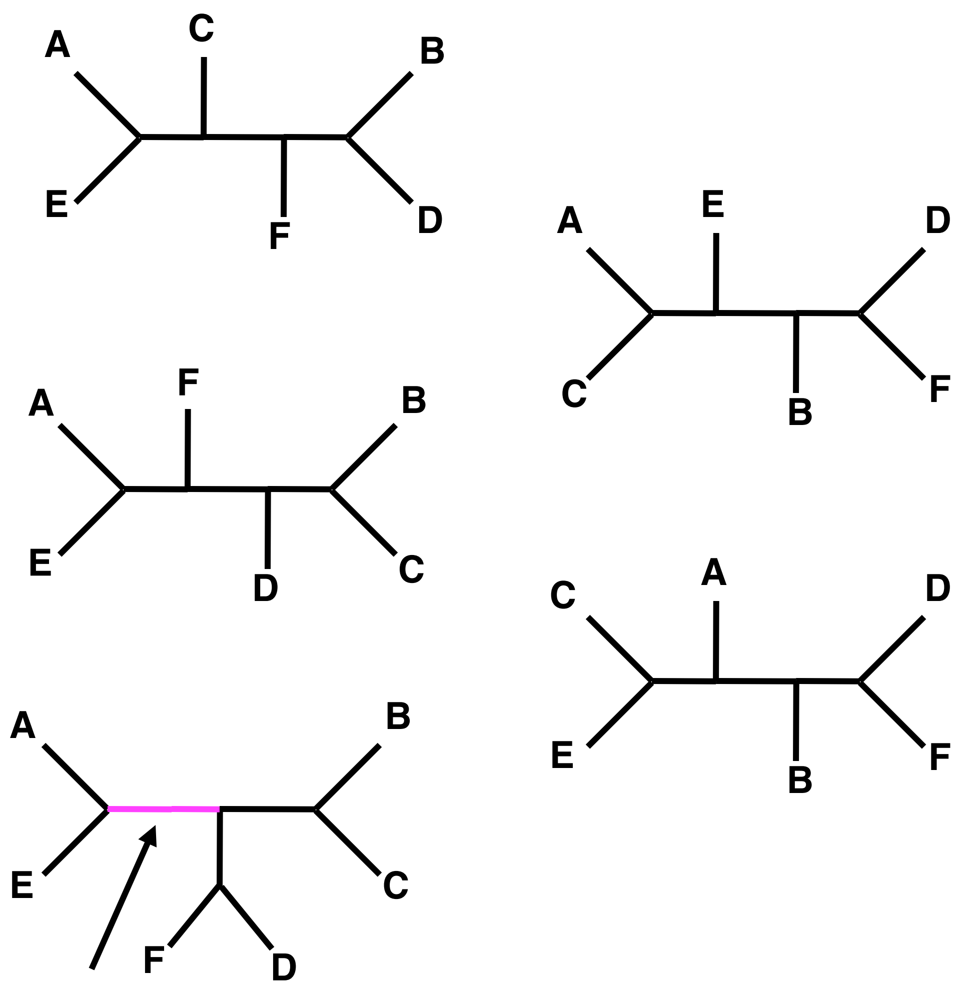
A	B	C	D	E	F	Count
*	-	-	-	*	-	II
*	-	*	-	*	-	III
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	I
*	-	-	-	*	*	I
-	*	*	-	-	-	I
-	-	*	-	*	-	I

Majority Rule Consensus Tree



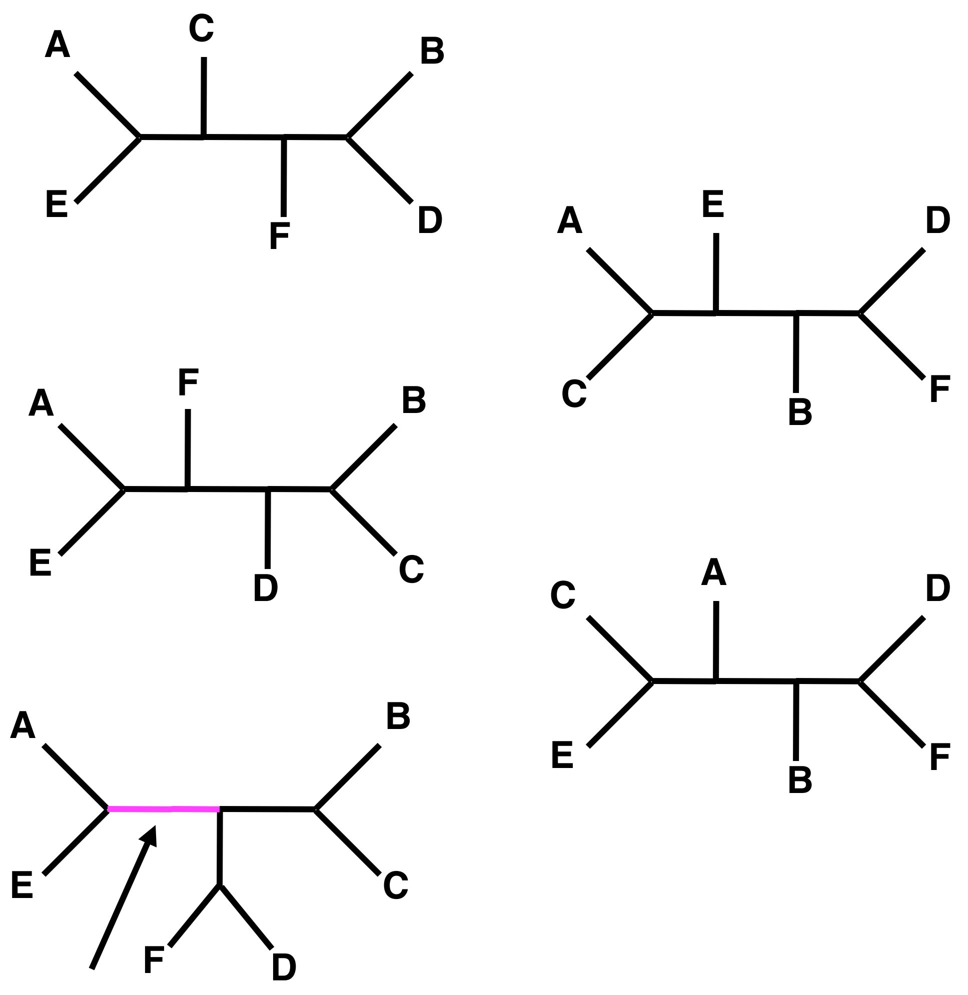
A	B	C	D	E	F	Count
*	-	-	-	*	-	II
*	-	*	-	*	-	III
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	II
*	-	-	-	*	*	I
-	*	*	-	-	-	I
-	-	*	-	*	-	I

Majority Rule Consensus Tree



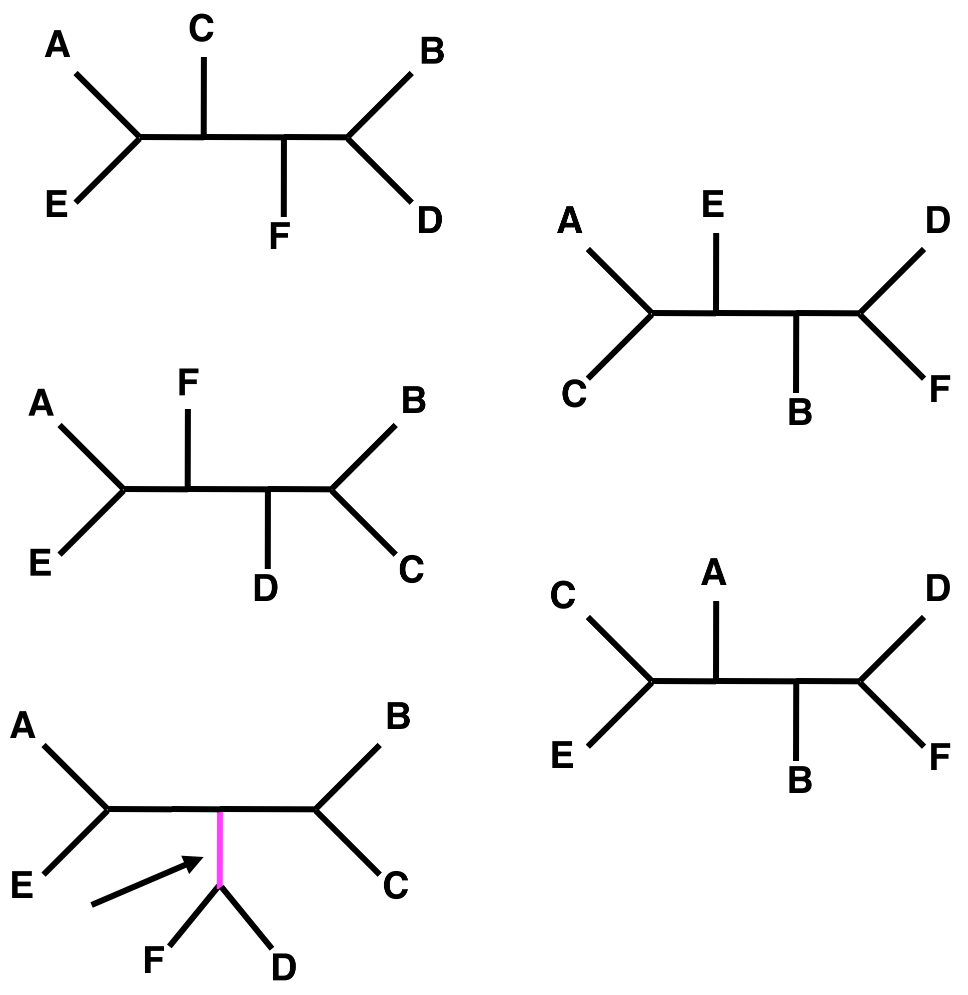
A	B	C	D	E	F	Count
*	–	–	–	*	–	II
*	–	*	–	*	–	III
–	*	–	*	–	–	I
*	–	*	–	–	–	I
–	–	–	*	–	*	II
*	–	–	–	*	*	I
–	*	*	–	–	–	I
–	–	*	–	*	–	I

Majority Rule Consensus Tree



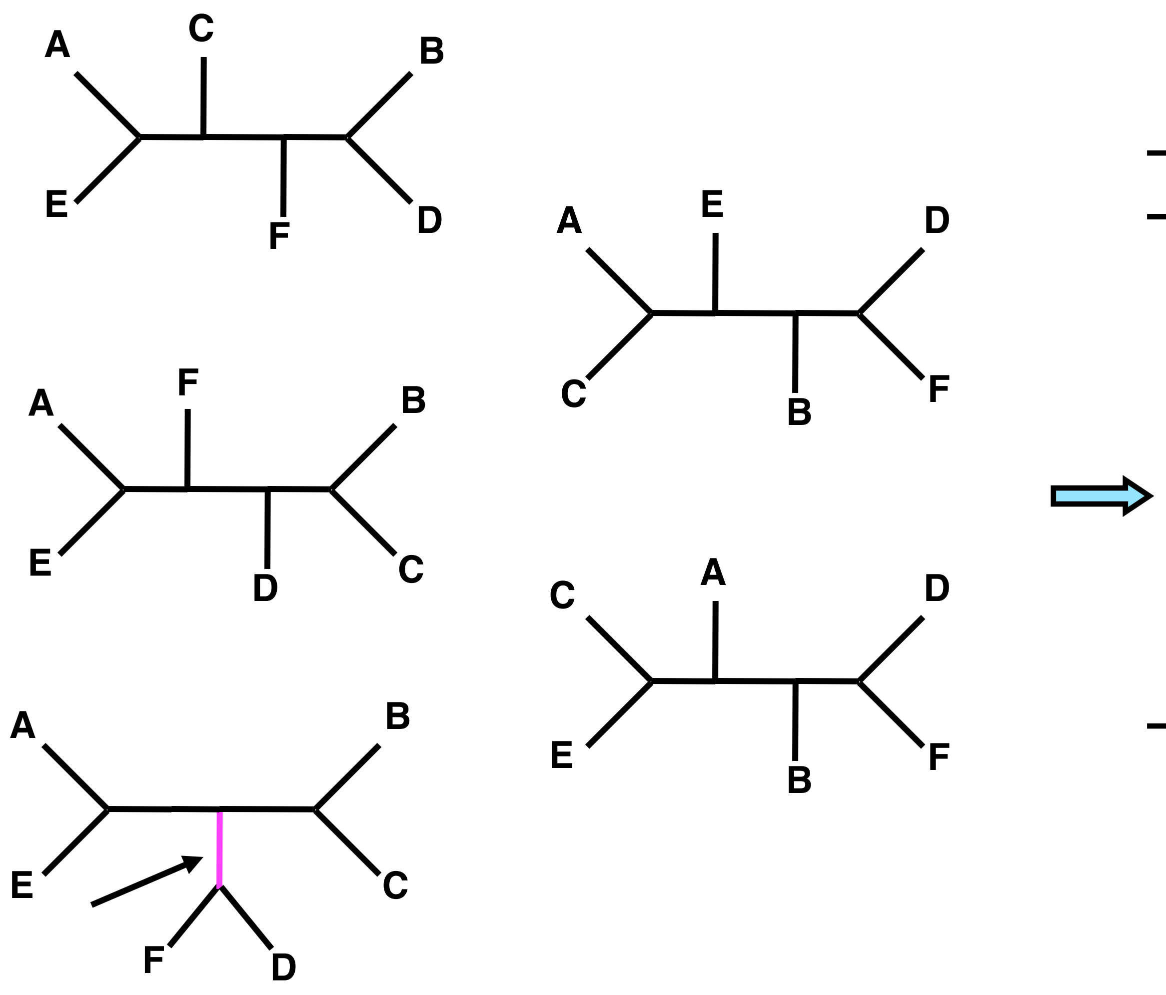
A	B	C	D	E	F	Count
*	-	-	-	*	-	III
*	-	*	-	*	-	III
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	II
*	-	-	-	*	*	I
-	*	*	-	-	-	I
-	-	*	-	*	-	I

Majority Rule Consensus Tree



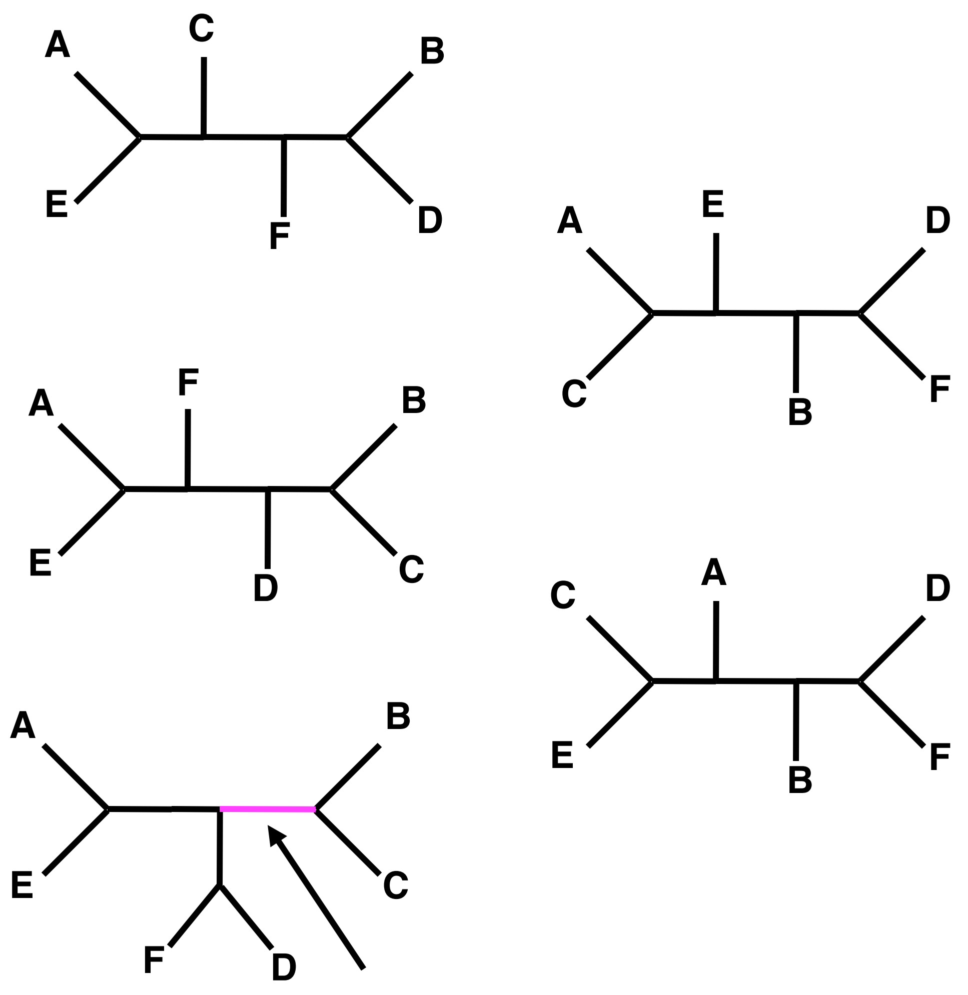
A	B	C	D	E	F	Count
*	-	-	-	*	-	III
*	-	*	-	*	-	III
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	II
*	-	-	-	*	*	I
-	*	*	-	-	-	I
-	-	*	-	*	-	I

Majority Rule Consensus Tree



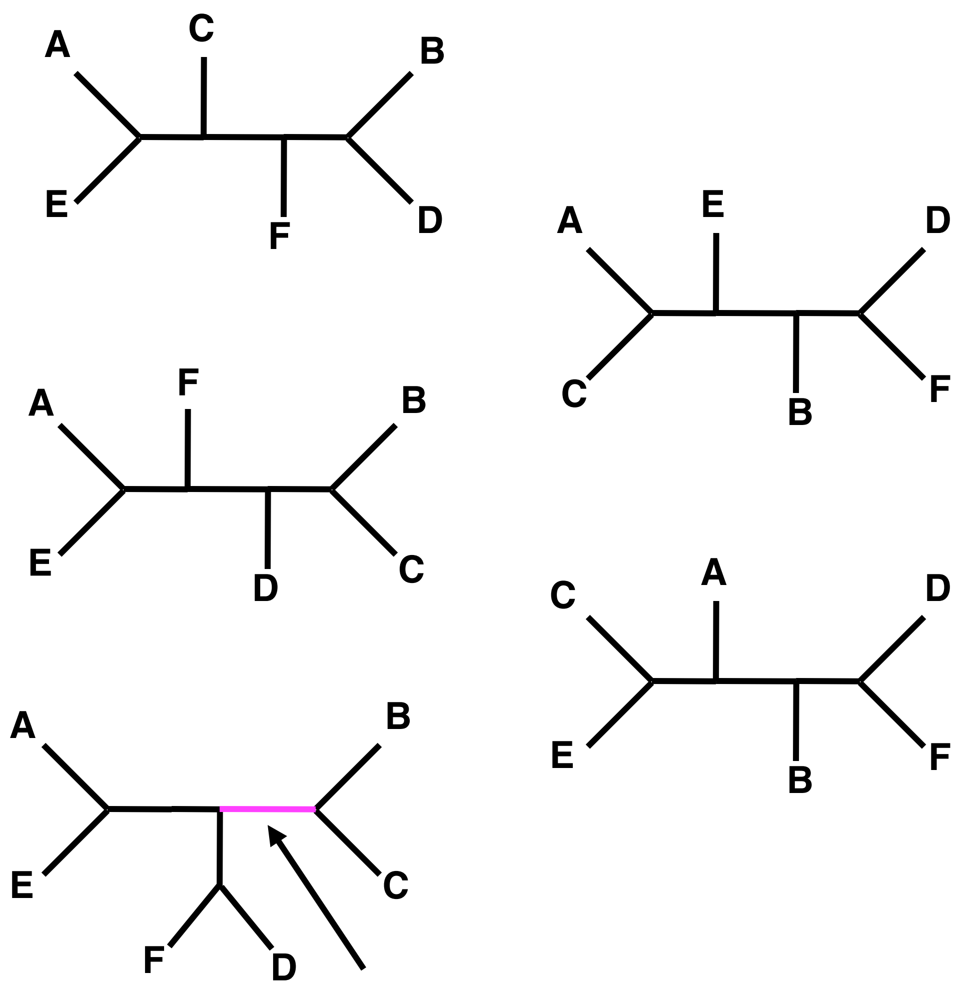
A	B	C	D	E	F	Count
*	-	-	-	*	-	III
*	-	*	-	*	-	III
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	III
*	-	-	-	*	*	I
-	*	*	-	-	-	I
-	-	*	-	*	-	I

Majority Rule Consensus Tree



A	B	C	D	E	F	Count
*	-	-	-	*	-	III
*	-	*	-	*	-	III
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	III
*	-	-	-	*	*	I
-	*	*	-	-	-	I
-	-	*	-	*	-	I

Majority Rule Consensus Tree



A	B	C	D	E	F	Count
*	-	-	-	*	-	III
*	-	*	-	*	-	III
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	III
*	-	-	-	*	*	I
-	*	*	-	-	-	II
-	-	*	-	*	-	I

Majority Rule Consensus Tree

A	B	C	D	E	F	Count
*	-	-	-	*	-	III
*	-	*	-	*	-	III
-	*	-	*	-	-	I
*	-	*	-	-	-	I
-	-	-	*	-	*	III
*	-	-	-	*	*	I
-	*	*	-	-	-	II
-	-	*	-	*	-	I

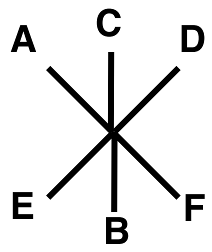
Majority Rule Consensus Tree

A	B	C	D	E	F	Freq
*	-	-	-	*	-	60
*	-	*	-	*	-	60
-	*	-	*	-	-	20
*	-	*	-	-	-	20
-	-	-	*	-	*	60
*	-	-	-	*	*	20
-	*	*	-	-	-	40
-	-	*	-	*	-	20

Majority Rule Consensus Tree

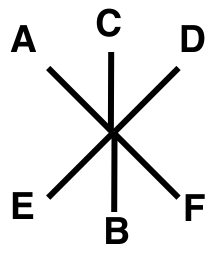
A	B	C	D	E	F	Freq
*	-	-	-	*	-	60
*	-	*	-	*	-	60
-	-	-	*	-	*	60
-	*	*	-	-	-	40
-	*	-	*	-	-	20
*	-	-	-	*	*	20
*	-	*	-	-	-	20
-	-	*	-	*	-	20

Majority Rule Consensus Tree



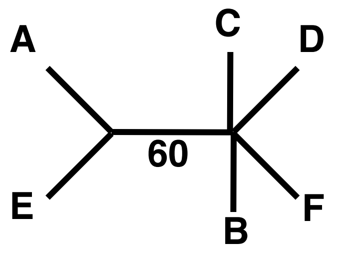
A	B	C	D	E	F	Freq
*	-	-	-	*	-	60
*	-	*	-	*	-	60
-	-	-	*	-	*	60
-	*	*	-	-	-	40
-	*	-	*	-	-	20
*	-	-	-	*	*	20
*	-	*	-	-	-	20
-	-	*	-	*	-	20

Majority Rule Consensus Tree



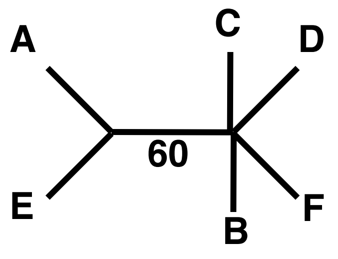
A	B	C	D	E	F	Freq
*	-	-	-	*	-	60
*	-	*	-	*	-	60
-	-	-	*	-	*	60
-	*	*	-	-	-	40
-	*	-	*	-	-	20
*	-	-	-	*	*	20
*	-	*	-	-	-	20
-	-	*	-	*	-	20

Majority Rule Consensus Tree



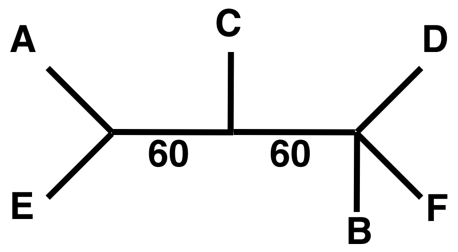
A	B	C	D	E	F	Freq
*	-	-	-	*	-	60
*	-	*	-	*	-	60
-	-	-	*	-	*	60
-	*	*	-	-	-	40
-	*	-	*	-	-	20
*	-	-	-	*	*	20
*	-	*	-	-	-	20
-	-	*	-	*	-	20

Majority Rule Consensus Tree



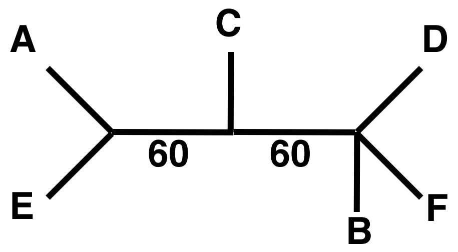
A	B	C	D	E	F	Freq
*	-	-	-	*	-	60
*	-	*	-	*	-	60
-	-	-	*	-	*	60
-	*	*	-	-	-	40
-	*	-	*	-	-	20
*	-	-	-	*	*	20
*	-	*	-	-	-	20
-	-	*	-	*	-	20

Majority Rule Consensus Tree



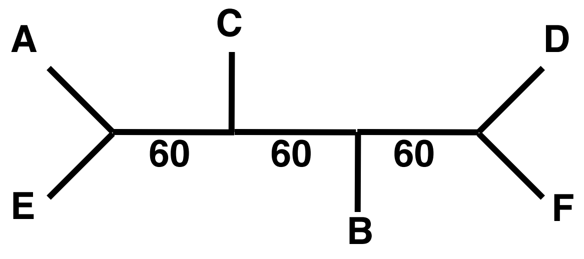
A	B	C	D	E	F	Freq
*	-	-	-	*	-	60
*	-	*	-	*	-	60
-	-	-	*	-	*	60
-	*	*	-	-	-	40
-	*	-	*	-	-	20
*	-	-	-	*	*	20
*	-	*	-	-	-	20
-	-	*	-	*	-	20

Majority Rule Consensus Tree



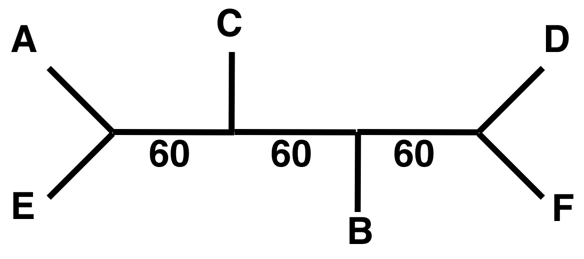
A	B	C	D	E	F	Freq
*	-	-	-	*	-	60
*	-	*	-	*	-	60
-	-	-	*	-	*	60
-	*	*	-	-	-	40
-	*	-	*	-	-	20
*	-	-	-	*	*	20
*	-	*	-	-	-	20
-	-	*	-	*	-	20

Majority Rule Consensus Tree



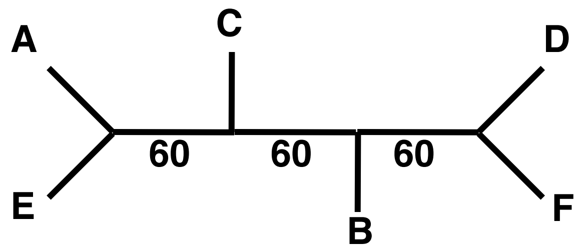
A	B	C	D	E	F	Freq
*	-	-	-	*	-	60
*	-	*	-	*	-	60
-	-	-	*	-	*	60
-	*	*	-	-	-	40
-	*	-	*	-	-	20
*	-	-	-	*	*	20
*	-	*	-	-	-	20
-	-	*	-	*	-	20

Majority Rule Consensus Tree



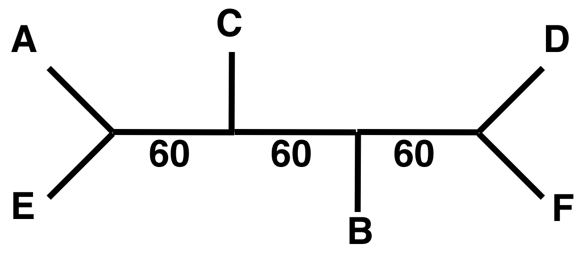
A	B	C	D	E	F	Freq
*	-	-	-	*	-	60
*	-	*	-	*	-	60
-	-	-	*	-	*	60
-	*	*	-	-	-	40
-	*	-	*	-	-	20
*	-	-	-	*	*	20
*	-	*	-	-	-	20
-	-	*	-	*	-	20

Majority Rule Consensus Tree



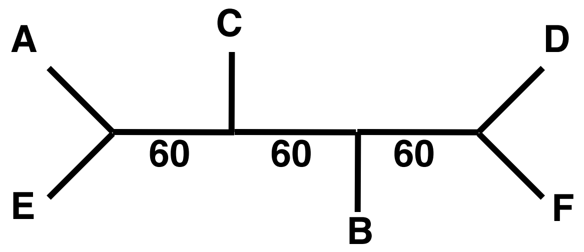
A	B	C	D	E	F	Freq
*	-	-	-	*	-	60
*	-	*	-	*	-	60
-	-	-	*	-	*	60
-	*	*	-	-	-	40
-	*	-	*	-	-	20
*	-	-	-	*	*	20
*	-	*	-	-	-	20
-	-	*	-	*	-	20

Majority Rule Consensus Tree



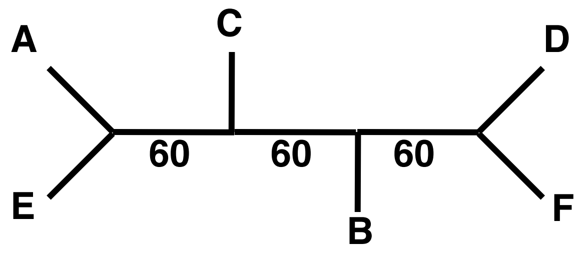
A	B	C	D	E	F	Freq
*	-	-	-	*	-	60
*	-	*	-	*	-	60
-	-	-	*	-	*	60
-	*	*	-	-	-	40
-	*	-	*	-	-	20
*	-	-	-	*	*	20 ÷
*	-	*	-	-	-	20
-	-	*	-	*	-	20

Majority Rule Consensus Tree



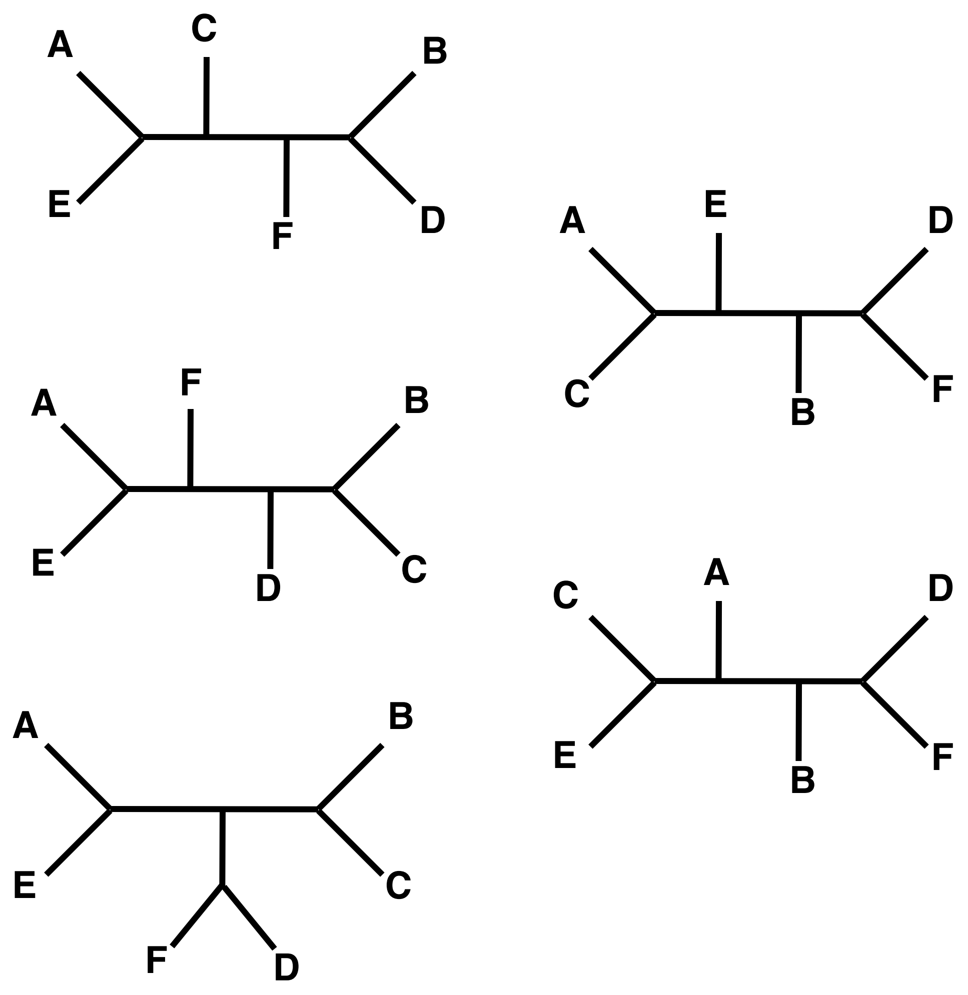
A	B	C	D	E	F	Freq
*	-	-	-	*	-	60
*	-	*	-	*	-	60
-	-	-	*	-	*	60
-	*	*	-	-	-	40
-	*	-	*	-	-	20
*	-	-	-	*	*	20
*	-	*	-	-	-	20 ÷
-	-	*	-	*	-	20

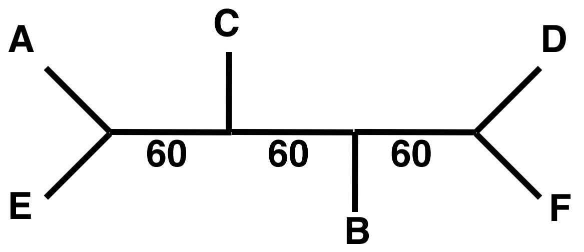
Majority Rule Consensus Tree



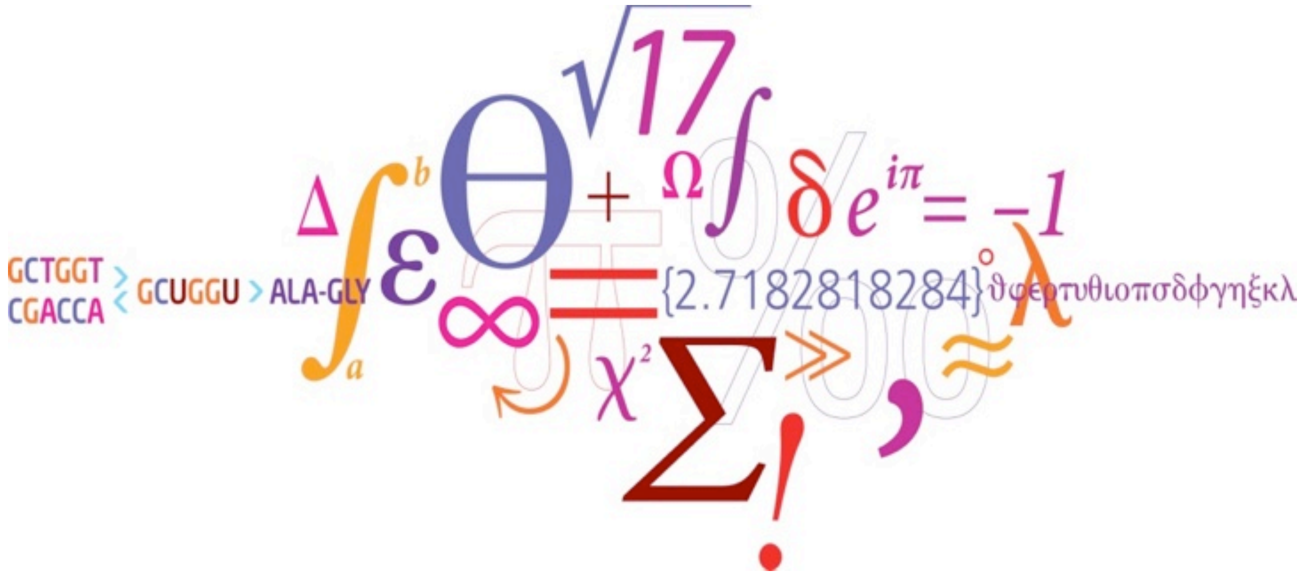
A	B	C	D	E	F	Freq
*	-	-	-	*	-	60
*	-	*	-	*	-	60
-	-	-	*	-	*	60
-	*	*	-	-	-	40
-	*	-	*	-	-	20
*	-	-	-	*	*	20
*	-	*	-	-	-	20
-	-	*	-	*	-	20

Majority Rule Consensus Tree





Distance Matrix Methods



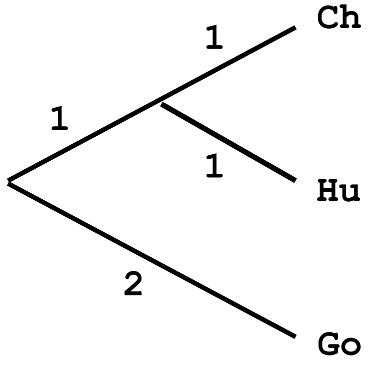
Distance Matrix Methods

```
Gorilla    :  ACGTCGTA
Human      :  ACGTTCCT
Chimpanzee:   ACGTTTCG
```

1. Construct multiple alignment of sequences

	Go	Hu	Ch
Go	-	4	4
Hu		-	2
Ch			-

2. Construct table listing all pairwise differences (distance matrix)

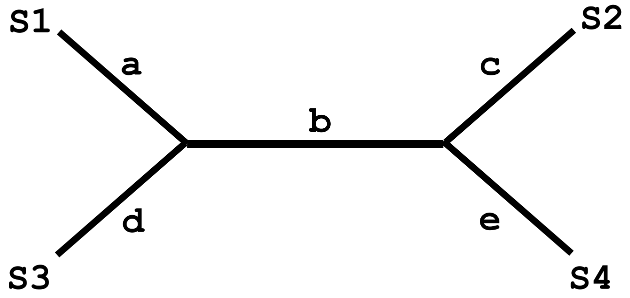


3. Construct tree from pairwise distances

Finding optimal branch lengths

	S_1	S_2	S_3	S_4
S_1	-	D_{12}	D_{13}	D_{14}
S_2		-	D_{23}	D_{24}
S_3			-	D_{34}
S_4				-

Observed distance



Distance along tree
(patristic distance)

Goal:

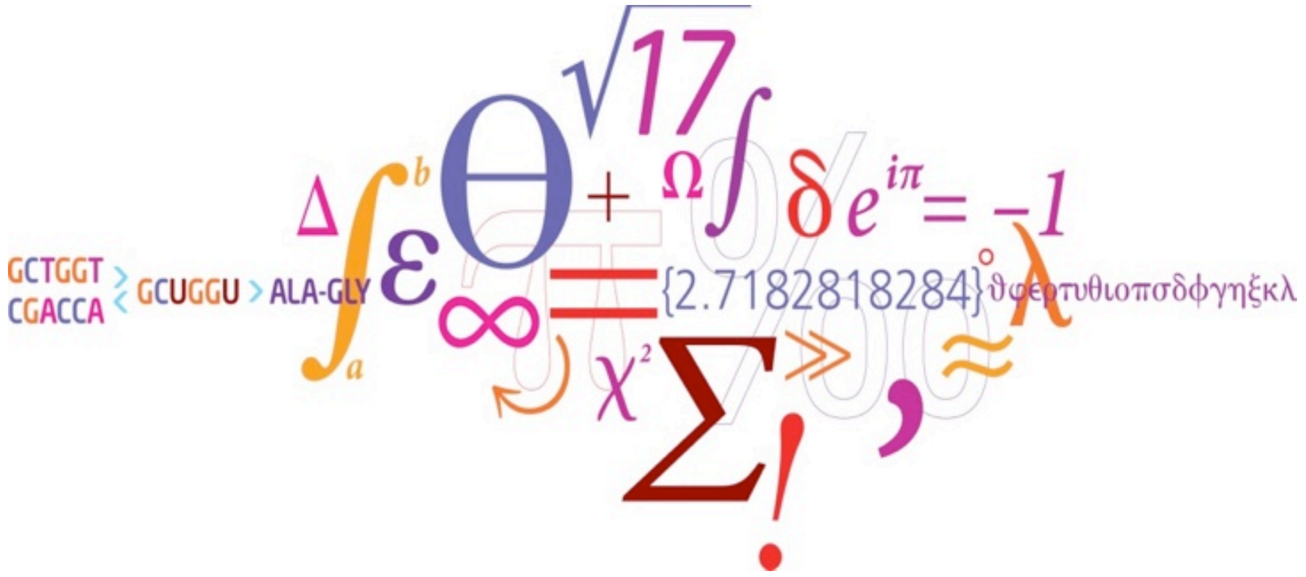
$$
\begin{aligned}
D_{12} &\approx d_{12} = a + b + c \\
D_{13} &\approx d_{13} = a + d \\
D_{14} &\approx d_{14} = a + b + e \\
D_{23} &\approx d_{23} = d + b + c \\
D_{24} &\approx d_{24} = c + e \\
D_{34} &\approx d_{34} = d + b + e
\end{aligned}
$$

Weekly Quiz: Manual Reconstruction of Phylogenetic Tree

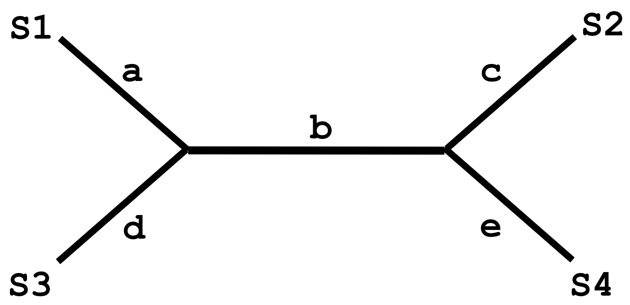
- Construct distance matrix (count different positions)

- Reconstruct tree and find best set of branch lengths

Distance Matrix Methods (continued)



Optimal Branch Lengths for a Given Tree: Least Squares

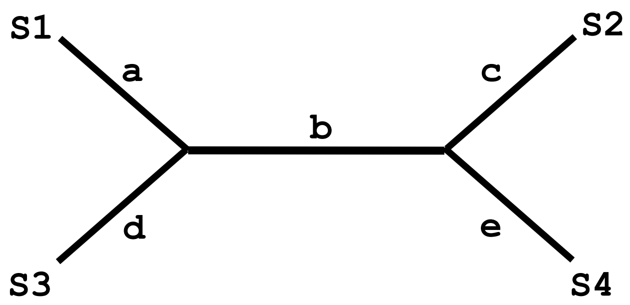


Distance along tree

$$
\begin{aligned}
D_{12} &\approx d_{12} = a + b + c \\
D_{13} &\approx d_{13} = a + d \\
\textbf{Goal:}\quad D_{14} &\approx d_{14} = a + b + e \\
D_{23} &\approx d_{23} = d + b + c \\
D_{24} &\approx d_{24} = c + e \\
D_{34} &\approx d_{34} = d + b + e
\end{aligned}
$$

- Fit between given tree and observed distances can be expressed as "sum of squared differences":
- $$Q = \sum_{j>i} (D_{ij} - d_{ij})^2$$
- Find branch lengths that minimize Q - this is the optimal set of branch lengths for this tree.

Optimal Branch Lengths: Least Squares



Distance along tree

- Longer distances associated with larger errors

- Squared deviation may be weighted so longer branches contribute less to Q:

$$Q = \sum_{j>i} \frac{(D_{ij} - d_{ij})^2}{D_{ij}^{\,n}}$$

- Power (n) is typically 1 or 2

Goal:

$$\begin{aligned}
D_{12} &\approx d_{12} = a + b + c \\
D_{13} &\approx d_{13} = a + d \\
D_{14} &\approx d_{14} = a + b + e \\
D_{23} &\approx d_{23} = d + b + c \\
D_{24} &\approx d_{24} = c + e \\
D_{34} &\approx d_{34} = d + b + e
\end{aligned}$$

Optimal Branch Lengths for a Given Tree: Least Squares Example

```
S1:   TCCGAGTCGATCAGC
S2:   ACCGAGTCGATCTGC
S3:   AAGTACCCGTTGATC
S4:   AAGTTGCCGTTCAGG
```

Multiple alignment

	S_1	S_2	S_3	S_4
S_1	-	2	9	8
S_2		-	9	8
S_3			-	5
S_4				-

Observed distance

Optimal Branch Lengths for a Given Tree: Least Squares Example

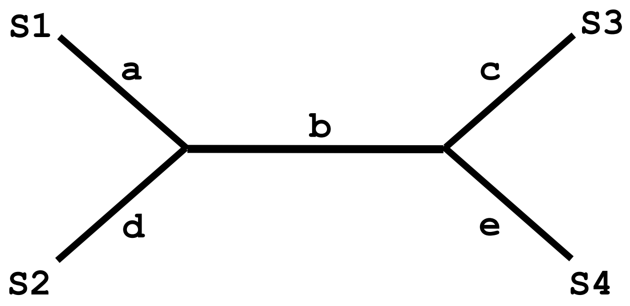
	s_1	s_2	s_3	s_4
s_1	-	2	9	8
s_2		-	9	8
s_3			-	5
s_4				-

Observed distance

Optimal Branch Lengths for a Given Tree: Least Squares Example

	S_1	S_2	S_3	S_4
S_1	-	2	9	8
S_2		-	9	8
S_3			-	5
S_4				-

Observed distance



Distance along tree

d_{12} = a + d
d_{13} = a + b + c
d_{14} = a + b + e
d_{23} = d + b + c
d_{24} = d + b + e
d_{34} = c + e

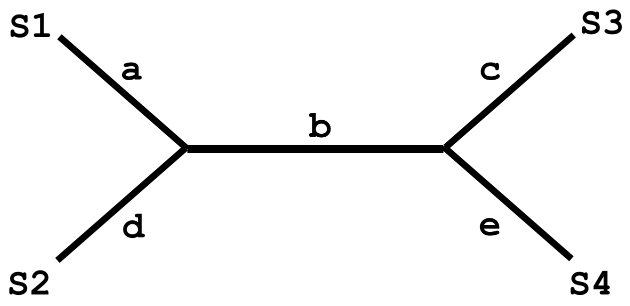
Goal: find branch lengths that minimize Q

$$
\begin{aligned}
Q &= \sum_{i<j}(D_{ij} - d_{ij})^2 \\
&= (D_{12} - d_{12})^2 + (D_{13} - d_{13})^2 + (D_{14} - d_{14})^2 + (D_{23} - d_{23})^2 + (D_{24} - d_{24})^2 + (D_{34} - d_{34})^2
\end{aligned}
$$

Optimal Branch Lengths for a Given Tree: Least Squares Example

	S_1	S_2	S_3	S_4
S_1	-	2	9	8
S_2		-	9	8
S_3			-	5
S_4				-

Observed distance



Distance along tree

$$
\begin{aligned}
d_{12} &= a + d \\
d_{13} &= a + b + c \\
d_{14} &= a + b + e \\
d_{23} &= d + b + c \\
d_{24} &= d + b + e \\
d_{34} &= c + e
\end{aligned}
$$

Goal: find branch lengths that minimize Q

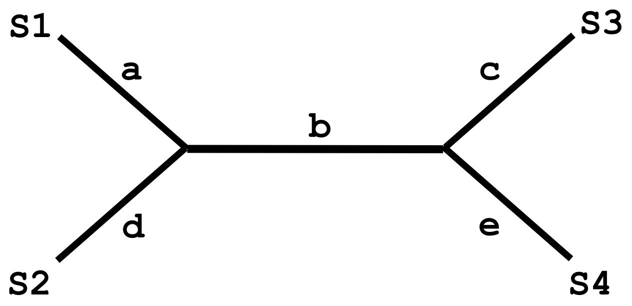
Substitute d-terms with sums of branch lengths

$$
\begin{aligned}
Q &= \sum_{i<j} (D_{ij} - d_{ij})^2 \\
&= (D_{12} - d_{12})^2 + (D_{13} - d_{13})^2 + (D_{14} - d_{14})^2 + (D_{23} - d_{23})^2 + (D_{24} - d_{24})^2 + (D_{34} - d_{34})^2 \\
&= (D_{12} - a - d)^2 + (D_{13} - a - b - c)^2 + (D_{14} - a - b - e)^2 + (D_{23} - d - b - c)^2 + (D_{24} - d - b - e)^2 + (D_{34} - c - e)^2
\end{aligned}
$$

Optimal Branch Lengths for a Given Tree: Least Squares Example

	S_1	S_2	S_3	S_4
S_1	–	2	9	8
S_2		–	9	8
S_3			–	5
S_4				–

Observed distance



Distance along tree

$d_{12} = a + d$
$d_{13} = a + b + c$
$d_{14} = a + b + e$
$d_{23} = d + b + c$
$d_{24} = d + b + e$
$d_{34} = c + e$

Goal: find branch lengths that minimize Q

Substitute D-terms with observed values

$$
\begin{aligned}
Q &= \sum_{i<j} (D_{ij} - d_{ij})^2 \\
&= (D_{12} - d_{12})^2 + (D_{13} - d_{13})^2 + (D_{14} - d_{14})^2 + (D_{23} - d_{23})^2 + (D_{24} - d_{24})^2 + (D_{34} - d_{34})^2 \\
&= (D_{12} - a - d)^2 + (D_{13} - a - b - c)^2 + (D_{14} - a - b - e)^2 + (D_{23} - d - b - c)^2 + (D_{24} - d - b - e)^2 + (D_{34} - c - e)^2 \\
&= (2 - a - d)^2 + (9 - a - b - c)^2 + (8 - a - b - e)^2 + (9 - d - b - c)^2 + (8 - d - b - e)^2 + (5 - c - e)^2
\end{aligned}
$$

Optimal Branch Lengths for a Given Tree: Least Squares Example

$$Q = (2 - a - d)^2 + (9 - a - b - c)^2 + (8 - a - b - e)^2 + (9 - d - b - c)^2 + (8 - d - b - e)^2 + (5 - c - e)^2$$

Transferred from previous slide

Optimal Branch Lengths for a Given Tree: Least Squares Example

$$Q = (2-a-d)^2 + (9-a-b-c)^2 + (8-a-b-e)^2 + (9-d-b-c)^2 + (8-d-b-e)^2 + (5-c-e)^2$$

$$= 319 - 38a - 68b - 42e - 46c - 38d + 2ce + 3a^2 + 2ad + 3d^2 + 4ab + 2ac + 4b^2 + 4bc + 3c^2 + 2ae + 4be + 3e^2 + 4db + 2dc + 2de$$

Expand by multiplying out parentheses

Optimal Branch Lengths for a Given Tree: Least Squares Example

$$Q = (2-a-d)^2 + (9-a-b-c)^2 + (8-a-b-e)^2 + (9-d-b-c)^2 + (8-d-b-e)^2 + (5-c-e)^2$$

$$= 319 - 38a - 68b - 42e - 46c - 38d + 2ce + 3a^2 + 2ad + 3d^2 + 4ab + 2ac + 4b^2 + 4bc + 3c^2 + 2ae + 4be + 3e^2 + 4db + 2dc + 2de$$

$$= 3a^2 + (4b + 2c + 2d - 38 + 2e)a + 319 - 68b - 42e - 46c - 38d + 2ce + 3d^2 + 4b^2 + 4bc + 3c^2 + 4be + 3e^2 + 4db + 2dc + 2de$$

Collect all terms with a

Optimal Branch Lengths for a Given Tree: Least Squares Example

$$Q = (2-a-d)^2 + (9-a-b-c)^2 + (8-a-b-e)^2 + (9-d-b-c)^2 + (8-d-b-e)^2 + (5-c-e)^2$$

$$= 319 - 38a - 68b - 42e - 46c - 38d + 2ce + 3a^2 + 2ad + 3d^2 + 4ab + 2ac + 4b^2 + 4bc + 3c^2 + 2ae + 4be + 3e^2 + 4db + 2dc + 2de$$

$$= 3a^2 + (4b + 2c + 2d - 38 + 2e)a + 319 - 68b - 42e - 46c - 38d + 2ce + 3d^2 + 4b^2 + 4bc + 3c^2 + 4be + 3e^2 + 4db + 2dc + 2de$$

Keep all other branch lengths constant: b, c, d, e

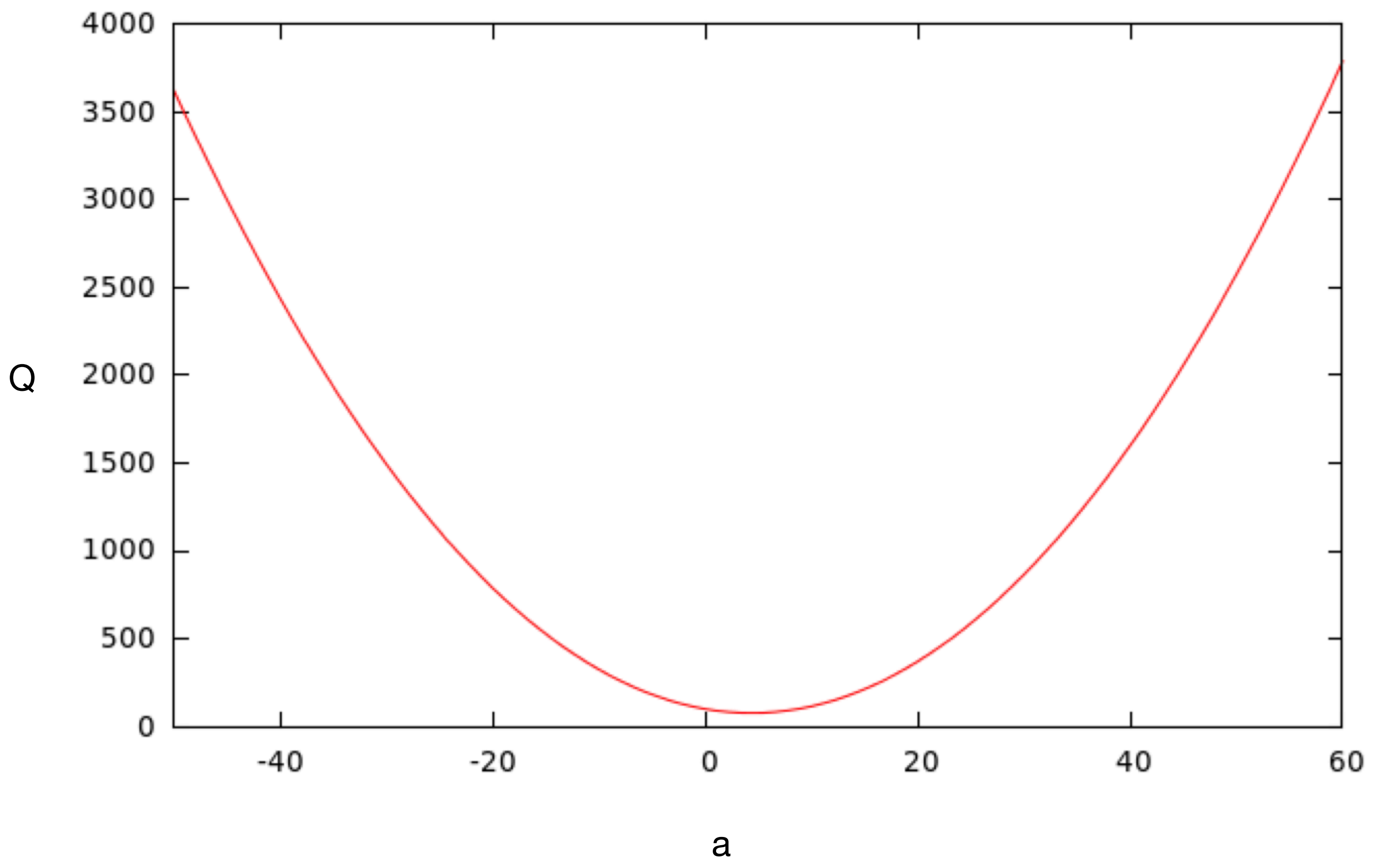
Optimal Branch Lengths for a Given Tree: Least Squares Example

$$Q = (2-a-d)^2 + (9-a-b-c)^2 + (8-a-b-e)^2 + (9-d-b-c)^2 + (8-d-b-e)^2 + (5-c-e)^2$$

$$= 319 - 38a - 68b - 42e - 46c - 38d + 2ce + 3a^2 + 2ad + 3d^2 + 4ab + 2ac + 4b^2 + 4bc + 3c^2 + 2ae + 4be + 3e^2 + 4db + 2dc + 2de$$

$$= 3a^2 + (4b + 2c + 2d - 38 + 2e)a + 319 - 68b - 42e - 46c - 38d + 2ce + 3d^2 + 4b^2 + 4bc + 3c^2 + 4be + 3e^2 + 4db + 2dc + 2de$$

Keep all other branch lengths constant: b, c, d, e



Optimal Branch Lengths for a Given Tree: Least Squares Example

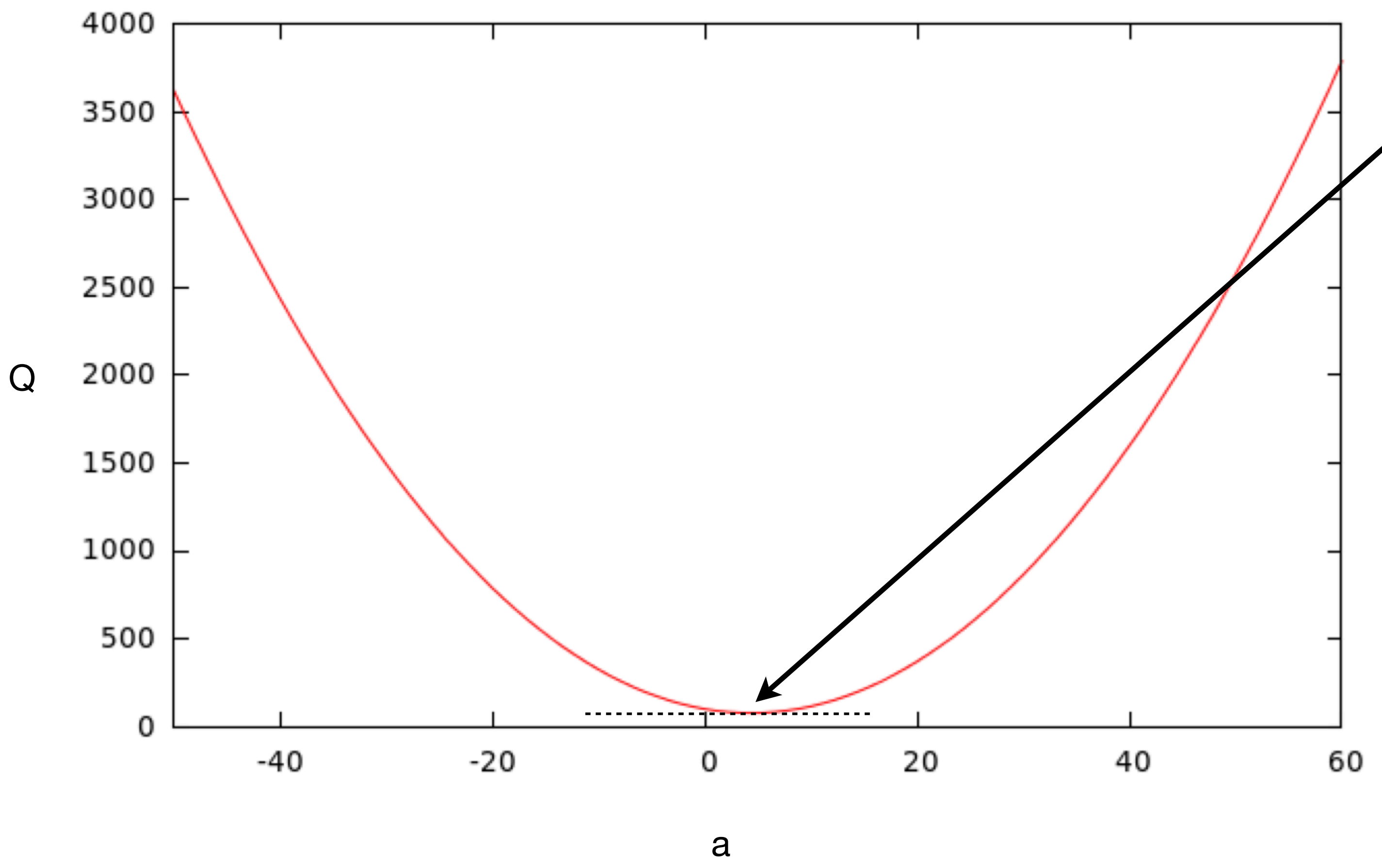
$$Q = (2-a-d)^2 + (9-a-b-c)^2 + (8-a-b-e)^2 + (9-d-b-c)^2 + (8-d-b-e)^2 + (5-c-e)^2$$

$$= 319 - 38a - 68b - 42e - 46c - 38d + 2ce + 3a^2 + 2ad + 3d^2 + 4ab + 2ac + 4b^2 + 4bc + 3c^2 + 2ae + 4be + 3e^2 + 4db + 2dc + 2de$$

$$= 3a^2 + (4b + 2c + 2d - 38 + 2e)a + 319 - 68b - 42e - 46c - 38d + 2ce + 3d^2 + 4b^2 + 4bc + 3c^2 + 4be + 3e^2 + 4db + 2dc + 2de$$

Keep all other branch lengths constant: b, c, d, e

Minimum: $\frac{\partial Q}{\partial a} = 0$



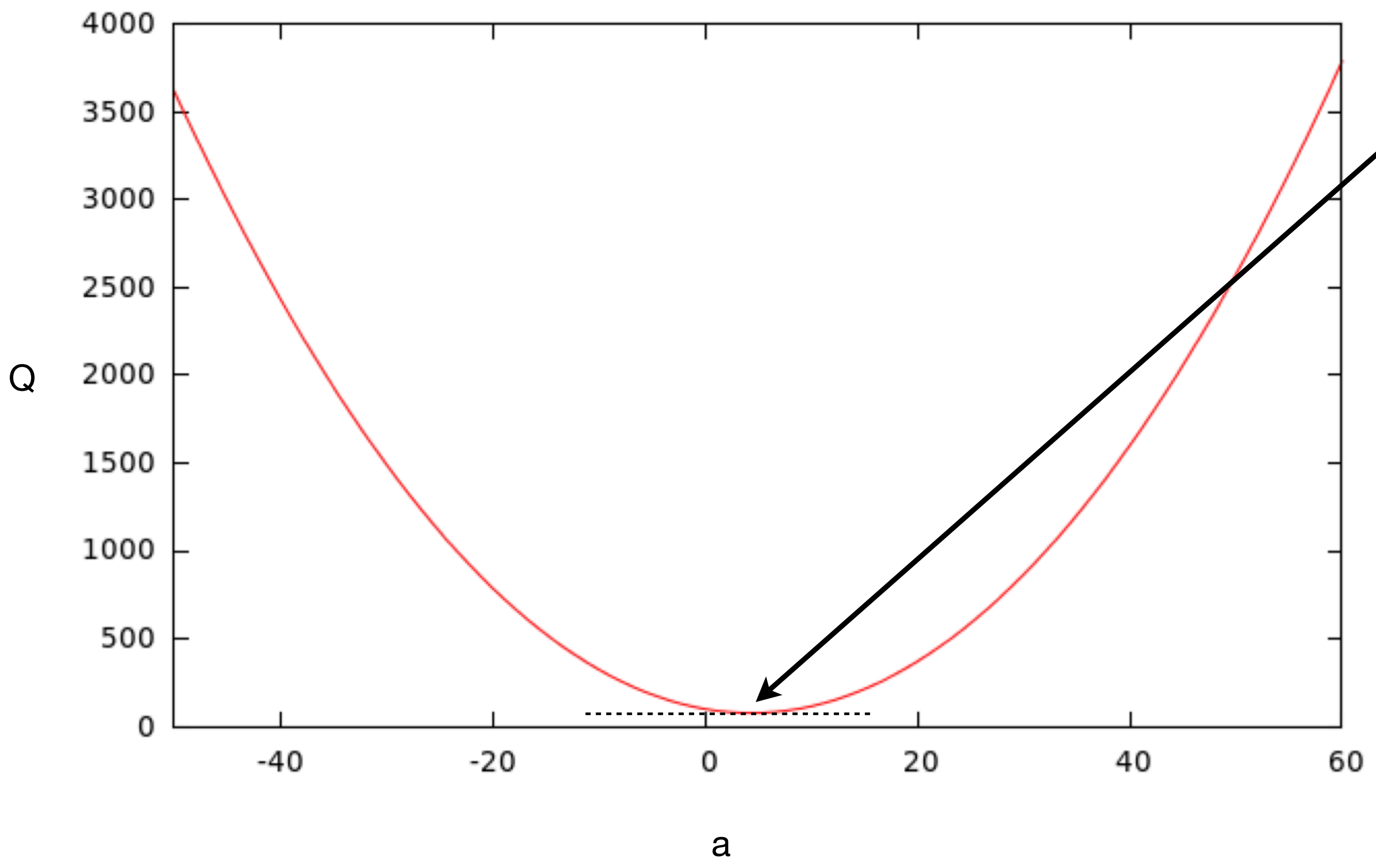
Optimal Branch Lengths for a Given Tree: Least Squares Example

$$Q = (2-a-d)^2 + (9-a-b-c)^2 + (8-a-b-e)^2 + (9-d-b-c)^2 + (8-d-b-e)^2 + (5-c-e)^2$$

$$= 319 - 38a - 68b - 42e - 46c - 38d + 2ce + 3a^2 + 2ad + 3d^2 + 4ab + 2ac + 4b^2 + 4bc + 3c^2 + 2ae + 4be + 3e^2 + 4db + 2dc + 2de$$

$$= 3a^2 + (4b + 2c + 2d - 38 + 2e)a + 319 - 68b - 42e - 46c - 38d + 2ce + 3d^2 + 4b^2 + 4bc + 3c^2 + 4be + 3e^2 + 4db + 2dc + 2de$$

Keep all other branch lengths constant: b, c, d, e



Minimum: $\dfrac{\partial Q}{\partial a} = 0$

$$6a + 4b + 2c + 2d - 38 + 2e = 0$$

Optimal Branch Lengths for a Given Tree: Least Squares Example

$$Q = (2-a-d)^2 + (9-a-b-c)^2 + (8-a-b-e)^2 + (9-d-b-c)^2 + (8-d-b-e)^2 + (5-c-e)^2$$

$$= 319 - 38a - 68b - 42e - 46c - 38d + 2ce + 3a^2 + 2ad + 3d^2 + 4ab + 2ac + 4b^2 + 4bc + 3c^2 + 2ae + 4be + 3e^2 + 4db + 2dc + 2de$$

$$= 3a^2 + (4b + 2c + 2d - 38 + 2e)a + 319 - 68b - 42e - 46c - 38d + 2ce + 3d^2 + 4b^2 + 4bc + 3c^2 + 4be + 3e^2 + 4db + 2dc + 2de$$

$$\frac{\partial Q}{\partial a} = 6a + 4b + 2c + 2d - 38 + 2e = 0$$

$$\frac{\partial Q}{\partial b} = -68 + 4a + 8b + 4c + 4e + 4d = 0$$

$$\frac{\partial Q}{\partial c} = -46 + 2a + 4b + 6c + 2d + 2e = 0$$

$$\frac{\partial Q}{\partial d} = -38 + 2a + 6d + 4b + 2c + 2e = 0$$

$$\frac{\partial Q}{\partial e} = -42 + 2a + 4b + 6e + 2d + 2c = 0$$

- System of 5 linear equations with 5 unknowns
- Can be solved for a, b, c, d, e

Optimal Branch Lengths for a Given Tree: Least Squares Example

$$Q = (2-a-d)^2 + (9-a-b-c)^2 + (8-a-b-e)^2 + (9-d-b-c)^2 + (8-d-b-e)^2 + (5-c-e)^2$$

$$= 319 - 38a - 68b - 42e - 46c - 38d + 2ce + 3a^2 + 2ad + 3d^2 + 4ab + 2ac + 4b^2 + 4bc + 3c^2 + 2ae + 4be + 3e^2 + 4db + 2dc + 2de$$

$$= 3a^2 + (4b + 2c + 2d - 38 + 2e)a + 319 - 68b - 42e - 46c - 38d + 2ce + 3d^2 + 4b^2 + 4bc + 3c^2 + 4be + 3e^2 + 4db + 2dc + 2de$$

$$\frac{\partial Q}{\partial a} = 6a + 4b + 2c + 2d - 38 + 2e = 0$$

$$\frac{\partial Q}{\partial b} = -68 + 4a + 8b + 4c + 4e + 4d = 0$$

$$\frac{\partial Q}{\partial c} = -46 + 2a + 4b + 6c + 2d + 2e = 0 \quad \Rightarrow \quad a = 1, b = 5, c = 3, d = 1, e = 2$$

$$\frac{\partial Q}{\partial d} = -38 + 2a + 6d + 4b + 2c + 2e = 0$$

$$\frac{\partial Q}{\partial e} = -42 + 2a + 4b + 6e + 2d + 2c = 0$$

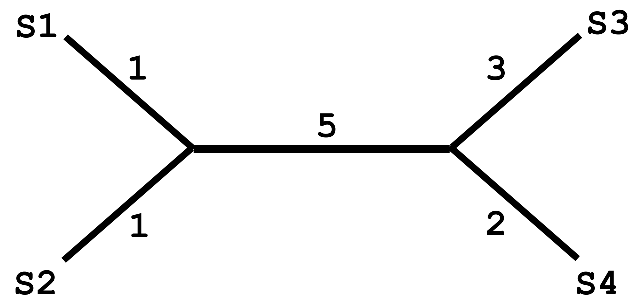
- System of 5 linear equations with 5 unknowns
- Can be solved for a, b, c, d, e

Finding Optimal Branch Lengths

	S_1	S_2	S_3	S_4
S_1	-	2	9	8
S_2		-	9	8
S_3			-	5
S_4				-

Observed distance

$$a = 1, b = 5, c = 3, d = 1, e = 2$$



Distance along tree

Least Squares Optimality Criterion

- Search through all (or many) tree topologies

- For each investigated tree, find best branch lengths using least squares criterion (solve N equations with N unknowns)

- Among all investigated trees, the best tree is the one with the **smallest sum of squared errors**.

- Least squares criterion used both for finding branch lengths on individual trees, and for finding best tree.

Minimum Evolution Optimality Criterion

- Search through all (or many) tree topologies

- For each investigated tree, find best branch lengths using least squares criterion (solve N equations with N unknowns)

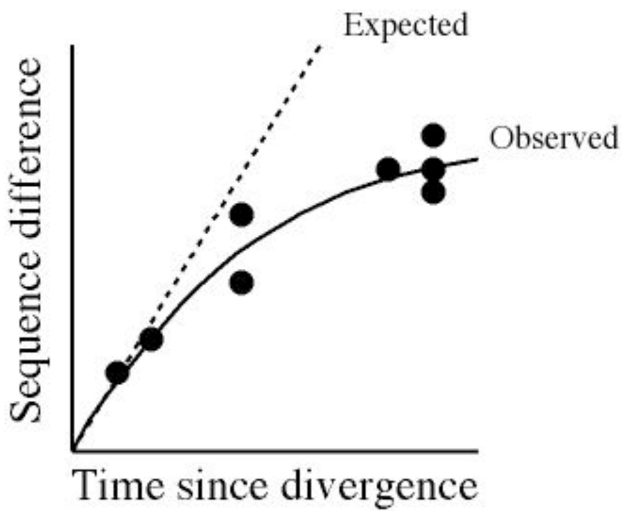
- Among all investigated trees, the best tree is the one with the **smallest sum of branch lengths (the shortest tree)**.

- Least squares criterion used for finding branch lengths on individual trees, minimum tree length used for finding best tree.

Superimposed Substitutions

```
ACGGTGC
↓    ↓|
C    T|
↓    ↓↓
GCGGTGA
```

- Actual number of evolutionary events: 5
- Observed number of differences: 2



- Distance is (almost) always underestimated

Model-based correction for superimposed substitutions

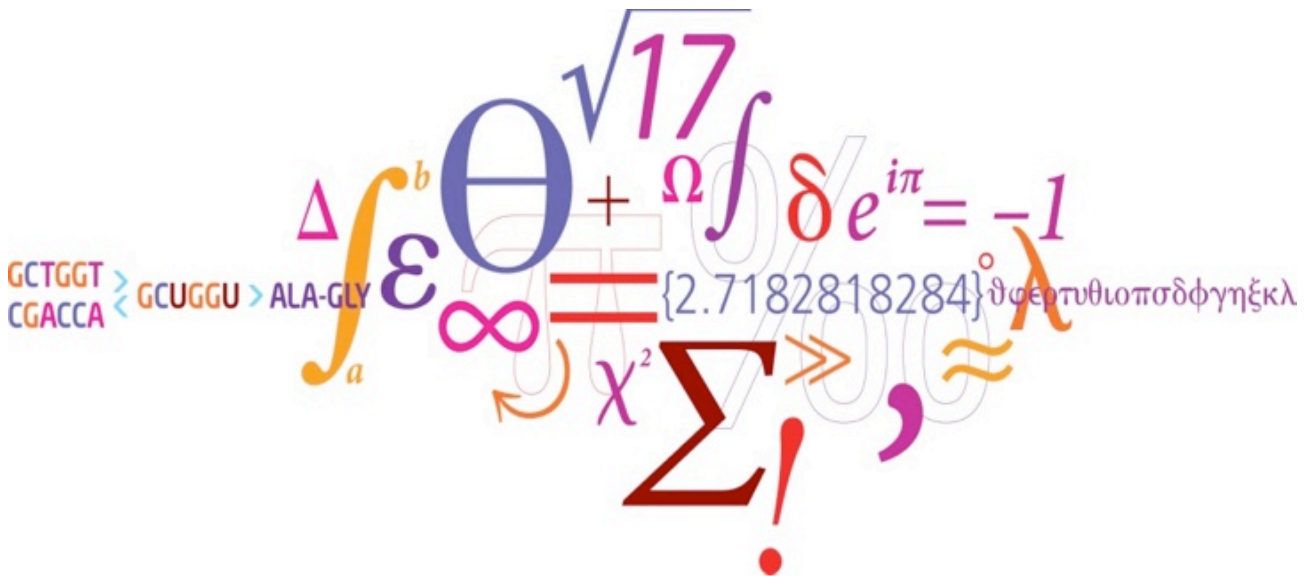
- Goal: try to infer the real number of evolutionary events (the real distance) based on
- Observed data (sequence alignment)
- A model of how evolution occurs

Jukes and Cantor Model

	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

- Four nucleotides assumed to be equally frequent (f=0.25)
- All 12 substitution rates assumed to be equal
- Under this model the corrected distance is: $D_{\mathrm{JC}} = -\frac{3}{4}\ln(1 - \frac{4}{3}D_{\mathrm{OBS}})$
- For instance: $D_{\mathrm{OBS}} = 0.42 \Rightarrow D_{\mathrm{JC}} = 0.62$

Clustering Algorithms: Neighbor Joining



Clustering Algorithms

- Starting point: Distance matrix
- Cluster the two nearest nodes:
    - Tree: connect pair of nodes to common ancestral node, compute branch lengths from ancestral node to both descendants
    - Distance matrix: replace the two joined nodes with the new (ancestral) node. Compute new distance matrix, by finding distance from new node to all other nodes
- Repeat until all nodes are linked in tree
- Results in only one tree, there is no measure of tree-goodness.

Neighbor Joining Algorithm

- For each tip compute $\mathbf{u_i} = \Sigma_j\ \mathbf{D_{ij}/(n-2)}$

  (essentially the average distance to all other tips, except the denominator is n-2 instead of n-1)

- Find the pair of tips, i and j, where $\mathbf{D_{ij}-u_i-u_j}$ is smallest

- Connect the tips i and j, forming a new ancestral node. The branch lengths from the ancestral node to i and j are:

$$v_i = 0.5\ D_{ij} + 0.5\ (u_i-u_j)$$

$$v_j = 0.5\ D_{ij} + 0.5\ (u_j-u_i)$$

- Update the distance matrix: Compute distance between new node and each remaining tip as follows:

$$D_{ij,k} = (D_{ik}+D_{jk}-D_{ij})/2$$

- Replace tips i and j by the new node which is now treated as a tip

- Repeat until only two nodes remain.

Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-

Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-

i	• $u_i = \Sigma_j D_{ij}/(n-2)$
A	(17+21+27)/2=32.5
B	(17+12+18)/2=23.5
C	(21+12+14)/2=23.5
D	(27+18+14)/2=29.5

Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-

i	• $u_i = \Sigma_j D_{ij}/(n-2)$
A	(17+21+27)/2=32.5
B	(17+12+18)/2=23.5
C	(21+12+14)/2=23.5
D	(27+18+14)/2=29.5

	A	B	C	D
A	-	-39	-35	-35
B		-	-35	-35
C			-	-39
D				-

$D_{ij}-u_i-u_j$

Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-

i	• $u_i = \Sigma_j D_{ij}/(n-2)$
A	(17+21+27)/2=32.5
B	(17+12+18)/2=23.5
C	(21+12+14)/2=23.5
D	(27+18+14)/2=29.5

	A	B	C	D
A	-	-39	-35	-35
B		-	-35	-35
C			-	-39
D				-

$D_{ij}-u_i-u_j$

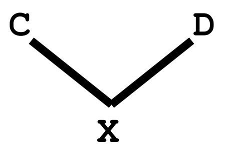
Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-

i	• $u_i = \Sigma_j D_{ij}/(n-2)$
A	(17+21+27)/2=32.5
B	(17+12+18)/2=23.5
C	(21+12+14)/2=23.5
D	(27+18+14)/2=29.5

	A	B	C	D
A	-	-39	-35	-35
B		-	-35	-35
C			-	-39
D				-

$D_{ij}-u_i-u_j$



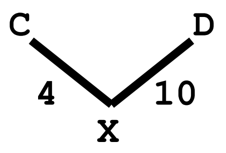
Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-

i	• $u_i = \Sigma_j D_{ij}/(n-2)$
A	(17+21+27)/2=32.5
B	(17+12+18)/2=23.5
C	(21+12+14)/2=23.5
D	(27+18+14)/2=29.5

	A	B	C	D
A	-	-39	-35	-35
B		-	-35	-35
C			-	-39
D				-

$D_{ij}-u_i-u_j$



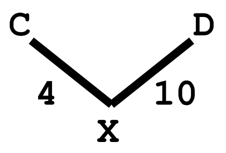
$$v_i = 0.5\ D_{ij} + 0.5\ (u_i-u_j)$$

$$v_j = 0.5\ D_{ij} + 0.5\ (u_j-u_i)$$

$$v_C = 0.5 \times 14 + 0.5 \times (23.5-29.5) = 4$$
$$v_D = 0.5 \times 14 + 0.5 \times (29.5-23.5) = 10$$

Neighbor Joining Algorithm

	A	B	C	D	X
A	-	17	21	27	
B		-	12	18	
C			-	14	
D				-	
X					-

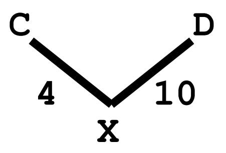


Neighbor Joining Algorithm

	A	B	C	D	X
A	-	17	21	27	
B		-	12	18	
C			-	14	
D				-	
X					-

$$D_{XA} = (D_{CA} + D_{DA} - D_{CD})/2$$
$$= (21 + 27 - 14)/2$$
$$= 17$$

$$D_{XB} = (D_{CB} + D_{DB} - D_{CD})/2$$
$$= (12 + 18 - 14)/2$$
$$= 8$$

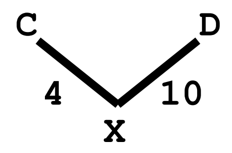


Neighbor Joining Algorithm

	A	B	C	D	X
A	-	17	21	27	17
B		-	12	18	8
C			-	14	
D				-	
X					-

$$
\begin{aligned}
D_{XA} &= (D_{CA} + D_{DA} - D_{CD})/2 \\
&= (21 + 27 - 14)/2 \\
&= 17
\end{aligned}
$$

$$
\begin{aligned}
D_{XB} &= (D_{CB} + D_{DB} - D_{CD})/2 \\
&= (12 + 18 - 14)/2 \\
&= 8
\end{aligned}
$$

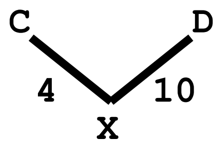


Neighbor Joining Algorithm

	A	B	X
A	-	17	17
B		-	8
X			-

$$D_{XA} = (D_{CA} + D_{DA} - D_{CD})/2$$
$$= (21 + 27 - 14)/2$$
$$= 17$$

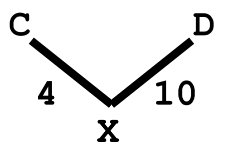
$$D_{XB} = (D_{CB} + D_{DB} - D_{CD})/2$$
$$= (12 + 18 - 14)/2$$
$$= 8$$



Neighbor Joining Algorithm

	A	B	X
A	-	17	17
B		-	8
X			-

i	• $u_i = \Sigma_j D_{ij} / (n-2)$
A	(17+17)/1 = 34
B	(17+8)/1 = 25
X	(17+8)/1 = 25



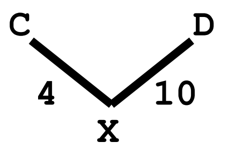
Neighbor Joining Algorithm

	A	B	X
A	-	17	17
B		-	8
X			-

i	• $u_i = \Sigma_j D_{ij}/(n-2)$
A	(17+17)/1 = 34
B	(17+8)/1 = 25
X	(17+8)/1 = 25

	A	B	X
A	-	-42	-28
B		-	-28
X			-

$D_{ij}-u_i-u_j$



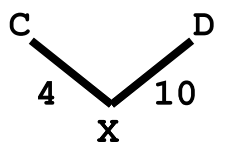
Neighbor Joining Algorithm

	A	B	X
A	-	17	17
B		-	8
X			-

i	• $u_i = \Sigma_j D_{ij}/(n-2)$
A	(17+17)/1 = 34
B	(17+8)/1 = 25
X	(17+8)/1 = 25

	A	B	X
A	-	-42	-28
B		-	-28
X			-

$D_{ij}-u_i-u_j$



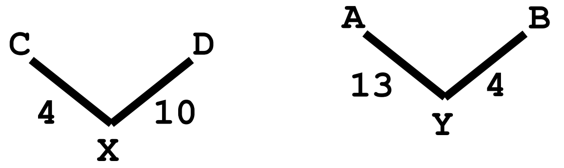
Neighbor Joining Algorithm

	A	B	X
A	-	17	17
B		-	8
X			-

i	• $u_i = \Sigma_j D_{ij}/(n-2)$
A	(17+17)/1 = 34
B	(17+8)/1 = 25
X	(17+8)/1 = 25

	A	B	X
A	-	-42	-28
B		-	-28
X			-

$D_{ij}-u_i-u_j$

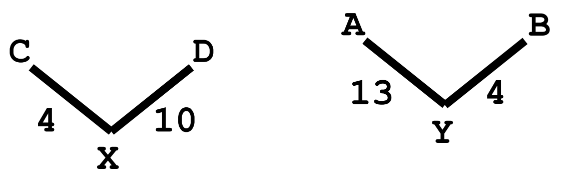


$v_A = 0.5 \times 17 + 0.5 \times (34-25) = 13$

$v_D = 0.5 \times 17 + 0.5 \times (25-34) = 4$

Neighbor Joining Algorithm

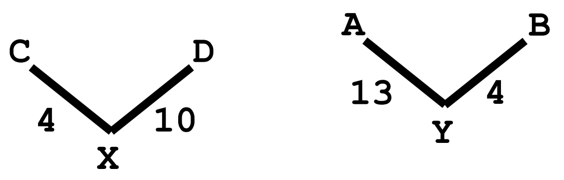
	A	B	X	Y
A	-	17	17	
B		-	8	
X			-	
Y				



Neighbor Joining Algorithm

	A	B	X	Y
A	-	17	17	
B		-	8	
X			-	4
Y				

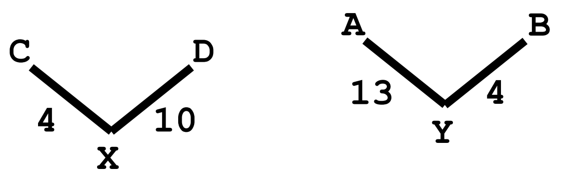
$$D_{YX} = (D_{AX} + D_{BX} - D_{AB})/2$$
$$= (17 + 8 - 17)/2$$
$$= 4$$



Neighbor Joining Algorithm

	X	Y
X	-	4
Y		-

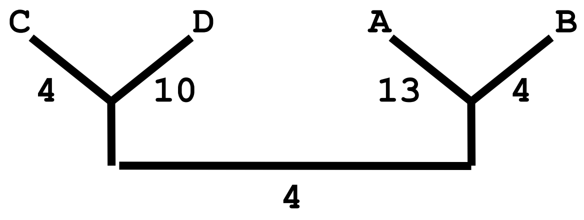
$$D_{YX} = (D_{AX} + D_{BX} - D_{AB})/2$$
$$= (17 + 8 - 17)/2$$
$$= 4$$



Neighbor Joining Algorithm

	X	Y
X	-	4
Y		-

$$D_{YX} = (D_{AX} + D_{BX} - D_{AB})/2$$
$$= (17 + 8 - 17)/2$$
$$= 4$$



Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-