

# Bio334

## Phylogenetic reconstruction Part I: Distance-based methods

Janko Tackmann & Lukas Malfertheiner

For questions, feel free to contact us on Slack or via:

[janko.tackmann@imls.uzh.ch](mailto:janko.tackmann@imls.uzh.ch) or  
[Lukas.malfertheiner@uzh.ch](mailto:Lukas.malfertheiner@uzh.ch)

# Exercise 1 – Questions

- Complexity

# Exercise 1 – Questions

- Complexity

$$\binom{n}{2} = \frac{n!}{2! (n-2)!} = \frac{n(n-1)}{2} \sim O(n^2)$$

Time: 1000 proteins  $\rightarrow$  1 s

10.000 proteins  $\rightarrow$  100 s

100.000 proteins  $\rightarrow$  10.000s ( $\sim$  3 hours)

# Exercise 1 – Questions

- Shortcomings of naïve Jaccard

# Exercise 1 – Questions

- Shortcomings of naïve Jaccard
  - All amino acid transitions have equal probability:  
unrealistic
    - Use explicit substitution matrices (BLOSUM, PAM)

# PAM Substitution Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-3	-1	-1	-3	-2	0	1	-3	-2	-3	-3	-2	-5	1	1	1	-7	-4	0
R	-3	7	-2	-4	-5	1	-3	-5	1	-3	-5	2	-1	-6	-1	-1	-3	1	-6	-4
N	-1	-2	5	3	-5	-1	1	-1	2	-3	-4	1	-4	-5	-2	1	0	-5	-2	-3
D	-1	-4	3	5	-7	0	4	-1	-1	-4	-6	-1	-5	-8	-3	-1	-2	-9	-6	-4
C	-3	-5	-5	-7	9	-8	-8	-5	-4	-3	-8	-8	-7	-7	-4	-1	-4	-9	-1	-3
Q	-2	1	-1	0	-8	6	2	-3	3	-4	-2	0	-2	-7	-1	-2	-2	-7	-6	-3
E	0	-3	1	4	-8	2	5	-1	-1	-3	-5	-1	-4	-8	-2	-1	-2	-9	-5	-3
G	1	-5	-1	-1	-5	-3	-1	5	-4	-5	-6	-3	-4	-6	-2	0	-2	-9	-7	-3
H	-3	1	2	-1	-4	3	-1	-4	7	-4	-3	-2	-4	-3	-1	-2	-3	-4	-1	-3
I	-2	-3	-3	-4	-3	-4	-3	-5	-4	6	1	-3	1	0	-4	-3	0	-7	-3	3
L	-3	-5	-4	-6	-8	-2	-5	-6	-3	1	6	-4	3	0	-4	-4	-3	-3	-3	0
K	-3	2	1	-1	-8	0	-1	-3	-2	-3	-4	5	0	-7	-3	-1	-1	-6	-6	-4
M	-2	-1	-4	-5	-7	-2	-4	-4	-4	1	3	0	9	-1	-4	-3	-1	-6	-5	1
F	-5	-6	-5	-8	-7	-7	-8	-6	-3	0	0	-7	-1	8	-6	-4	-5	-1	4	-3
P	1	-1	-2	-3	-4	-1	-2	-2	-1	-4	-4	-3	-4	-6	7	0	-1	-7	-7	-3
S	1	-1	1	-1	-1	-2	-1	0	-2	-3	-4	-1	-3	-4	0	4	2	-3	-4	-2
T	1	-3	0	-2	-4	-2	-2	-2	-3	0	-3	-1	-1	-5	-1	2	5	-7	-4	0
W	-7	1	-5	-9	-9	-7	-9	-9	-4	-7	-3	-6	-6	-1	-7	-3	-7	12	-2	-9
Y	-4	-6	-2	-6	-1	-6	-5	-7	-1	-3	-3	-6	-5	4	-7	-4	-4	-2	9	-4
V	0	-4	-3	-4	-3	-3	-3	-3	-3	3	0	-4	1	-3	-3	-2	0	-9	-4	5



# Exercise 1 – Questions

- Shortcomings of naïve Jaccard
  - All amino acid transitions have equal probability: unrealistic
    - Use explicit substitution matrices (BLOSUM, PAM)
  - Dependencies between neighboring sites (e.g. indels) are not considered
    - Smaller gap penalties, explicit modeling

AGTG-----ACTATAAT---CG---GAGGACAG--  
ATTCTGT---CCTATAAT---CG---GAGAAAAGCC  
AGTCTGT---ACTATAATGTTGG---GAGGAAAAGC  
AGTCCGTTGC--TATAAT---GG---GAGGAAAACC  
AATCTGT---AGTATAAT---GGTGTGAGGAAAGCC



# Exercise 1 – Questions

- Nucleotides vs Amino acids

# Exercise 1 – Questions

- Nucleotides vs Amino acids
  - Nucleotides
    - High resolution: small differences can be detected
    - More neutrally evolving regions: less biased similarity estimates
    - Distinction between synonymous and non-synonymous mutations

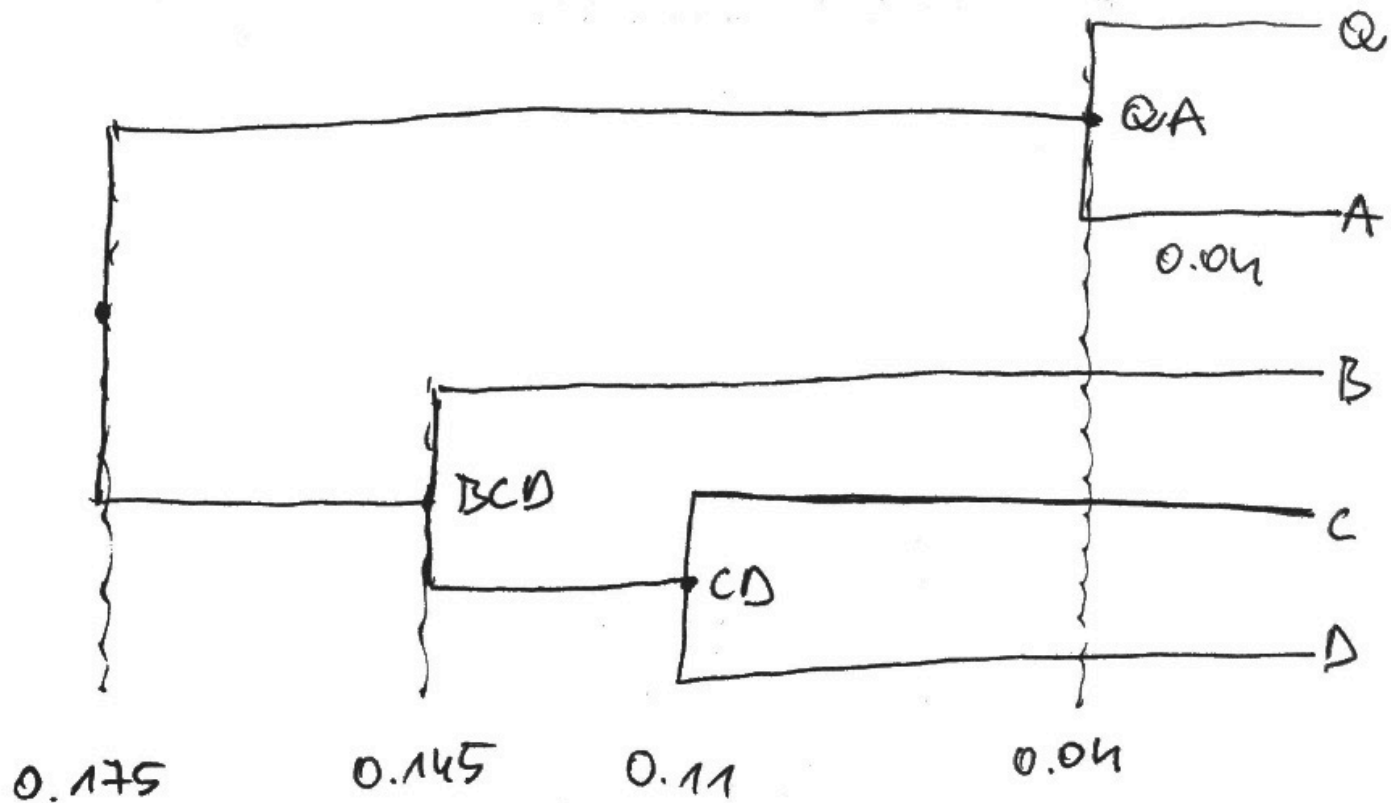
# Exercise 1 – Questions

- Nucleotides vs Amino acids
  - Nucleotides
    - High resolution: small differences can be detected
    - More neutrally evolving regions: less biased similarity estimates
    - Distinction between synonymous and non-synonymous mutations
  - Amino acids
    - Stable over longer evolutionary time frames, used to address questions on non-recent evolution

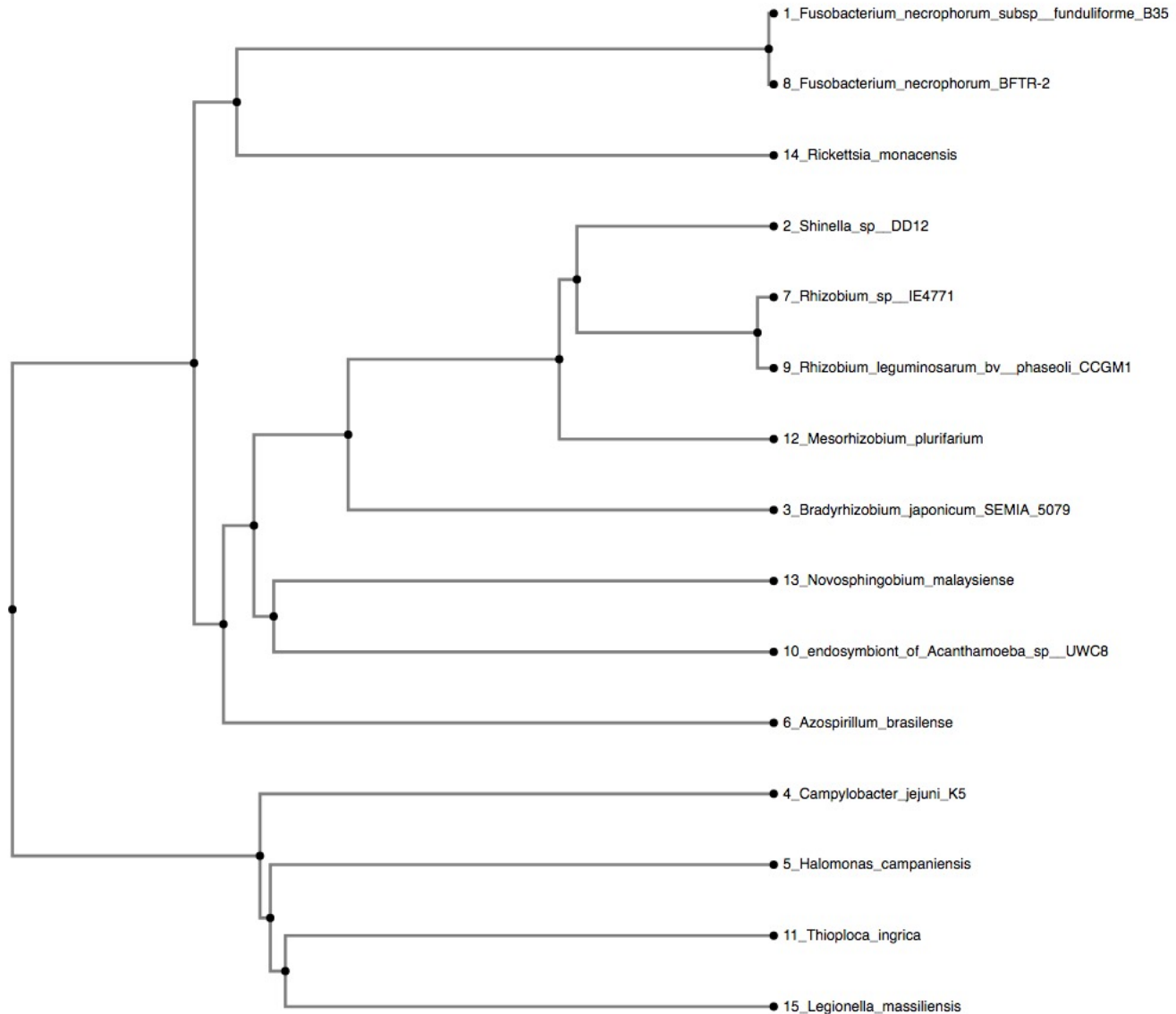
# Exercise 2 – Questions

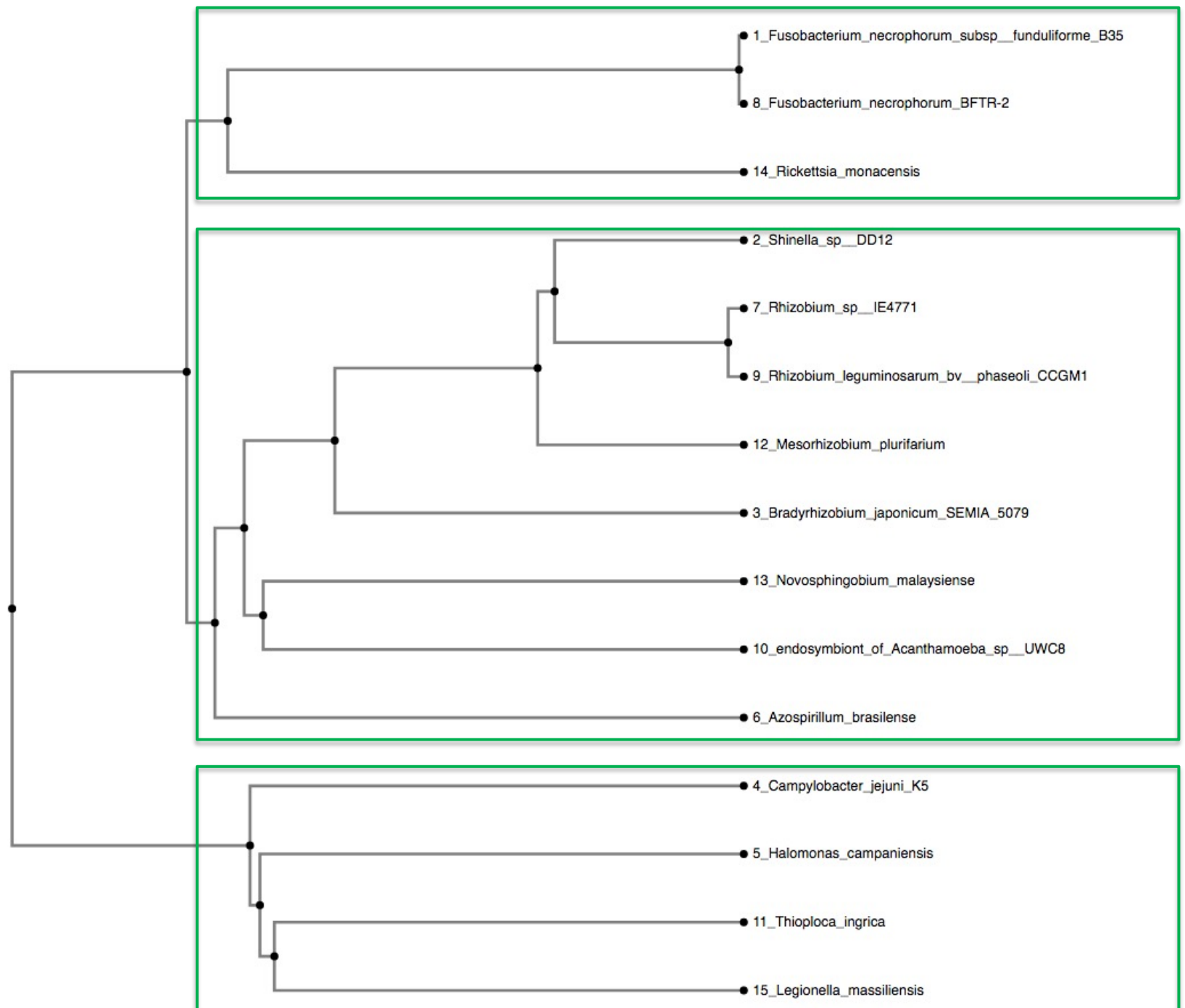
[illegible]

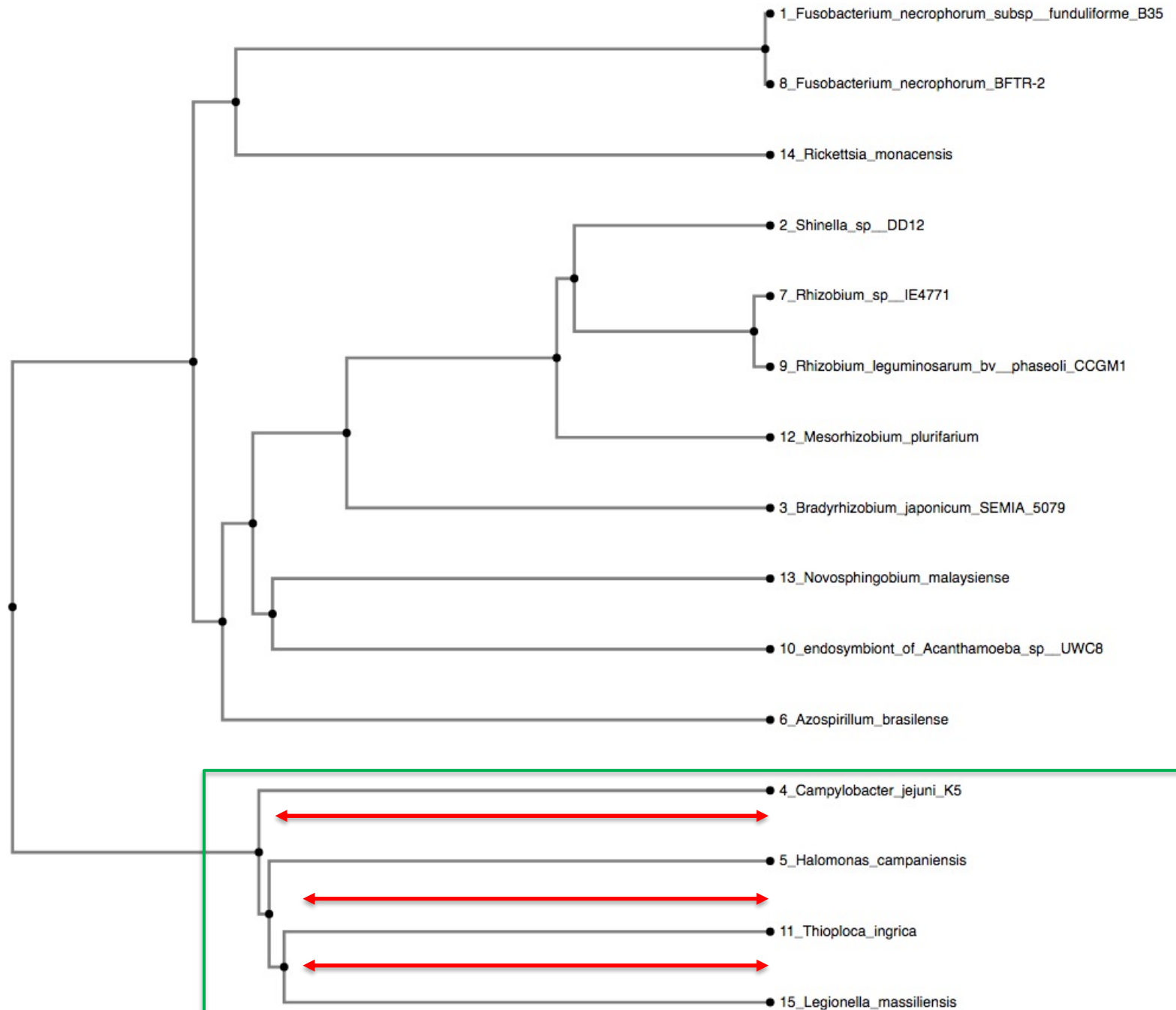
QABCD



# Exercise 3 – Questions









**Fusobacterium necrophorum subsp. funduliforme B35**

**Fusobacterium**

Genus

**necrophorum**

Species

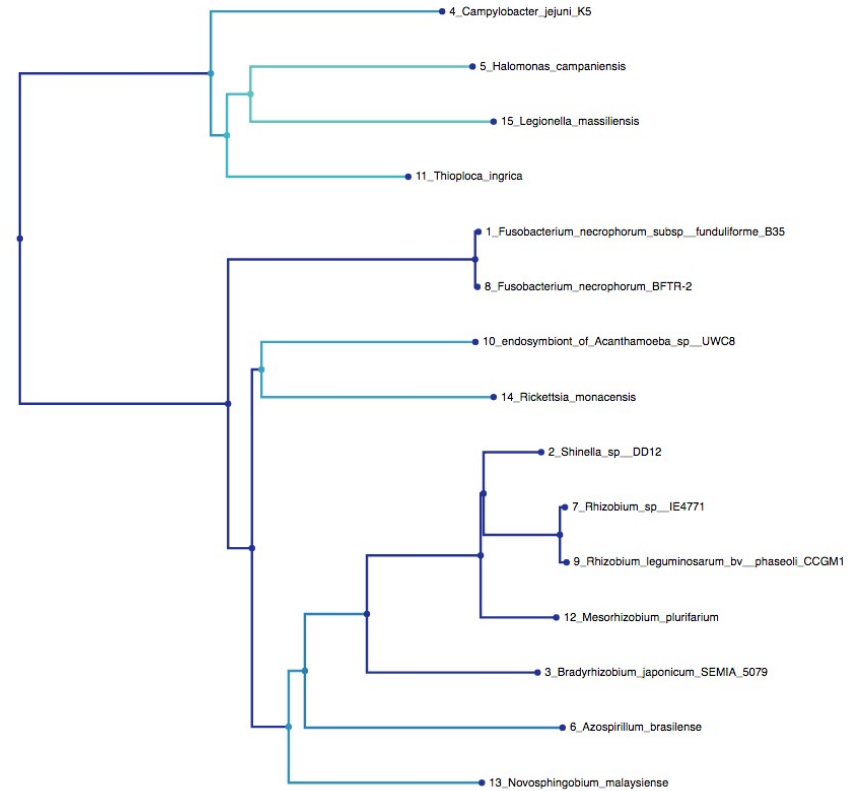
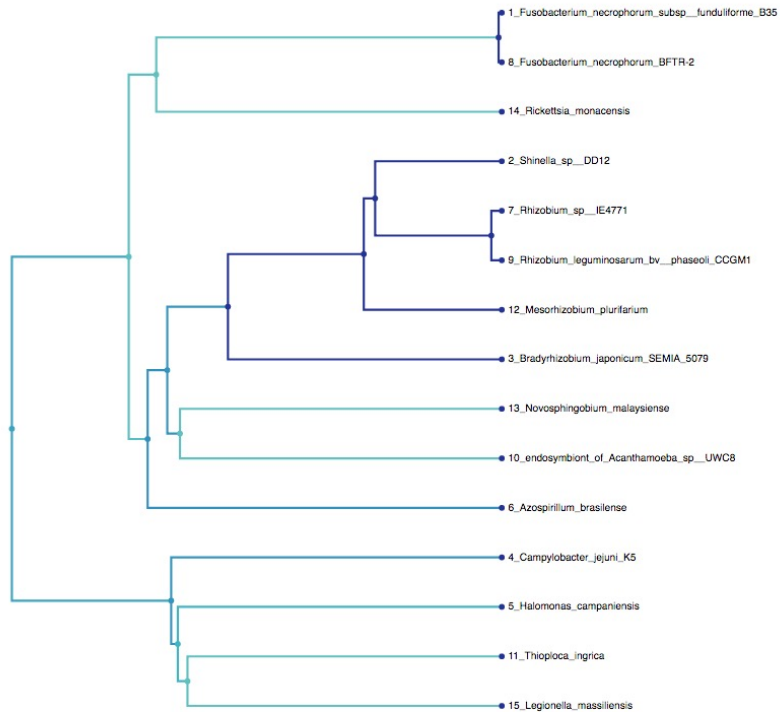
**subsp. funduliforme**

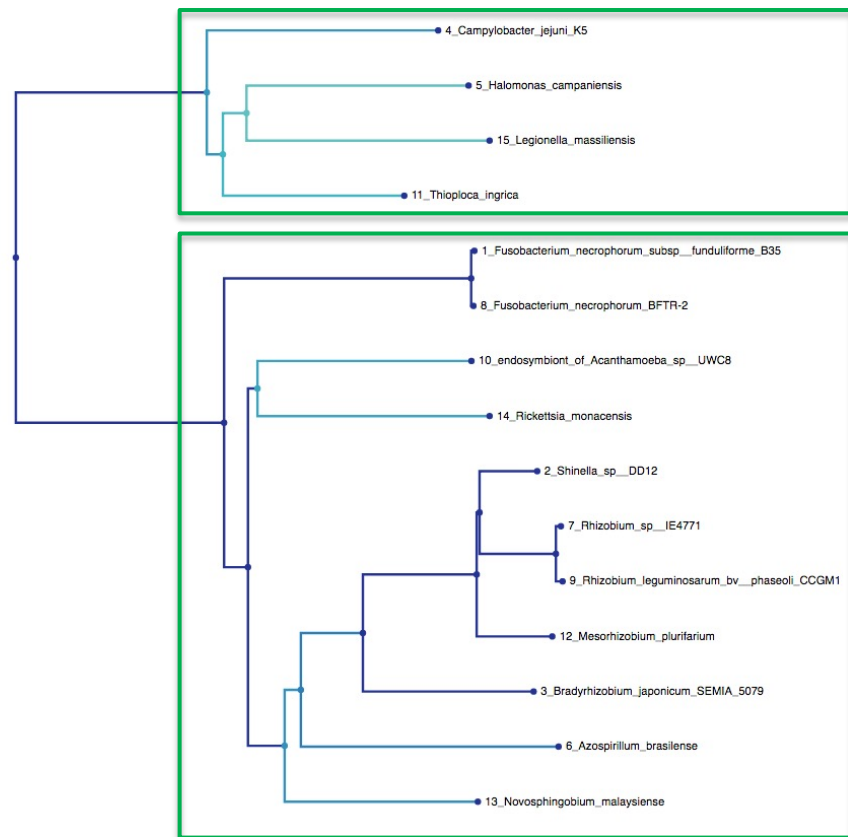
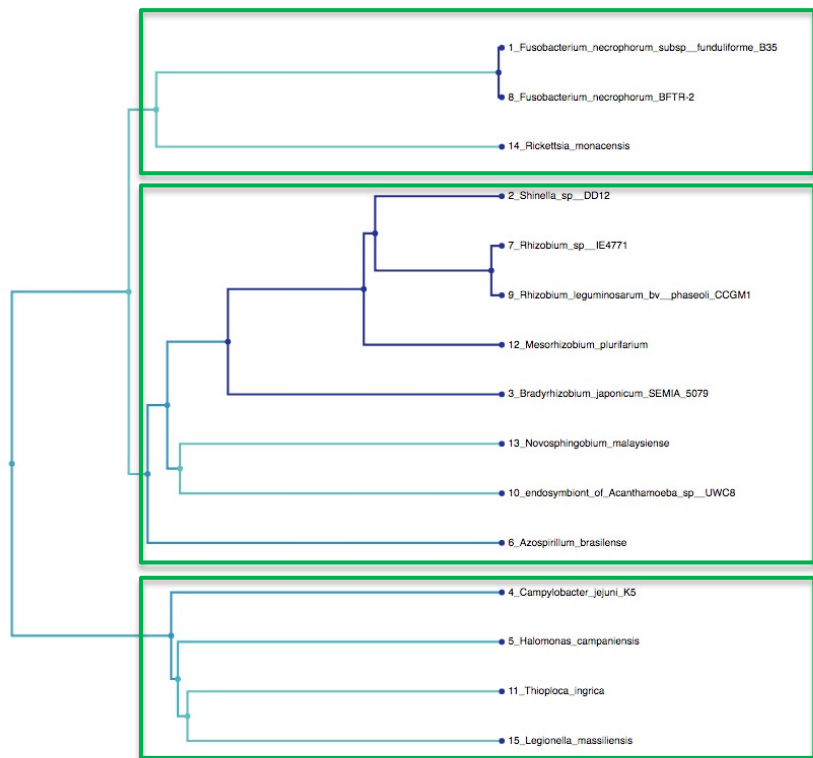
Subspecies

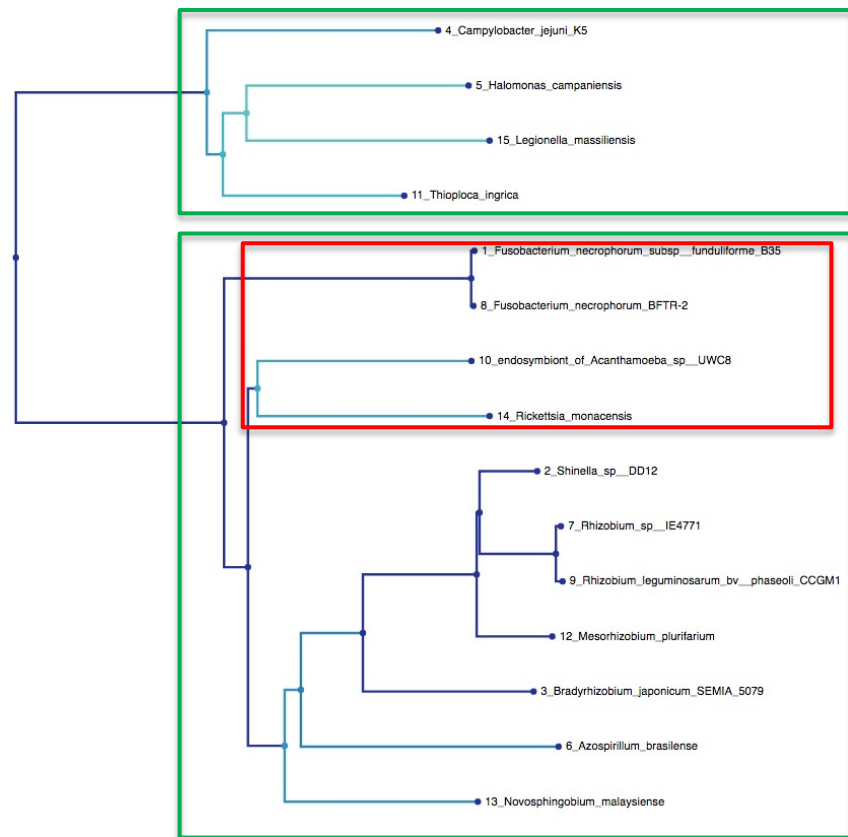
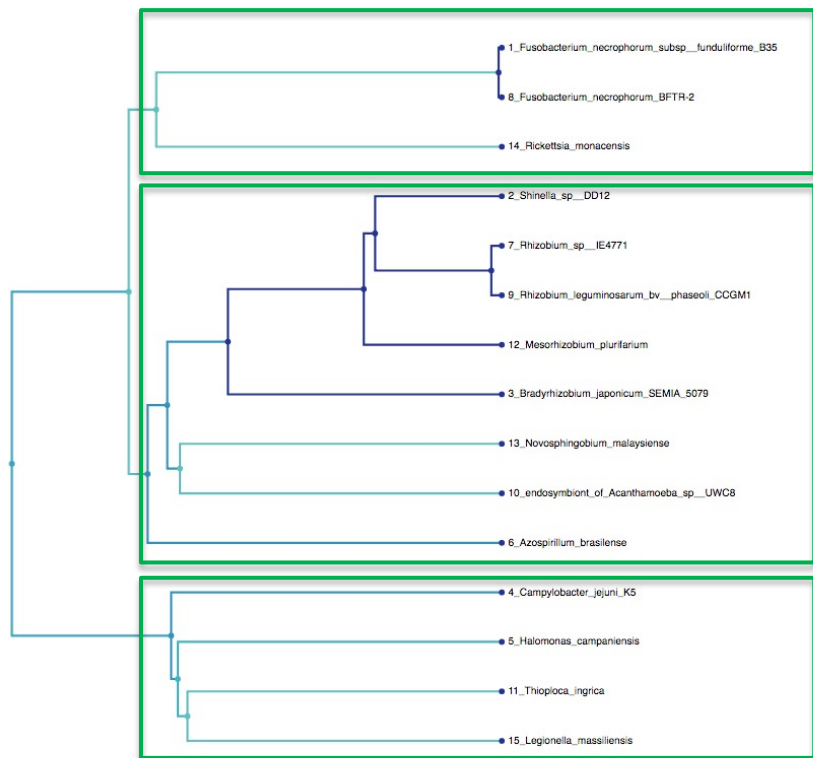
**B35**

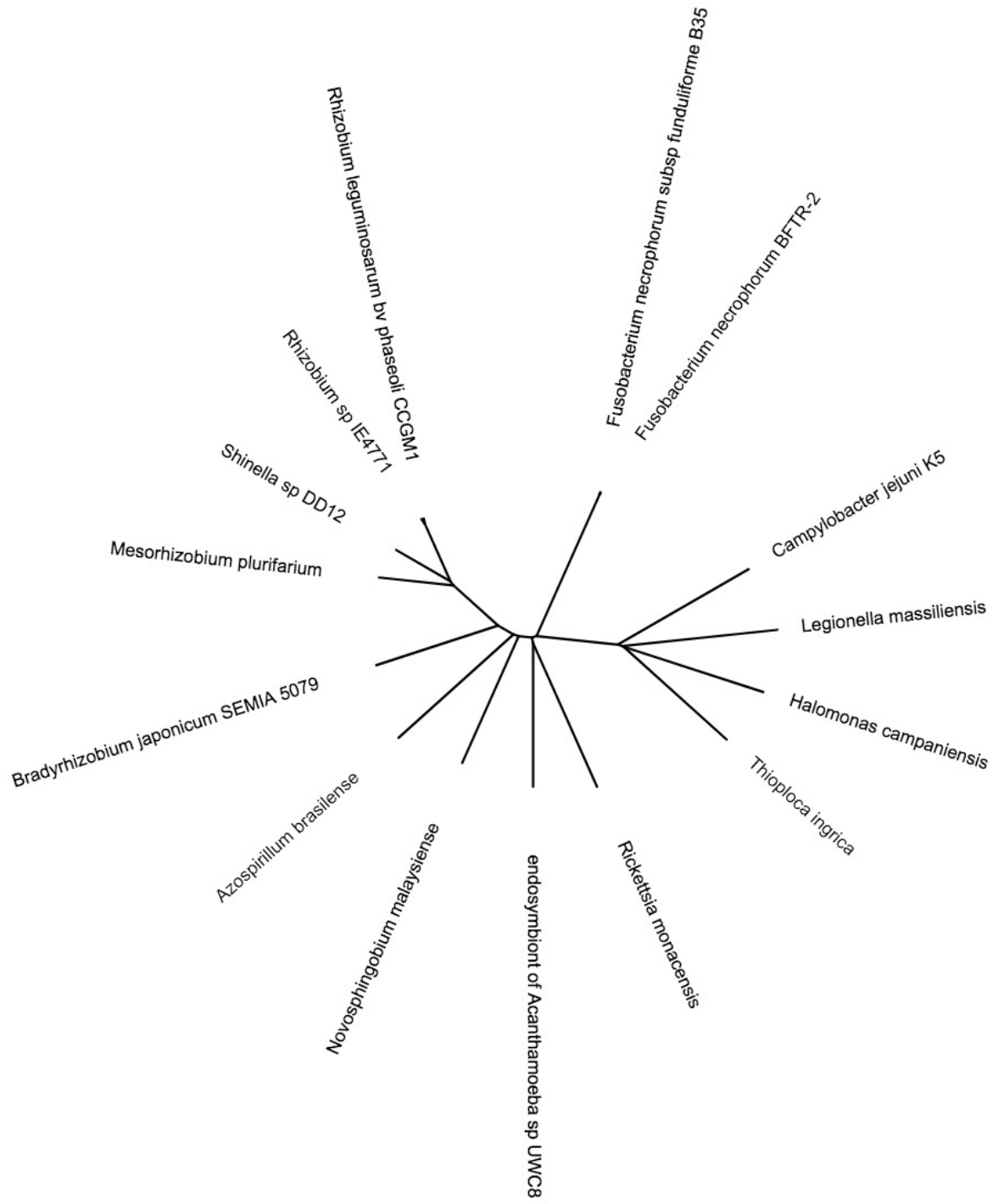
Type strain

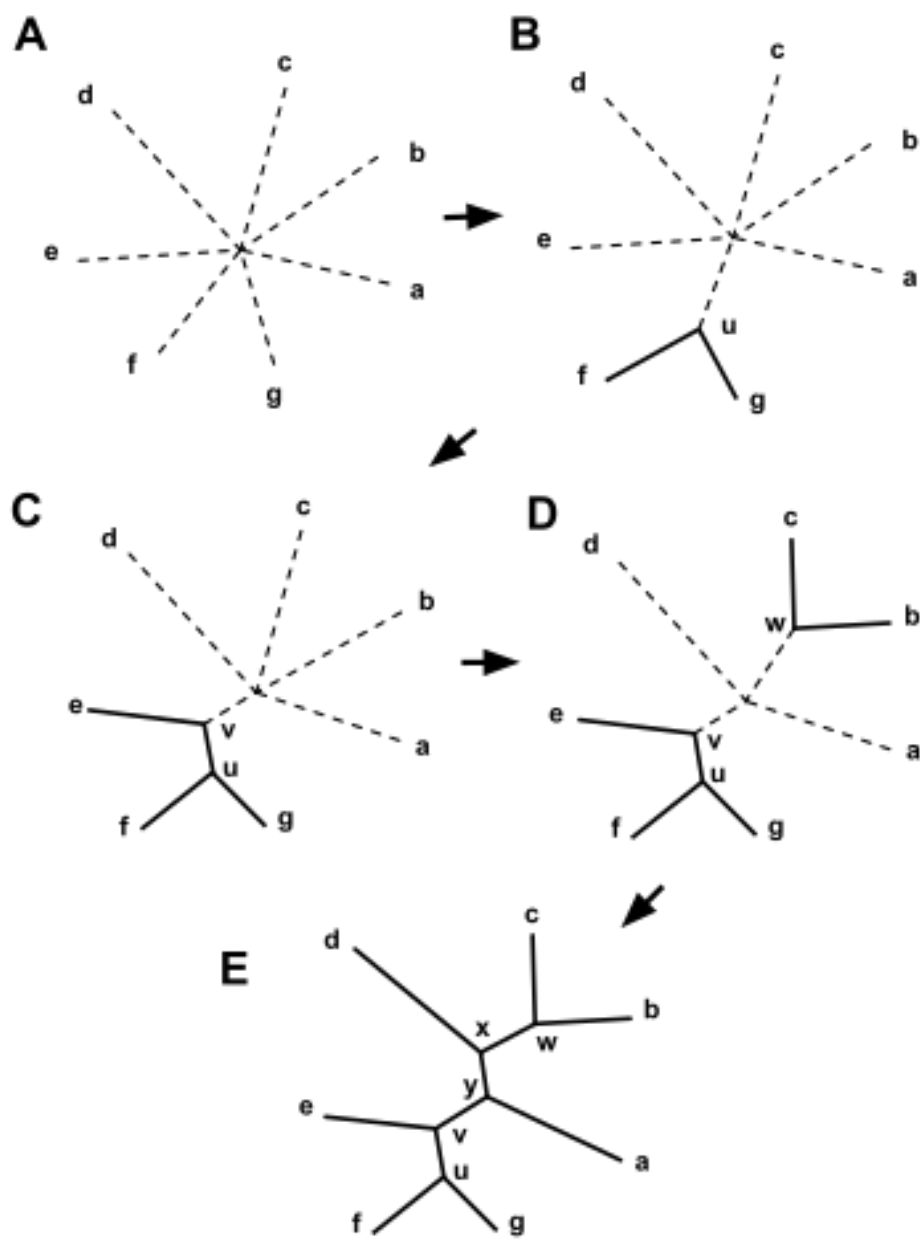
# Exercise 4 – Questions











$$Q(i, j) =$$

$$(n - 2)d(i, j) - \sum_{k=1}^n d(i, k) - \sum_{k=1}^n d(j, k)$$