Bio334 – Building Maximum Likelihood Trees

Solutions to the exercises

Marija Dmitrijeva and João Matias Rodrigues

For remaining doubts feel free to contact us on Slack or at:
marija.dmitrijeva@uzh.ch or joao.rodrigues@imls.uzh.ch

Exercise 1 – Part 1: The grammar

Is the tree specified as rooted or unrooted?

((A,(B,(C,D))),(E,F));

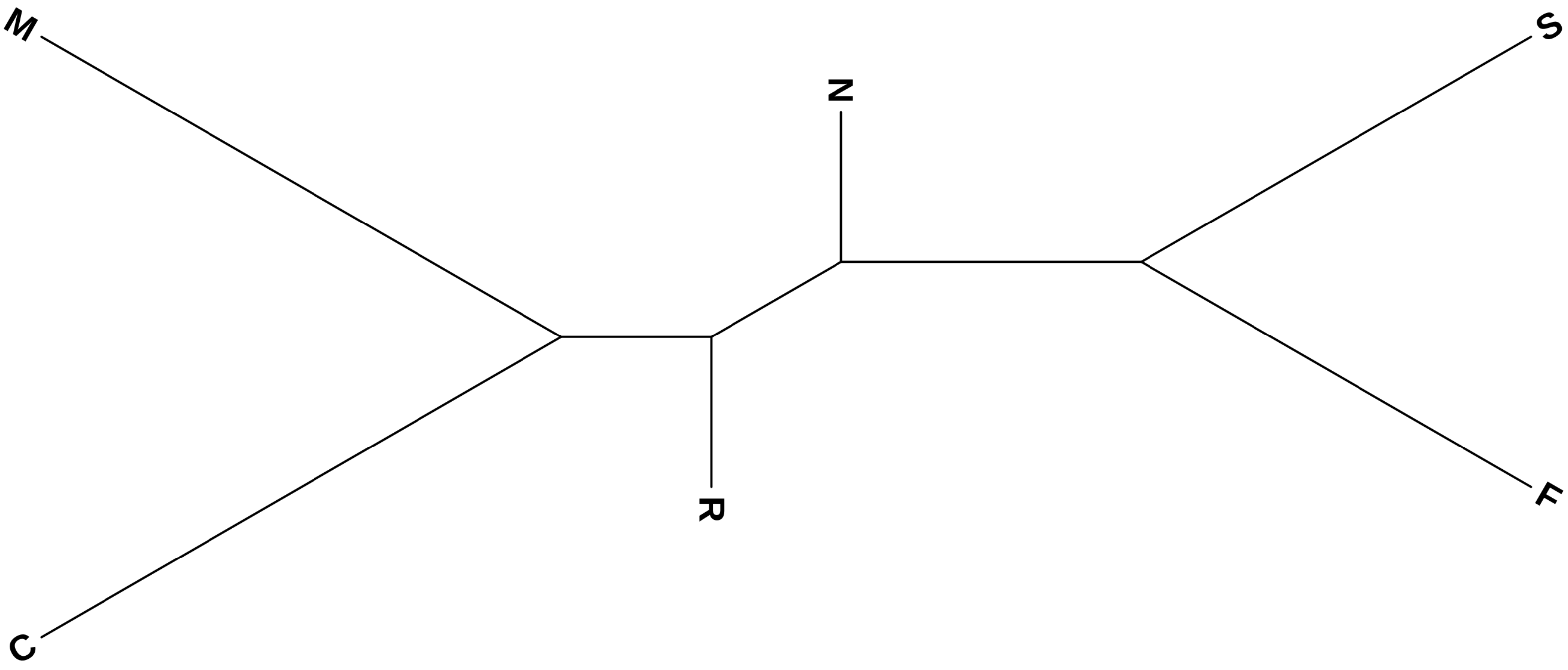
Exercise 1 – Part 1: The grammar

Is the tree specified as rooted or unrooted?

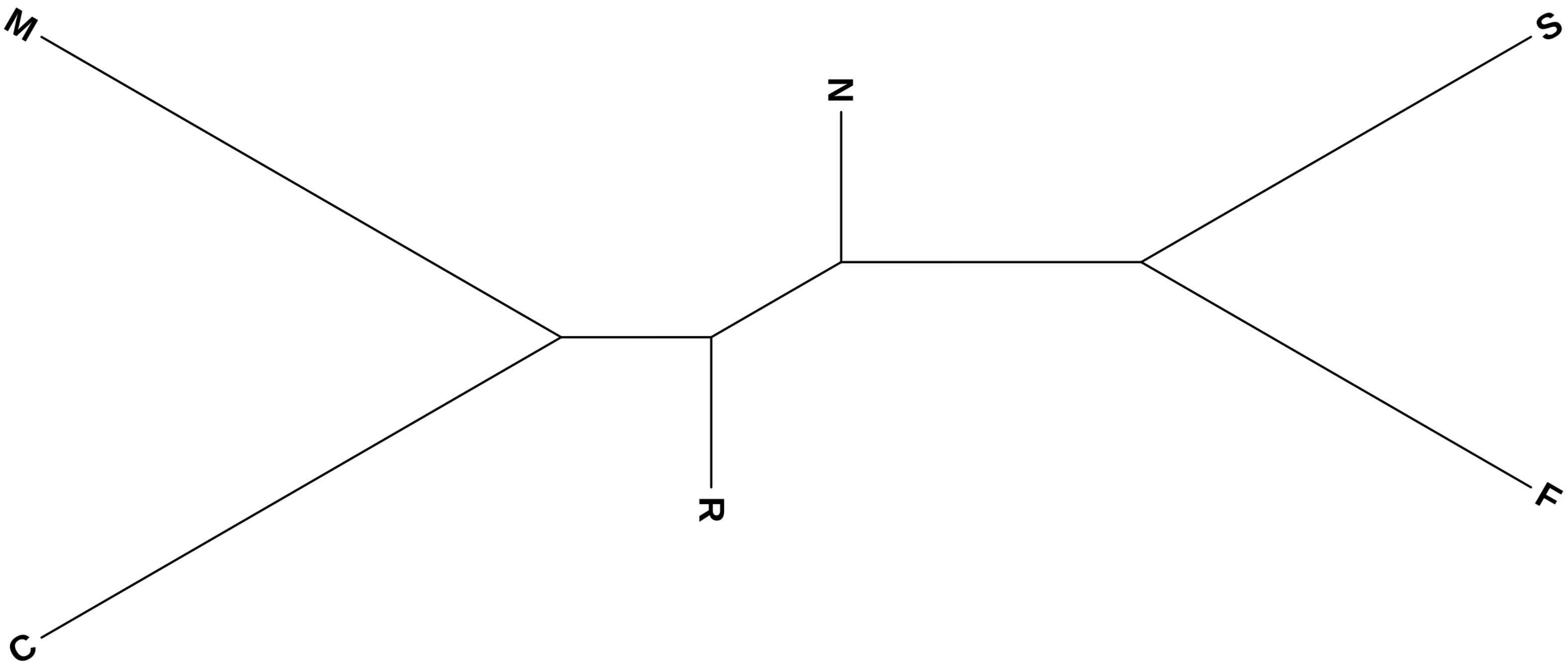
((A,(B,(C,D))),(E,F));

-> The tree is rooted

Exercise 1 – Part 2: Specifying branch length

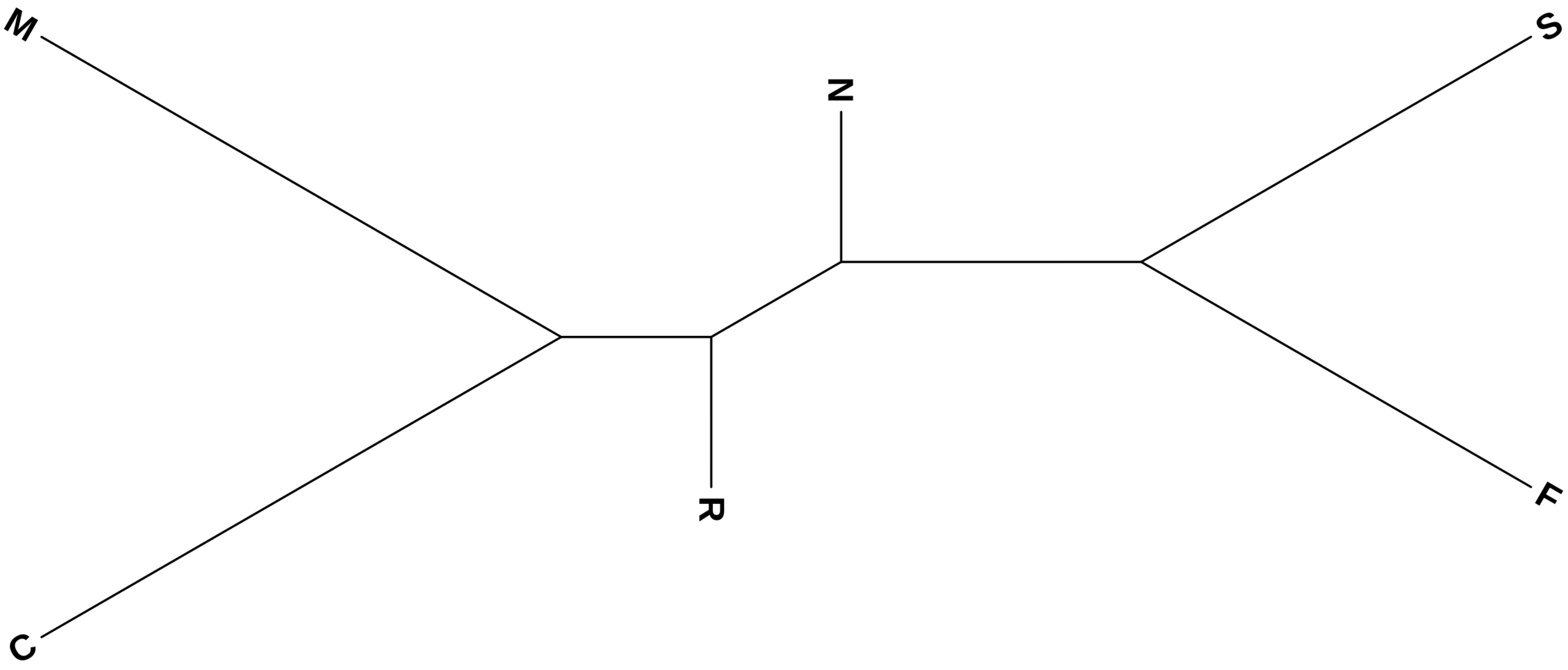


Exercise 1 – Part 2: Specifying branch length



(N,(R,(C,M)),(S,F));

Exercise 1 – Part 2: Specifying branch length



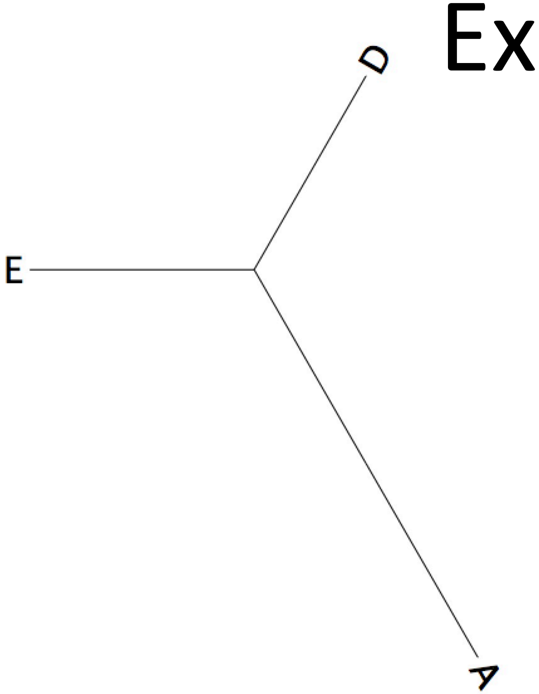
(N:0.5,(R:0.5,(C:2,M:2):0.5):0.5,(S:1.5,F:1.5):1);

Exercise 1 – Part 3: Common errors in Newick representations

Exercise 1 – Part 3.a

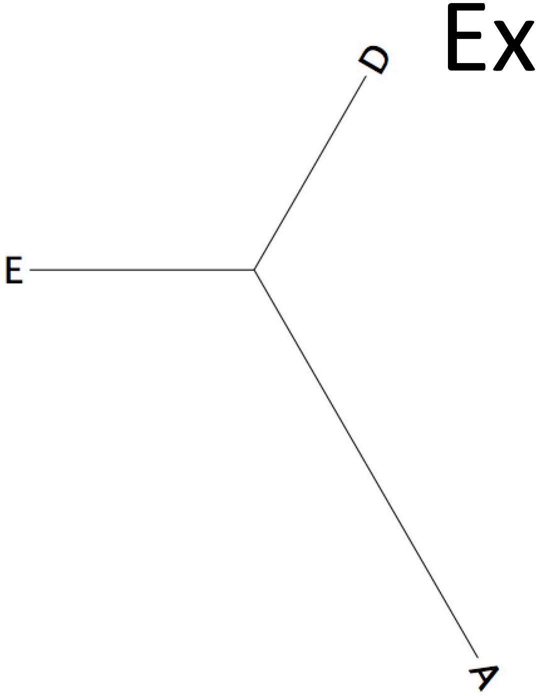
(A,(E,D)),(C,(B,F));

Exercise 1 – Part 3.a



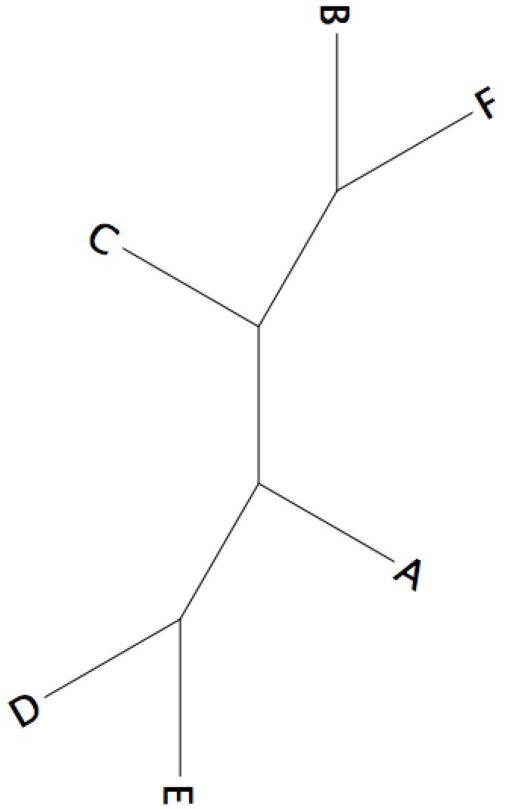
(A,(E,D)),(C,(B,F));

Exercise 1 – Part 3.a



(A,(E,D)),(C,(B,F));

(A,(E,D),(C,(B,F)));

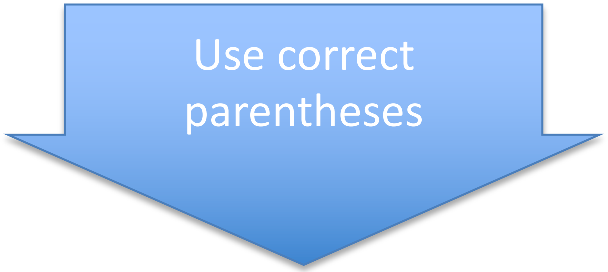


Exercise 1 – Part 3.b

```
(A:1,D:6,([E:1.01,F:1.2]:1,B:2):0.21,(C:4,G:2.2):2):1);
```

Exercise 1 – Part 3.b

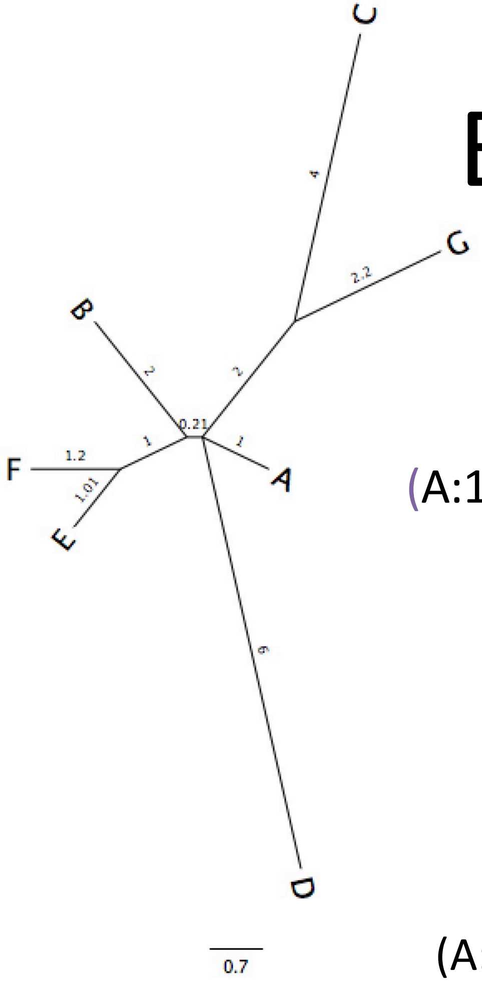
(A:1,D:6,([E:1.01,F:1.2]:1,B:2):0.21,(C:4,G:2.2):2):1);



Use correct parentheses

(A:1,(D:6,((E:1.01,F:1.2):1,B:2):0.21),(C:4,G:2.2):2):1);

Exercise 1 – Part 3.b

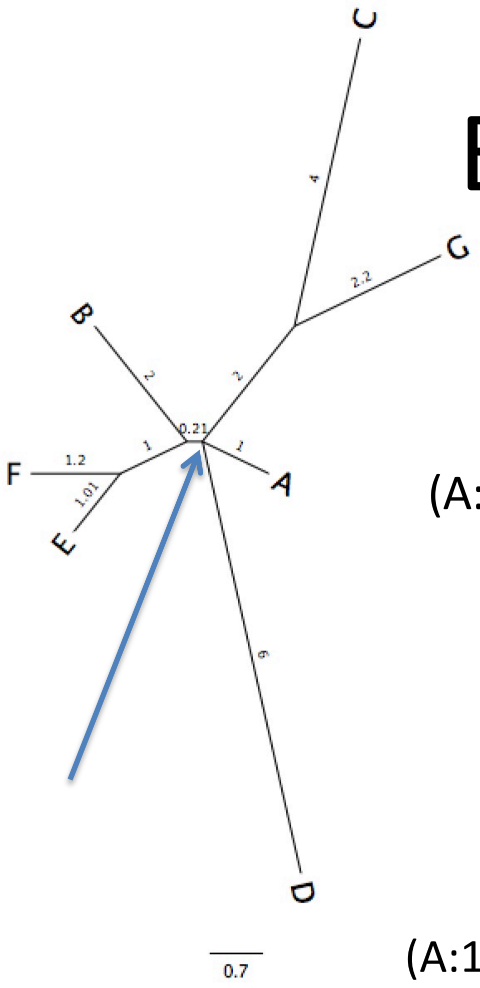


(A:1,D:6,((E:1.01,F:1.2):1,B:2):0.21,(C:4,G:2.2):2):1);

Remove incorrect branch length

(A:1,(D:6,((E:1.01,F:1.2):1,B:2):0.21),(C:4,G:2.2):2);

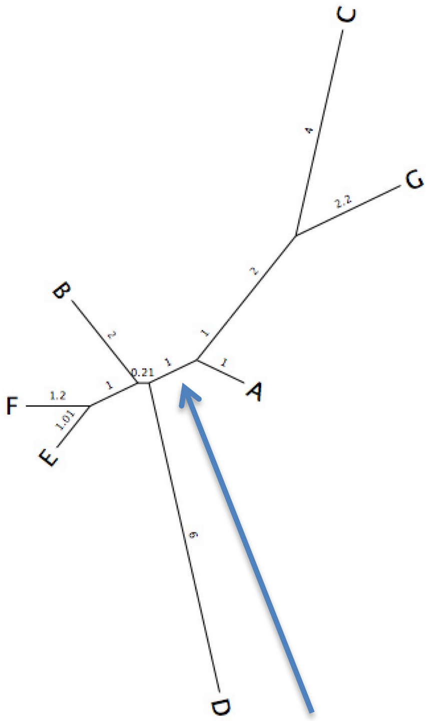
Exercise 1 – Part 3.b



(A:1,D:6,((E:1.01,F:1.2):1,B:2):0.21,(C:4,G:2.2):2);

Resolve multi-branching point

(A:1,**(**D:6,((E:1.01,F:1.2):1,B:2):0.21**):1**,(C:4,G:2.2):2);



Note: There are many ways to solve the multifurcation, this is just one of the many.

Exercise 2 – Overview

Small theory recap

Phylogenetic trees are used to represent the evolutionary relationships between a group of related sequences/species/genes. For tree construction, several computational methods are available. These are as follows:

1. **Distance Based Method**: Neighbor Joining, etc. These require a distance measure between the sequences.
2. **Maximum Parsimony**: The tree based on this method will provide the minimum number of evolutionary steps to produce the sequences.
3. **Maximum Likelihood**: This method uses an expected pattern of mutational changes from one DNA base to another with probability calculations to find the most likely arrangement of branches that generates the set of sequences.

$$L = p(\text{data} \mid \text{tree, branch lengths, model})$$

The ML algorithm searches different trees and branch lengths to find the $\mathbf{L_{max}}$. It works as following

LOOP OVER

Generate tree topology ➔ Optimize branch lengths ➔ Retain if result improved

Exercise 2 – Part 1.d

Can you identify which command line arguments are required (not-optional) for RAxML to be executed?

-s alignment_file_name

-n output_file_name_extension

-m model_of_amino_acid_substitution

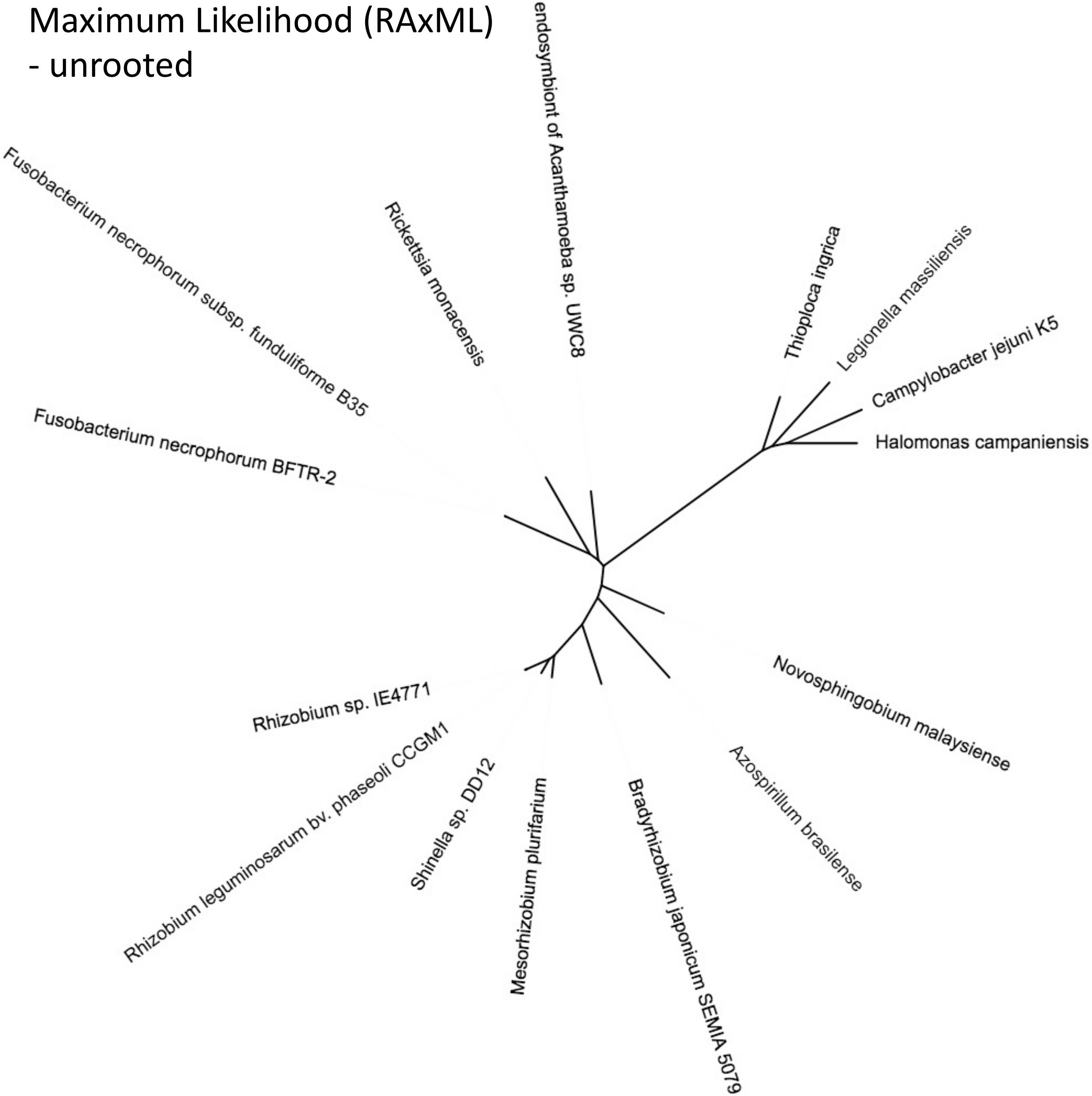
-p random_seed

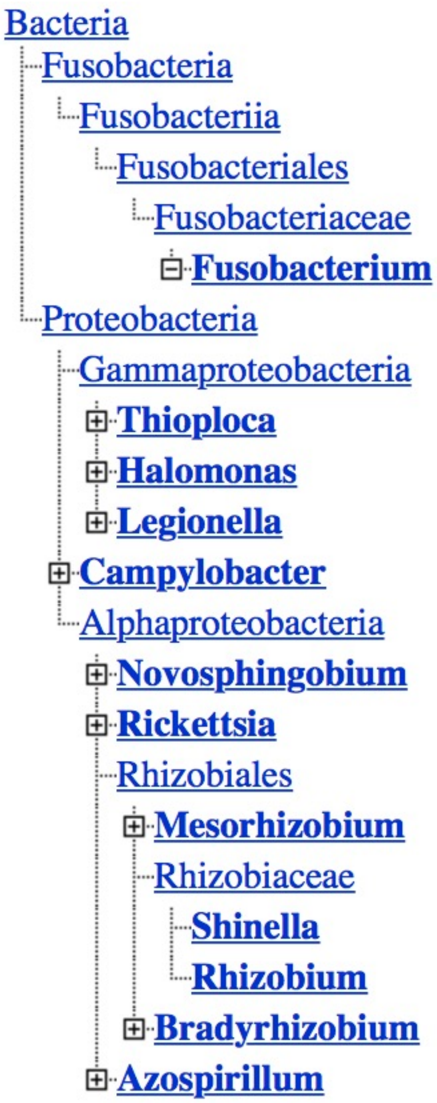
Exercise 2 – Part 2

Using RAxML on the MFS-1 dataset

```
raxmlHPC
-s mfs_domain_proteins_aligned_taxnames.fa
-n msf.auto.txt
-m PROTGAMMAAUTO
-p 123
```

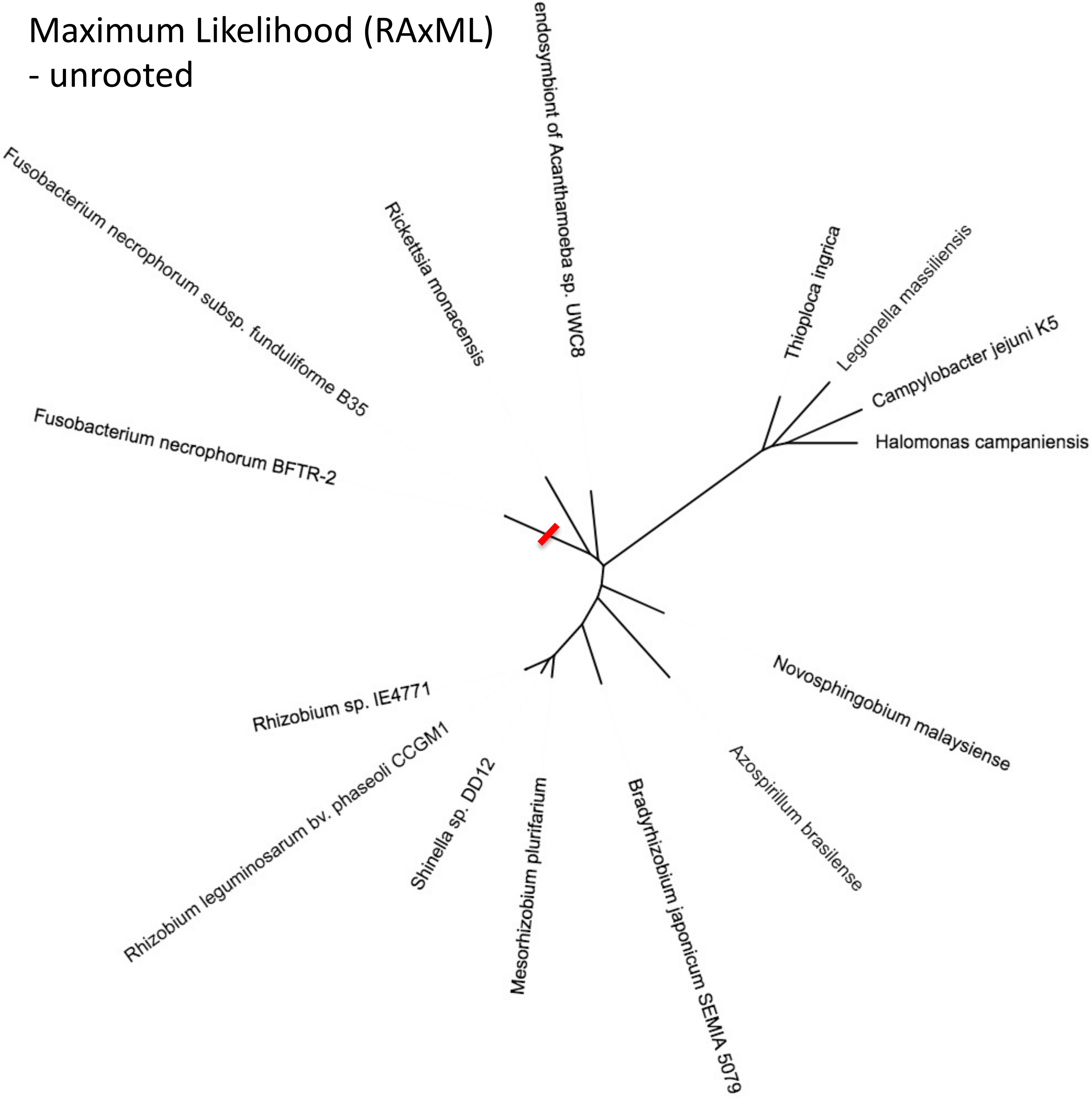
Maximum Likelihood (RAxML) - unrooted

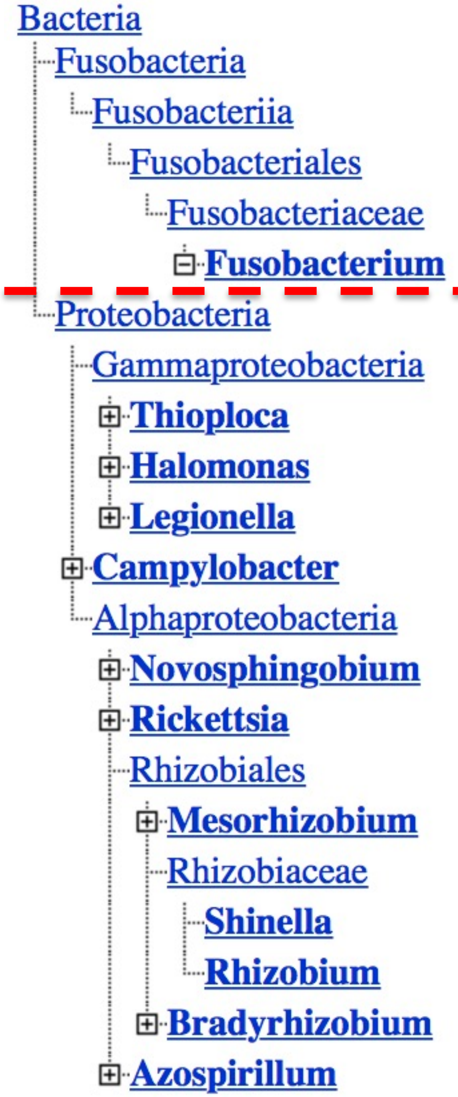

endosymbiont of Acanthamoeba sp. UWC8
Rickettsia monacensis
Fusobacterium necrophorum subsp. funduliforme B35
Fusobacterium necrophorum BFTR-2
Thioploca ingrica
Legionella massiliensis
Campylobacter jejuni K5
Halomonas campaniensis
Novosphingobium malaysiense
Azospirillum brasilense
Bradyrhizobium japonicum SEMIA 5079
Mesorhizobium plurifarium
Shinella sp. DD12
Rhizobium leguminosarum bv. phaseoli CCGM1
Rhizobium sp. IE4771



- Bacteria
  - Fusobacteria
    - Fusobacteriia
      - Fusobacteriales
        - Fusobacteriaceae
          - ⊟ **Fusobacterium**
  - Proteobacteria
    - Gammaproteobacteria
      - ⊞ **Thioploca**
      - ⊞ **Halomonas**
      - ⊞ **Legionella**
    - ⊞ **Campylobacter**
    - Alphaproteobacteria
      - ⊞ **Novosphingobium**
      - ⊞ **Rickettsia**
      - Rhizobiales
        - ⊞ **Mesorhizobium**
        - Rhizobiaceae
          - **Shinella**
          - **Rhizobium**
        - ⊞ **Bradyrhizobium**
      - ⊞ **Azospirillum**

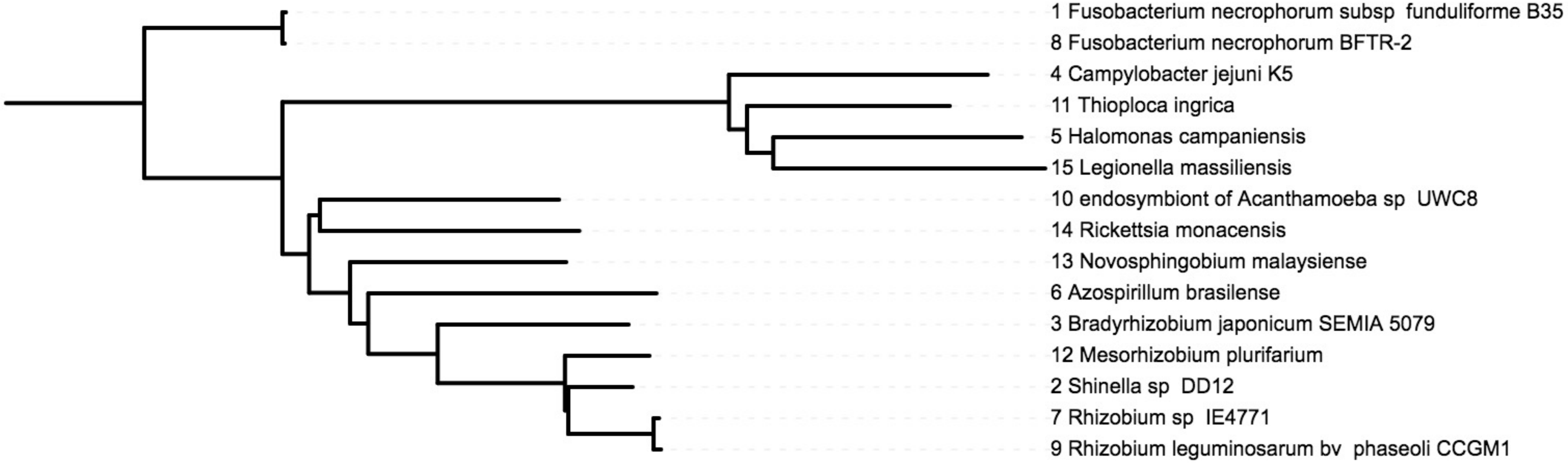
Maximum Likelihood (RAxML) - unrooted


endosymbiont of Acanthamoeba sp. UWC8
Fusobacterium necrophorum subsp. funduliforme B35
Rickettsia monacensis
Fusobacterium necrophorum BFTR-2
Thioploca ingrica
Legionella massiliensis
Campylobacter jejuni K5
Halomonas campaniensis
Novosphingobium malaysiense
Rhizobium sp. IE4771
Rhizobium leguminosarum bv. phaseoli CCGM1
Shinella sp. DD12
Mesorhizobium plurifarium
Bradyrhizobium japonicum SEMIA 5079
Azospirillum brasilense

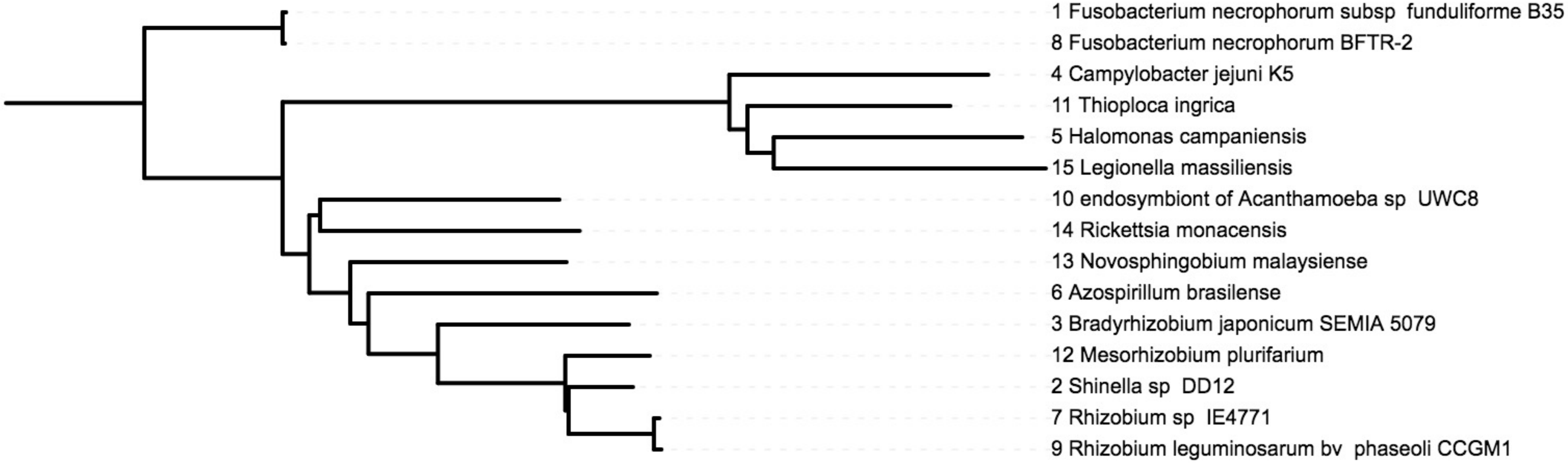


- Bacteria
  - Fusobacteria
    - Fusobacteriia
      - Fusobacteriales
        - Fusobacteriaceae
          - ⊟ **Fusobacterium**
  - Proteobacteria
    - Gammaproteobacteria
      - ⊞ **Thioploca**
      - ⊞ **Halomonas**
      - ⊞ **Legionella**
    - ⊞ **Campylobacter**
    - Alphaproteobacteria
      - ⊞ **Novosphingobium**
      - ⊞ **Rickettsia**
      - Rhizobiales
        - ⊞ **Mesorhizobium**
        - Rhizobiaceae
          - **Shinella**
          - **Rhizobium**
        - ⊞ **Bradyrhizobium**
      - ⊞ **Azospirillum**

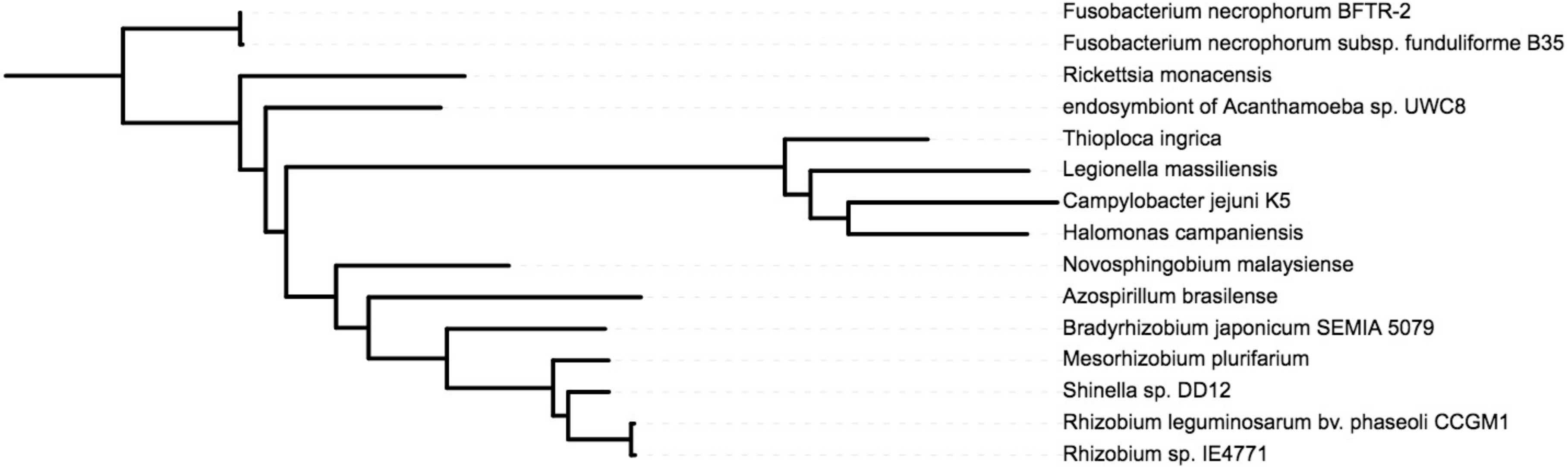
Maximum Likelihood (RAxML)
- rooted on Fusobacteria


1 Fusobacterium necrophorum subsp funduliforme B35
8 Fusobacterium necrophorum BFTR-2
4 Campylobacter jejuni K5
11 Thioploca ingrica
5 Halomonas campaniensis
15 Legionella massiliensis
10 endosymbiont of Acanthamoeba sp UWC8
14 Rickettsia monacensis
13 Novosphingobium malaysiense
6 Azospirillum brasilense
3 Bradyrhizobium japonicum SEMIA 5079
12 Mesorhizobium plurifarium
2 Shinella sp DD12
7 Rhizobium sp IE4771
9 Rhizobium leguminosarum bv phaseoli CCGM1

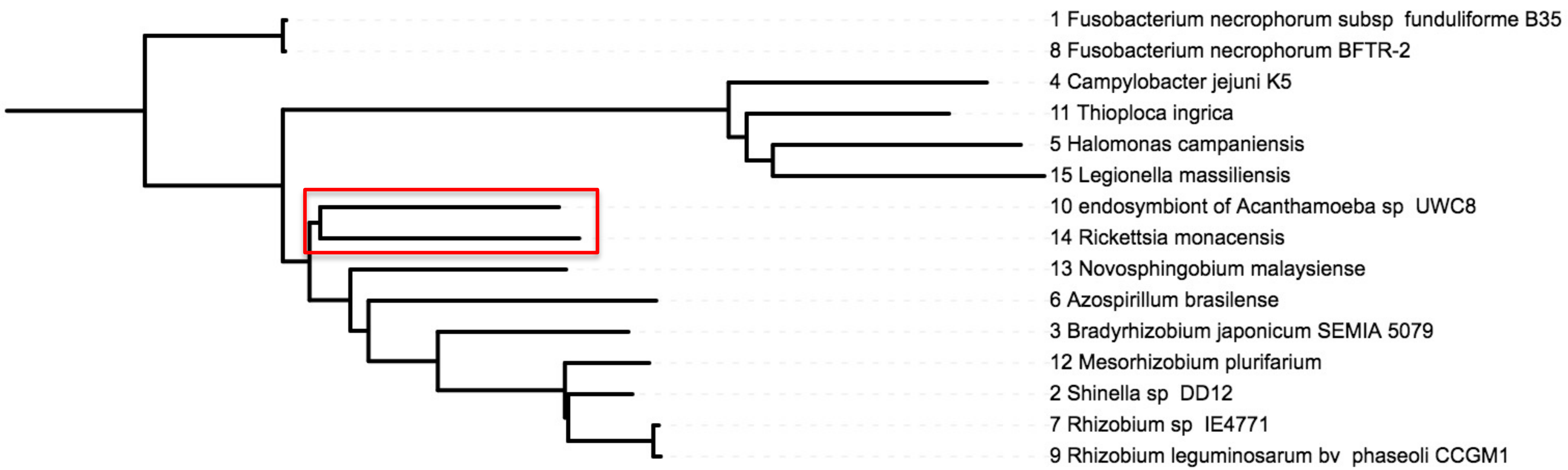
Neighbor joining (rooted on Fusobacteria)


1 Fusobacterium necrophorum subsp funduliforme B35
8 Fusobacterium necrophorum BFTR-2
4 Campylobacter jejuni K5
11 Thioploca ingrica
5 Halomonas campaniensis
15 Legionella massiliensis
10 endosymbiont of Acanthamoeba sp UWC8
14 Rickettsia monacensis
13 Novosphingobium malaysiense
6 Azospirillum brasilense
3 Bradyrhizobium japonicum SEMIA 5079
12 Mesorhizobium plurifarium
2 Shinella sp DD12
7 Rhizobium sp IE4771
9 Rhizobium leguminosarum bv phaseoli CCGM1

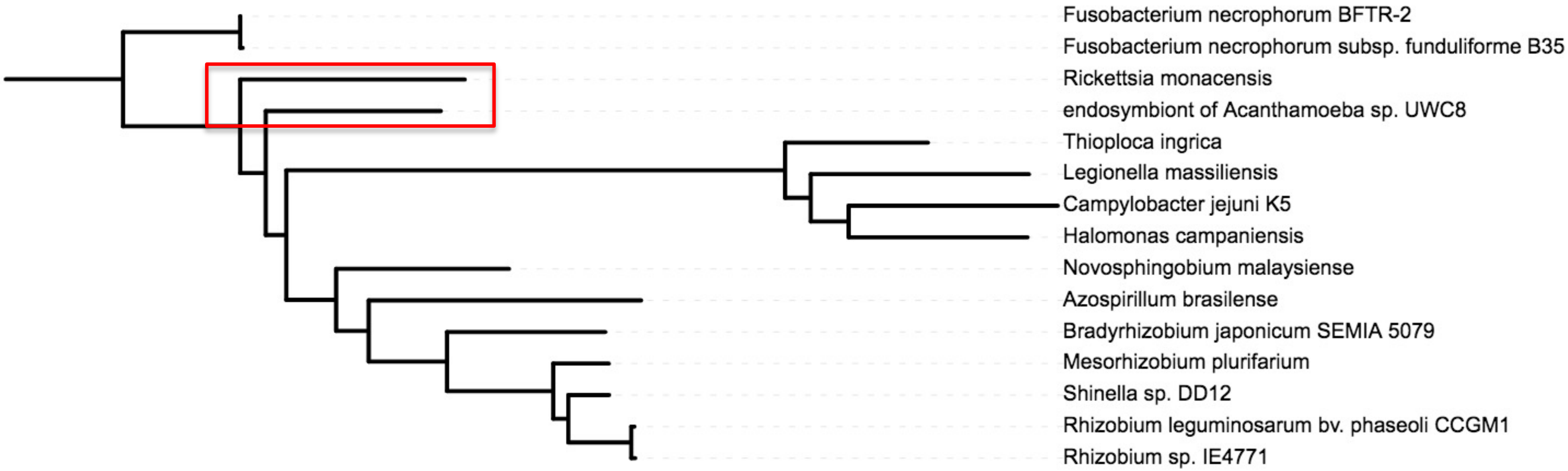
Maximum Likelihood (RAxML, rooted on Fusobacteria)


Fusobacterium necrophorum BFTR-2
Fusobacterium necrophorum subsp. funduliforme B35
Rickettsia monacensis
endosymbiont of Acanthamoeba sp. UWC8
Thioploca ingrica
Legionella massiliensis
Campylobacter jejuni K5
Halomonas campaniensis
Novosphingobium malaysiense
Azospirillum brasilense
Bradyrhizobium japonicum SEMIA 5079
Mesorhizobium plurifarium
Shinella sp. DD12
Rhizobium leguminosarum bv. phaseoli CCGM1
Rhizobium sp. IE4771

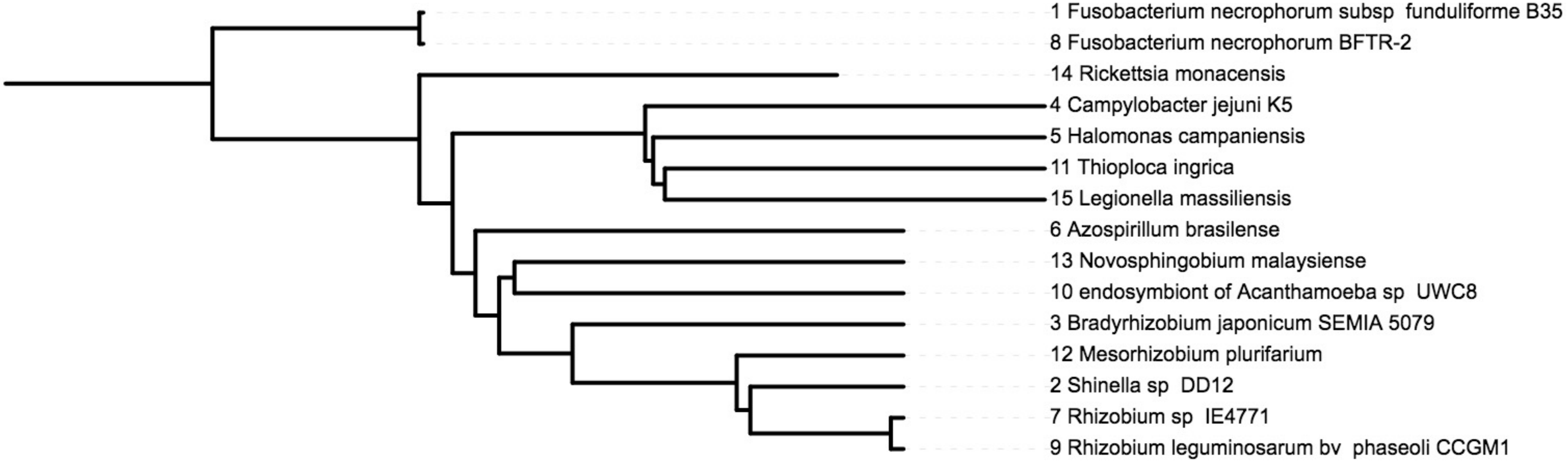
Neighbor joining (rooted on Fusobacteria)


1 Fusobacterium necrophorum subsp funduliforme B35
8 Fusobacterium necrophorum BFTR-2
4 Campylobacter jejuni K5
11 Thioploca ingrica
5 Halomonas campaniensis
15 Legionella massiliensis
10 endosymbiont of Acanthamoeba sp UWC8
14 Rickettsia monacensis
13 Novosphingobium malaysiense
6 Azospirillum brasilense
3 Bradyrhizobium japonicum SEMIA 5079
12 Mesorhizobium plurifarium
2 Shinella sp DD12
7 Rhizobium sp IE4771
9 Rhizobium leguminosarum bv phaseoli CCGM1

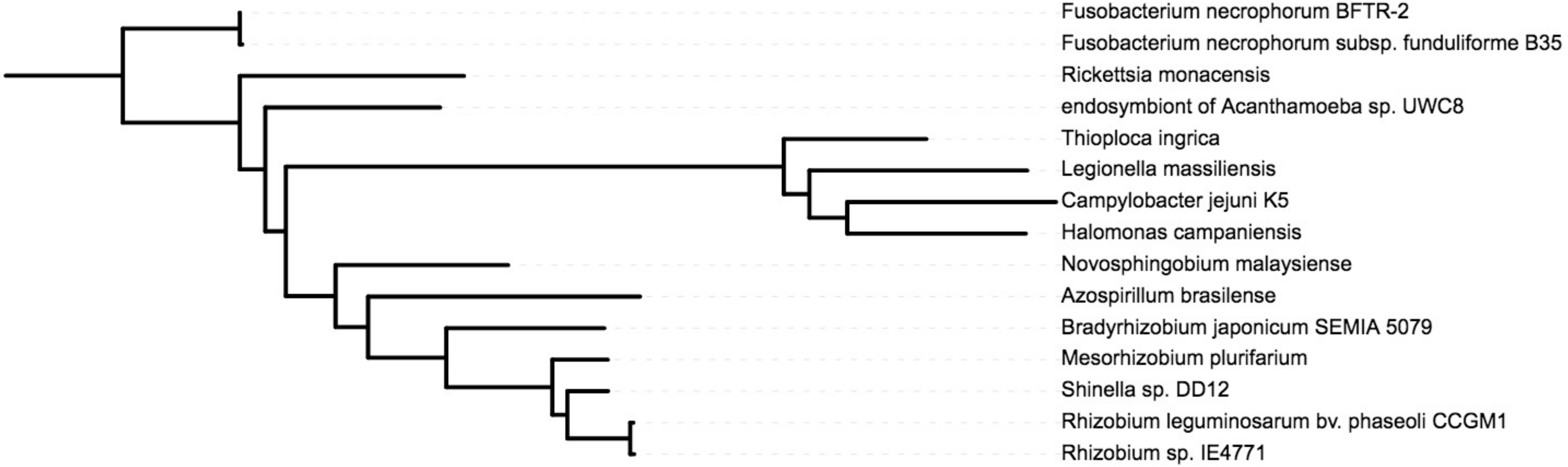
Maximum Likelihood (RAxML, rooted on Fusobacteria)


Fusobacterium necrophorum BFTR-2
Fusobacterium necrophorum subsp. funduliforme B35
Rickettsia monacensis
endosymbiont of Acanthamoeba sp. UWC8
Thioploca ingrica
Legionella massiliensis
Campylobacter jejuni K5
Halomonas campaniensis
Novosphingobium malaysiense
Azospirillum brasilense
Bradyrhizobium japonicum SEMIA 5079
Mesorhizobium plurifarium
Shinella sp. DD12
Rhizobium leguminosarum bv. phaseoli CCGM1
Rhizobium sp. IE4771

UPGMA (rooted on Fusobacteria)

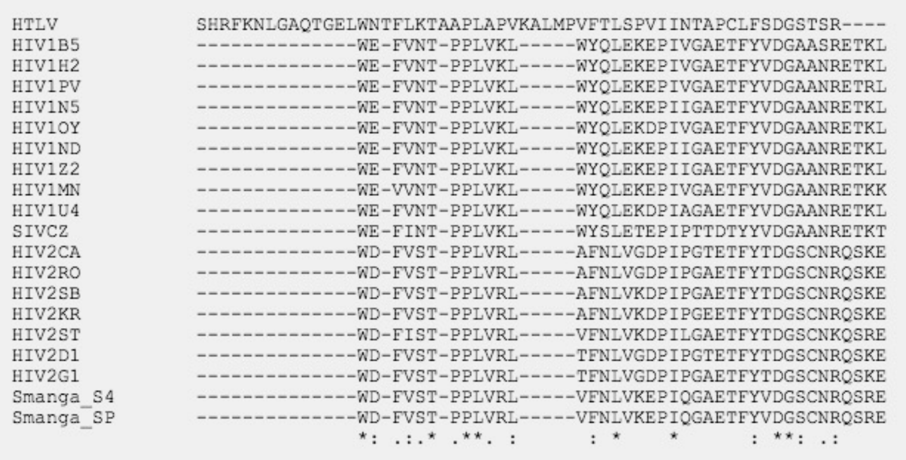

1 Fusobacterium necrophorum subsp funduliforme B35
8 Fusobacterium necrophorum BFTR-2
14 Rickettsia monacensis
4 Campylobacter jejuni K5
5 Halomonas campaniensis
11 Thioploca ingrica
15 Legionella massiliensis
6 Azospirillum brasilense
13 Novosphingobium malaysiense
10 endosymbiont of Acanthamoeba sp UWC8
3 Bradyrhizobium japonicum SEMIA 5079
12 Mesorhizobium plurifarium
2 Shinella sp DD12
7 Rhizobium sp IE4771
9 Rhizobium leguminosarum bv phaseoli CCGM1

Maximum Likelihood (RAxML, rooted on Fusobacteria)

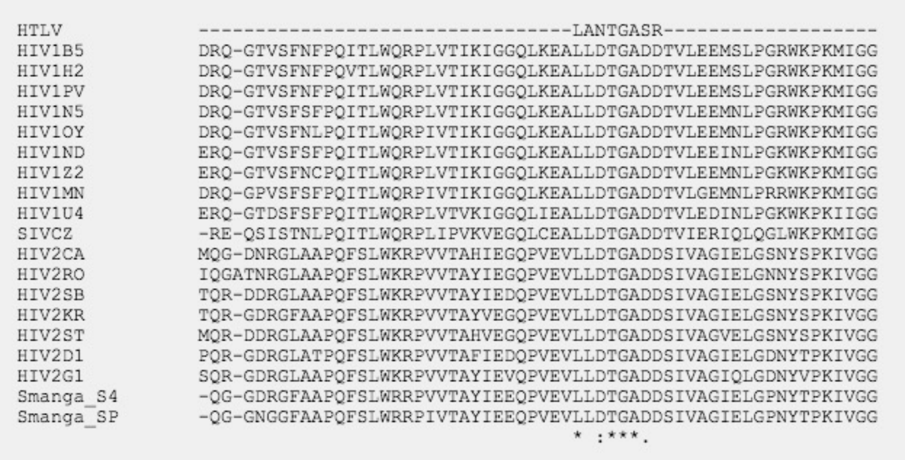

Fusobacterium necrophorum BFTR-2
Fusobacterium necrophorum subsp. funduliforme B35
Rickettsia monacensis
endosymbiont of Acanthamoeba sp. UWC8
Thioploca ingrica
Legionella massiliensis
Campylobacter jejuni K5
Halomonas campaniensis
Novosphingobium malaysiense
Azospirillum brasilense
Bradyrhizobium japonicum SEMIA 5079
Mesorhizobium plurifarium
Shinella sp. DD12
Rhizobium leguminosarum bv. phaseoli CCGM1
Rhizobium sp. IE4771

Exercise 3 – Part 1

Sequence alignment



```
HTLV              SHRFKNLGAQTGELWNTFLKTAAPLAPVKALMPVFTLSPVIINTAPCLFSDGSTSR----
HIV1B5            --------------WE-FVNT-PPLVKL-----WYQLEKEPIVGAETFYVDGAASRETKL
HIV1H2            --------------WE-FVNT-PPLVKL-----WYQLEKEPIVGAETFYVDGAANRETKL
HIV1PV            --------------WE-FVNT-PPLVKL-----WYQLEKEPIVGAETFYVDGAANRETRL
HIV1N5            --------------WE-FVNT-PPLVKL-----WYQLEKEPIIGAETFYVDGAANRETKL
HIV1OY            --------------WE-FVNT-PPLVKL-----WYQLEKDPIVGAETFYVDGAANRETKL
HIV1ND            --------------WE-FVNT-PPLVKL-----WYQLEKEPIIGAETFYVDGAANRETKL
HIV1Z2            --------------WE-FVNT-PPLVKL-----WYQLEKEPIIGAETFYVDGAANRETKL
HIV1MN            --------------WE-VVNT-PPLVKL-----WYQLEKEPIVGAETFYVDGAANRETKK
HIV1U4            --------------WE-FVNT-PPLVKL-----WYQLEKDPIAGAETFYVDGAANRETKL
SIVCZ             --------------WE-FINT-PPLVKL-----WYSLETEPIPTTDTYYVDGAANRETKT
HIV2CA            --------------WD-FVST-PPLVRL-----AFNLVGDPIPGTETFYTDGSCNRQSKE
HIV2RO            --------------WD-FVST-PPLVRL-----AFNLVGDPIPGAETFYTDGSCNRQSKE
HIV2SB            --------------WD-FVST-PPLVRL-----AFNLVKDPIPGAETFYTDGSCNRQSKE
HIV2KR            --------------WD-FVST-PPLVRL-----AFNLVKDPIPGEETFYTDGSCNRQSKE
HIV2ST            --------------WD-FIST-PPLVRL-----VFNLVKDPILGAETFYTDGSCNKQSRE
HIV2D1            --------------WD-FVST-PPLVRL-----TFNLVGDPIPGTETFYTDGSCNRQSKE
HIV2G1            --------------WD-FVST-PPLVRL-----TFNLVGDPIPGAETFYTDGSCNRQSKE
Smanga_S4         --------------WD-FVST-PPLVRL-----VFNLVKEPIQGAETFYVDGSCNRQSRE
Smanga_SP         --------------WD-FVST-PPLVRL-----VFNLVKEPIQGAETFYVDGSCNRQSRE
                                *: .:.* .**. :      : *    *      : **: .:
```



```
HTLV              ---------------------------------LANTGASR-------------------
HIV1B5            DRQ-GTVSFNFPQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGG
HIV1H2            DRQ-GTVSFNFPQVTLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGG
HIV1PV            DRQ-GTVSFNFPQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGG
HIV1N5            DRQ-GTVSFSFPQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGG
HIV1OY            DRQ-GTVSFNLPQITLWQRPIVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGG
HIV1ND            ERQ-GTVSFSFPQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEINLPGKWKPKMIGG
HIV1Z2            ERQ-GTVSFNCPQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGKWKPKMIGG
HIV1MN            DRQ-GPVSFSFPQITLWQRPIVTIKIGGQLKEALLDTGADDTVLGEMNLPRRWKPKMIGG
HIV1U4            ERQ-GTDSFSFPQITLWQRPLVTVKIGGQLIEALLDTGADDTVLEDINLPGKWKPKIIGG
SIVCZ             -RE-QSISTNLPQITLWQRPLIPVKVEGQLCEALLDTGADDTVIERIQLQGLWKPKMIGG
HIV2CA            MQG-DNRGLAAPQFSLWKRPVVTAHIEGQPVEVLLDTGADDSIVAGIELGSNYSPKIVGG
HIV2RO            IQGATNRGLAAPQFSLWKRPVVTAYIEGQPVEVLLDTGADDSIVAGIELGNNYSPKIVGG
HIV2SB            TQR-DDRGLAAPQFSLWKRPVVTAYIEDQPVEVLLDTGADDSIVAGIELGSNYSPKIVGG
HIV2KR            TQR-GDRGFAAPQFSLWKRPVVTAYVEGQPVEVLLDTGADDSIVAGIELGSNYSPKIVGG
HIV2ST            MQR-DDRGLAAPQFSLWKRPVVTAHVEGQPVEVLLDTGADDSIVAGVELGSNYSPKIVGG
HIV2D1            PQR-GDRGLATPQFSLWKRPVVTAFIEDQPVEVLLDTGADDSIVAGIELGDNYTPKIVGG
HIV2G1            SQR-GDRGLAAPQFSLWKRPVVTAYIEVQPVEVLLDTGADDSIVAGIQLGDNYVPKIVGG
Smanga_S4         -QG-GDRGFAAPQFSLWRRPVVTAYIEEQPVEVLLDTGADDSIVAGIELGPNYTPKIVGG
Smanga_SP         -QG-GNGGFAAPQFSLWRRPIVTAYIEEQPVEVLLDTGADDSIVAGIELGPNYTPKIVGG
                                                   * :***.
```

Exercise 3 – Part 2

Tree reconstruction with RAxML
While the program is running, can you identify what parameters we are using?

-s

-n

-m

-f

-N

-x

-p

Exercise 3 – Part 2

While the program is running, can you identify what parameters we are using?

-s name of the alignment data file

-n name of the output file

-m Model of Binary (Morphological), Nucleotide, Multi-State, or Amino Acid Substitution

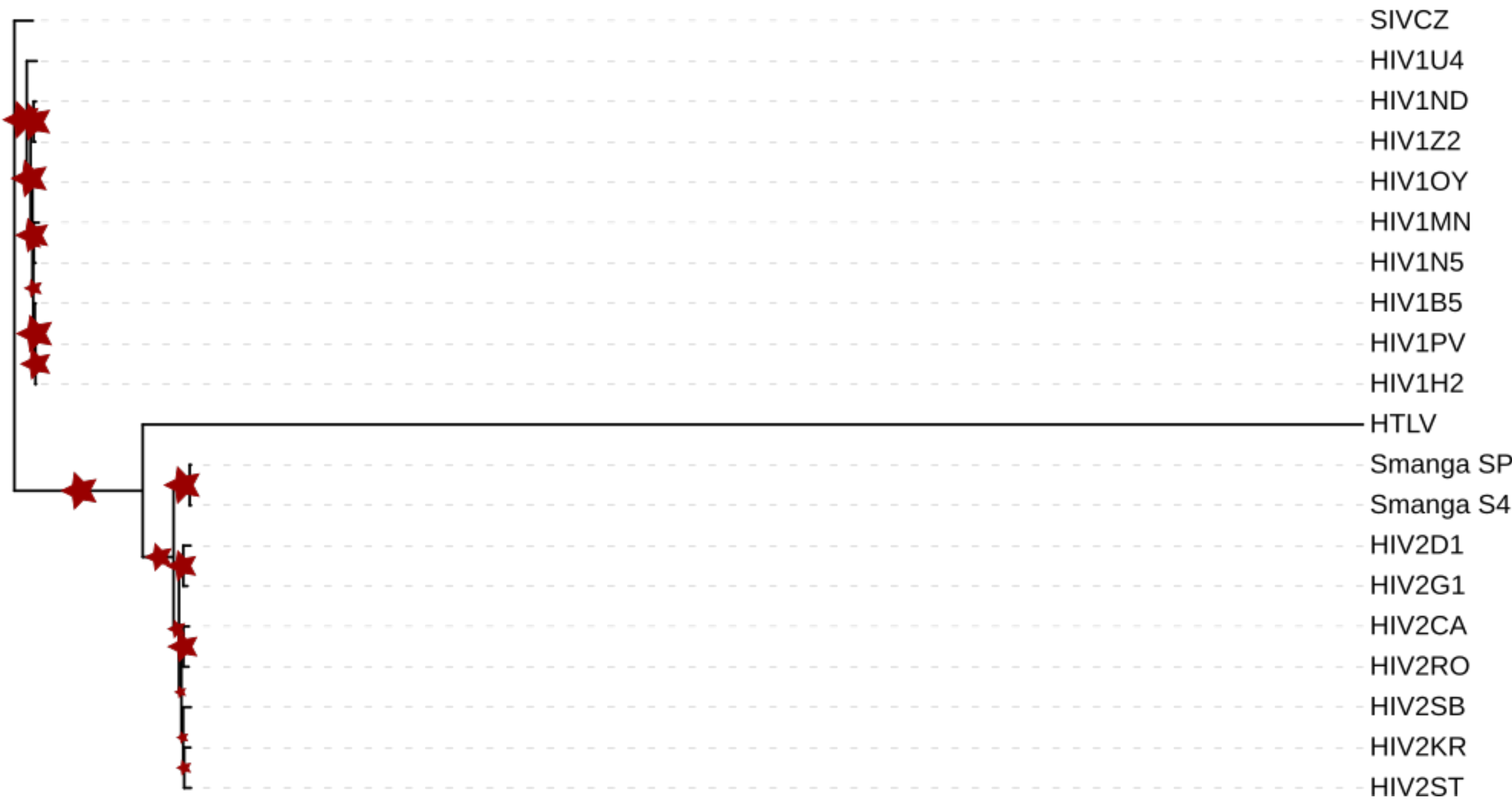
-f select algorithm

-N number of alternative runs on distinct starting trees

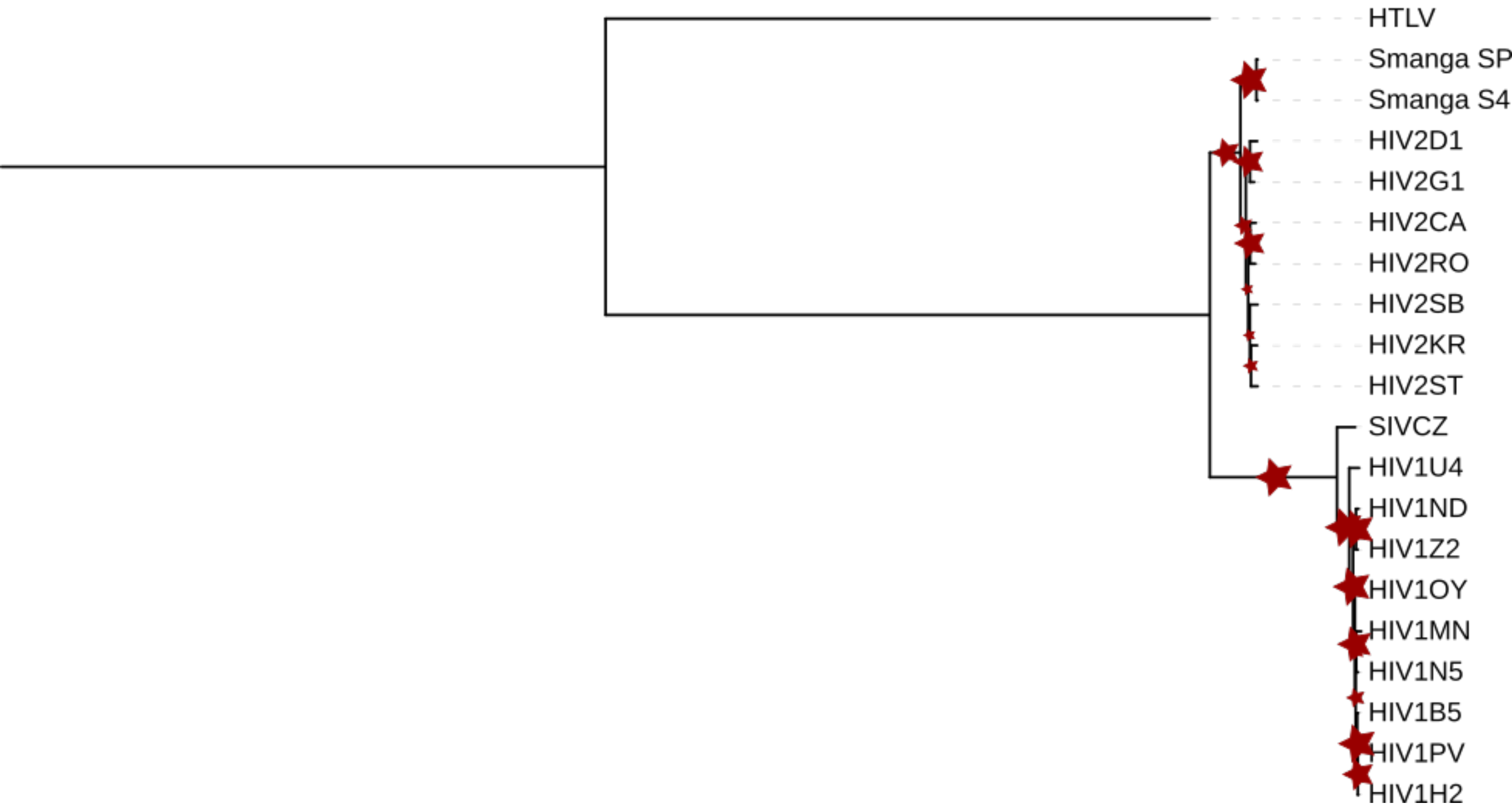
-x random seed for rapid bootstrapping

-p random number seed for the parsimony inferences

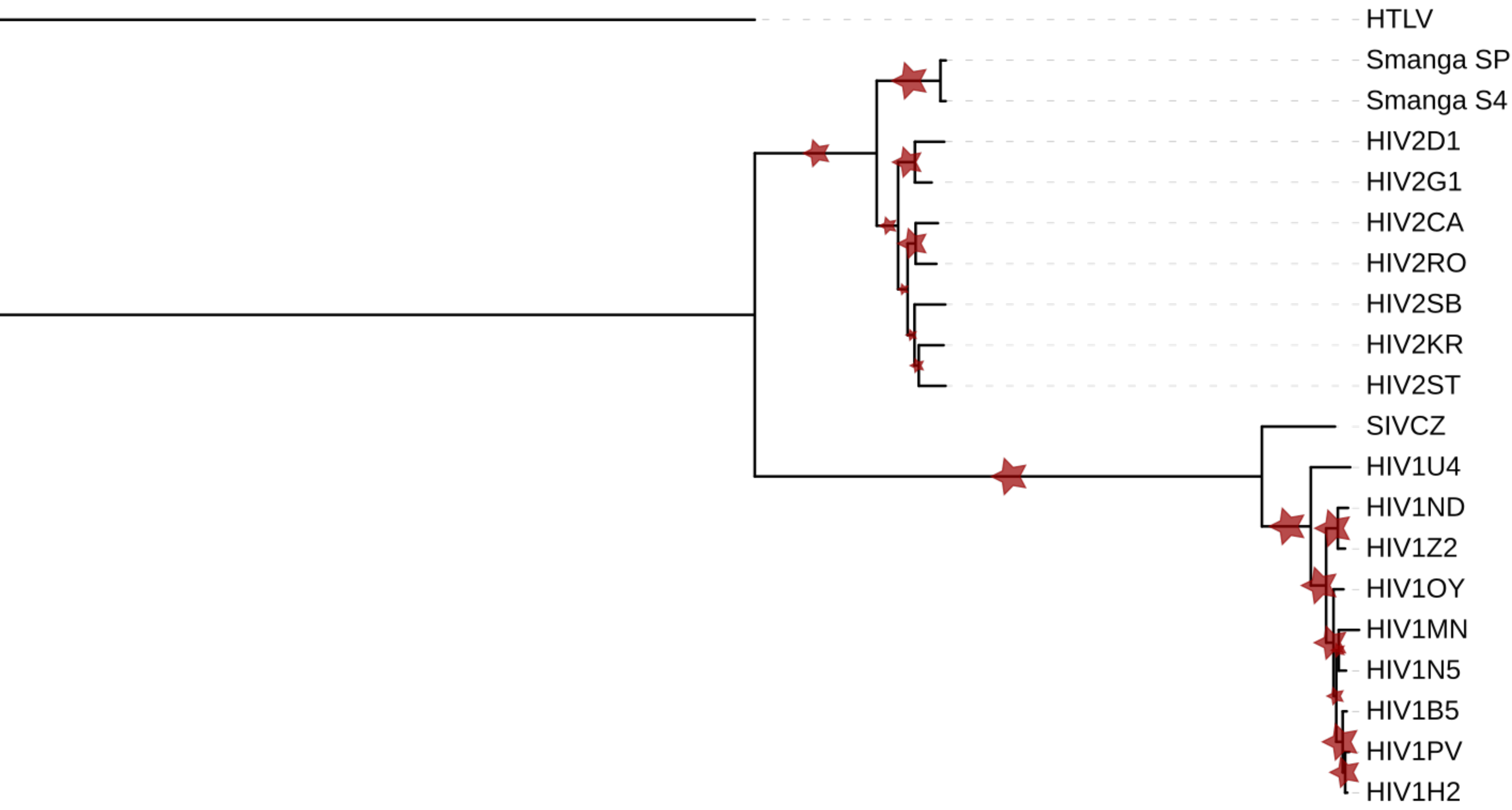
Exercise 3 – Part 3



Exercise 3 – Part 4



Exercise 3 – Part 4


HTLV
Smanga SP
Smanga S4
HIV2D1
HIV2G1
HIV2CA
HIV2RO
HIV2SB
HIV2KR
HIV2ST
SIVCZ
HIV1U4
HIV1ND
HIV1Z2
HIV1OY
HIV1MN
HIV1N5
HIV1B5
HIV1PV
HIV1H2