hw2

September 29, 2020

0.0.1 Imports secion

[1]:
```
import matplotlib
import re
import requests

import matplotlib.pyplot as plt
import pandas as pd

from random import randint
from Bio import Phylo, Entrez
from ete3 import Tree, TreeStyle, NodeStyle
from io import StringIO
```

0.0.2 Define some functions

[2]:
```
def get_tree(url: str, ete=False):
    tree_text = requests.get(url).text
    if ete:
        return Tree(tree_text, format=1)
    return Phylo.read(StringIO(tree_text), 'newick')

def plot_tree(tree: Phylo.Newick.Tree, out_file=None, show=False) -> None:
    matplotlib.rc('font', size=6)
    fig = plt.figure(figsize=(10, 20), dpi=100)
    axes = fig.add_subplot(1, 1, 1)
    Phylo.draw(tree, axes=axes, do_show=show)
    if out_file:
        plt.savefig(out_file, dpi=100)
        plt.close()
def entrez_gene(term: str, db='nucleotide') -> dict:
    res = Entrez.esearch(db='nucleotide', term=term)
    return Entrez.read(res)

def entrez_summary(gene_id: str, db='nucleotide') -> pd.DataFrame:
    summary = Entrez.esummary(db='nucleotide', id=gene_id)
    res = Entrez.read(summary)
    df = pd.DataFrame(res)[['Id', 'Caption', 'Length']]
```

```
    df.index = [gene_id]
    return df

def get_fasta(gene_id: str, db='nucleotide') -> str:
    return Entrez.efetch(db=db, id=gene_id, rettype="fasta", retmode="text").
 ↪read()
```

0.1 PART 1

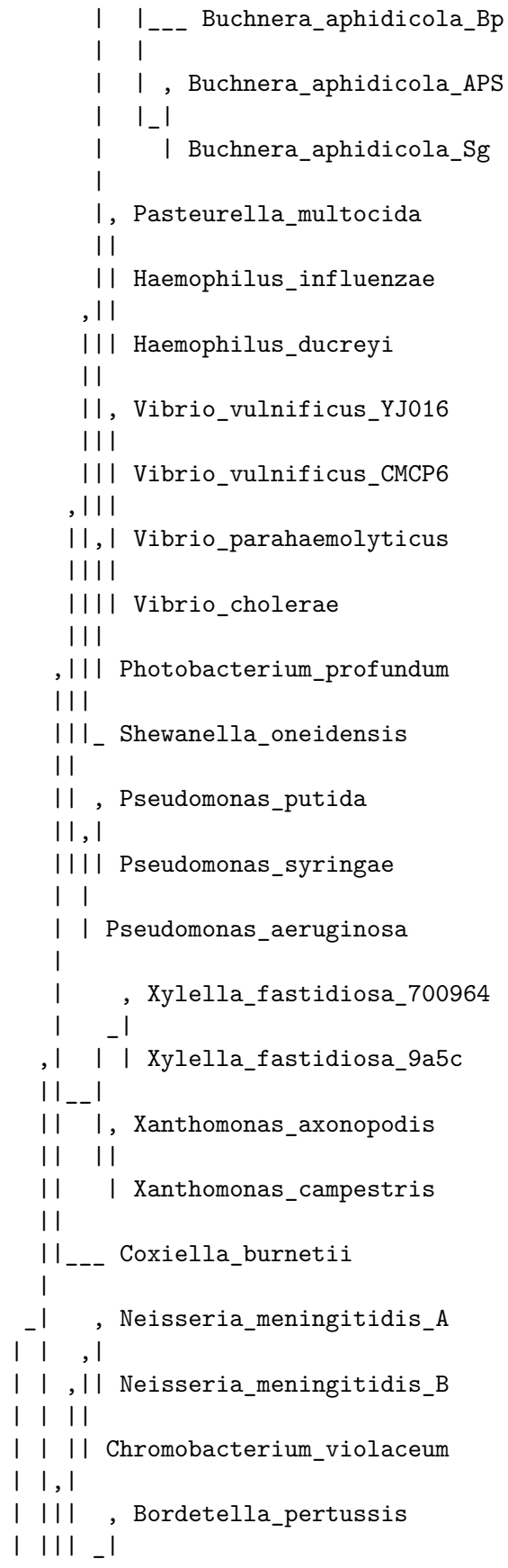
0.1.1 1.1 Read the tree

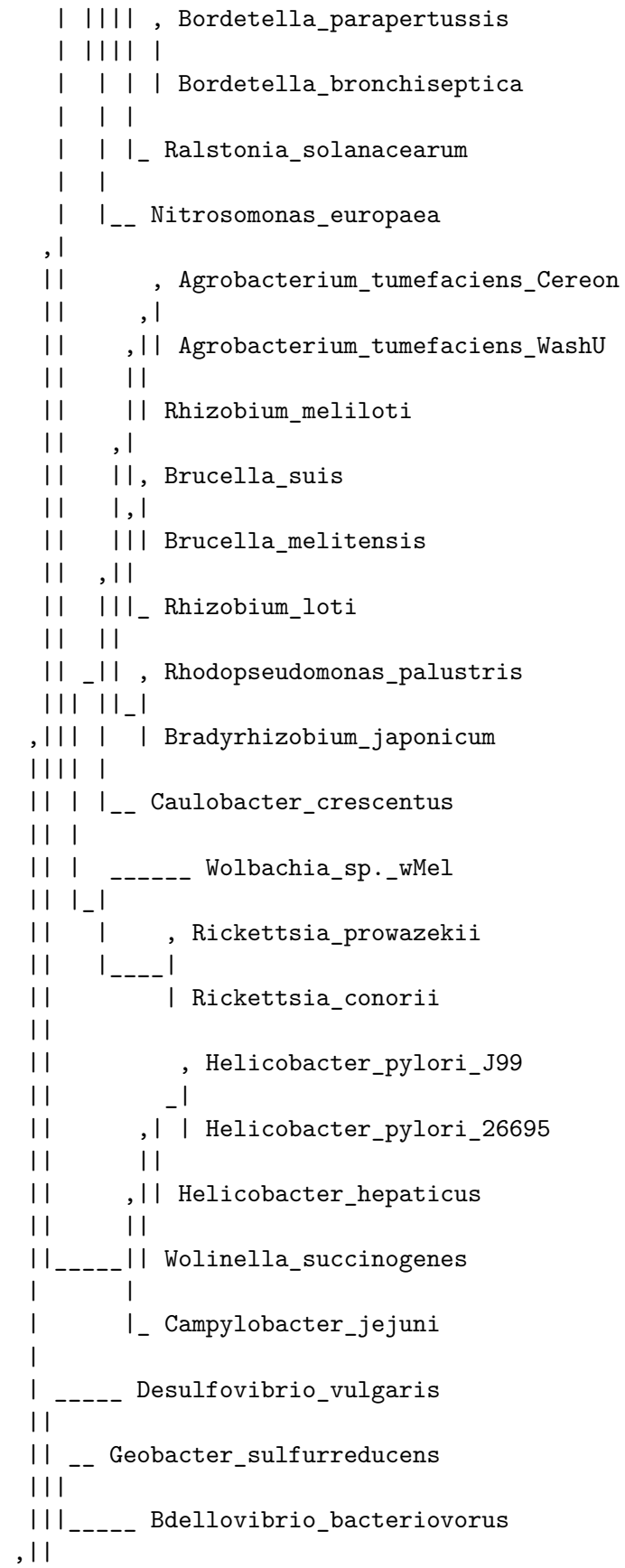
[3]:
```
tree = get_tree('https://www.jasondavies.com/tree-of-life/life.txt')
```

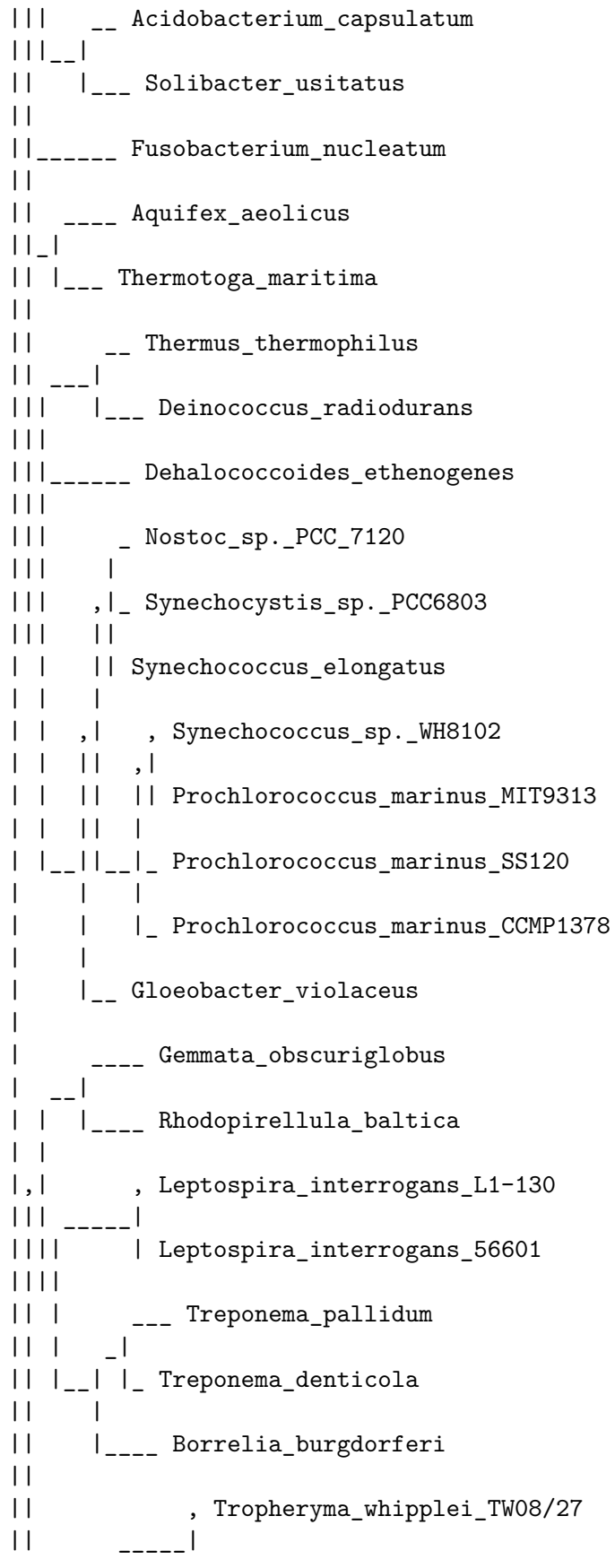
0.1.2 1.2 draw the tree with pseudographics

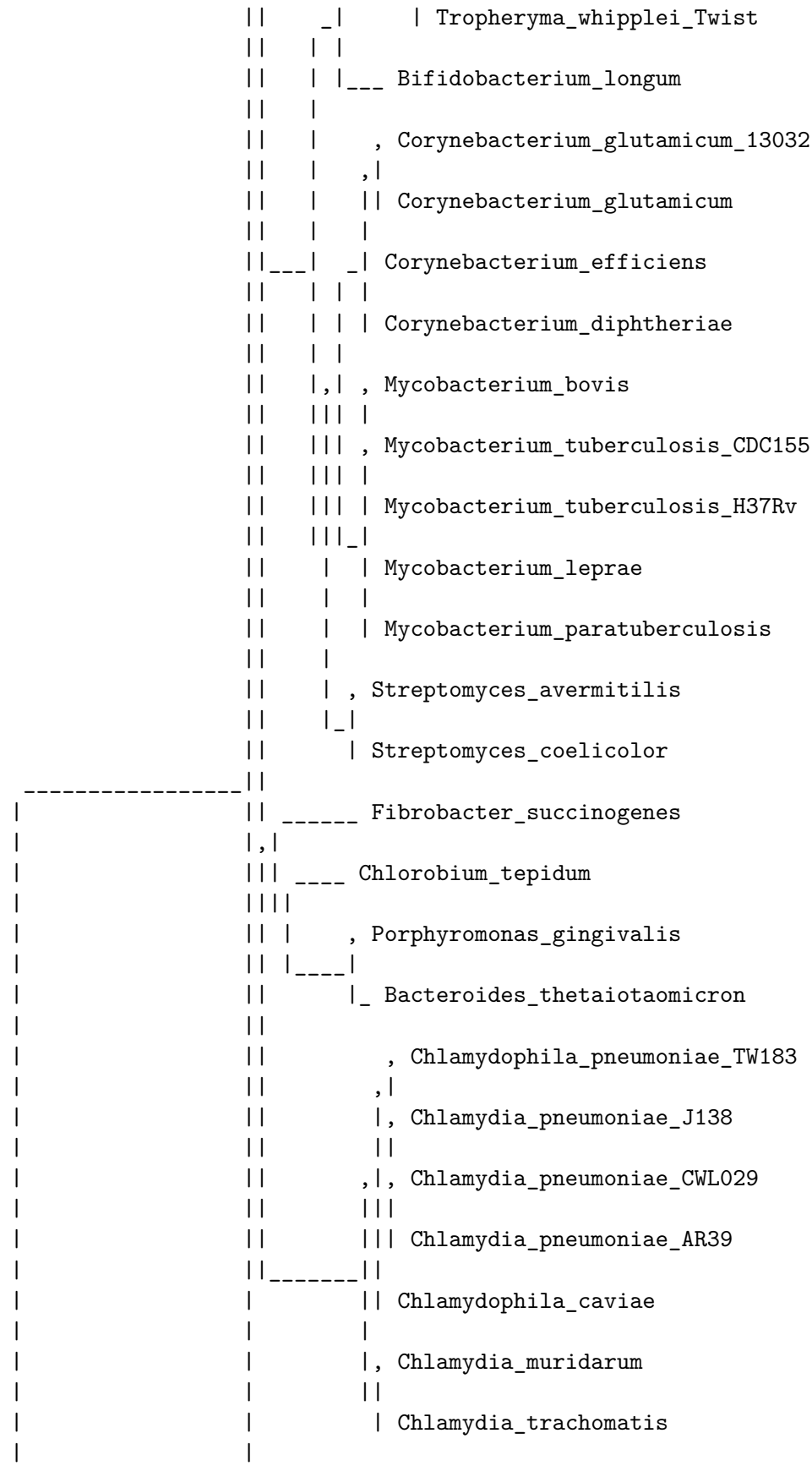
[4]:
```
Phylo.draw_ascii(tree)
```

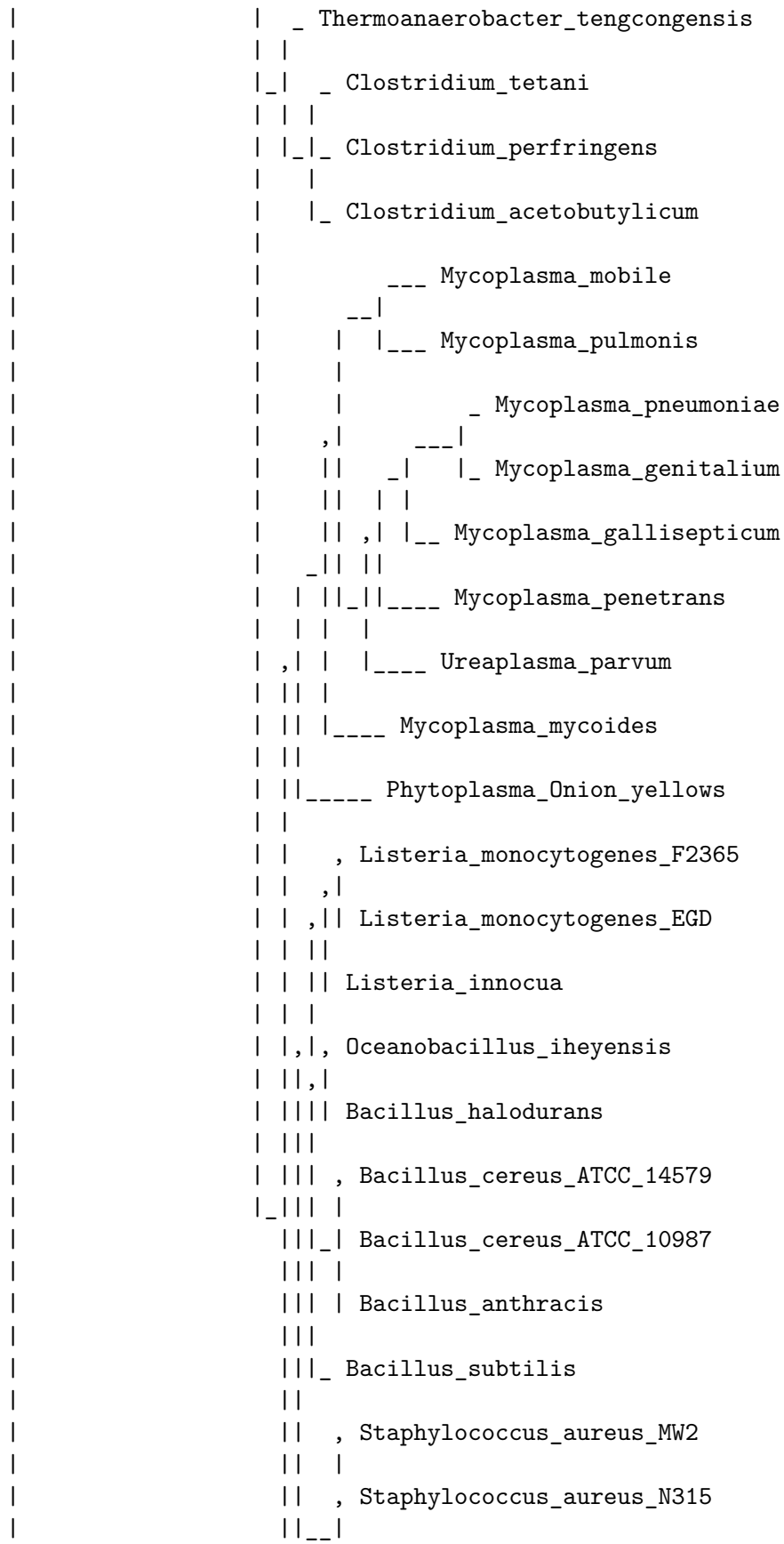
```
                                  , Escherichia_coli_EDL933
                                  |
                                  | Escherichia_coli_O157_H7
                                  |
                                  , Escherichia_coli_O6
                                  |
                                  | Escherichia_coli_K12
                                  |
                                  , Shigella_flexneri_2a_2457T
                                  |
                                  | Shigella_flexneri_2a_301
                                  |
                                  , Salmonella_enterica
                                  |
                                  | Salmonella_typhi
                                  |
                                  | Salmonella_typhimurium
                                  |
                                  , Yersinia_pestis_Medievalis
                                  |
                                  , Yersinia_pestis_KIM
                                  |
                                 ,| Yersinia_pestis_CO92
                                 ||
                                 || Photorhabdus_luminescens
                                 ||
                                 ||   ___ Blochmannia_floridanus
                                 || ,|
                                 || ||____ Wigglesworthia_brevipalpis
                                 ||_|
```

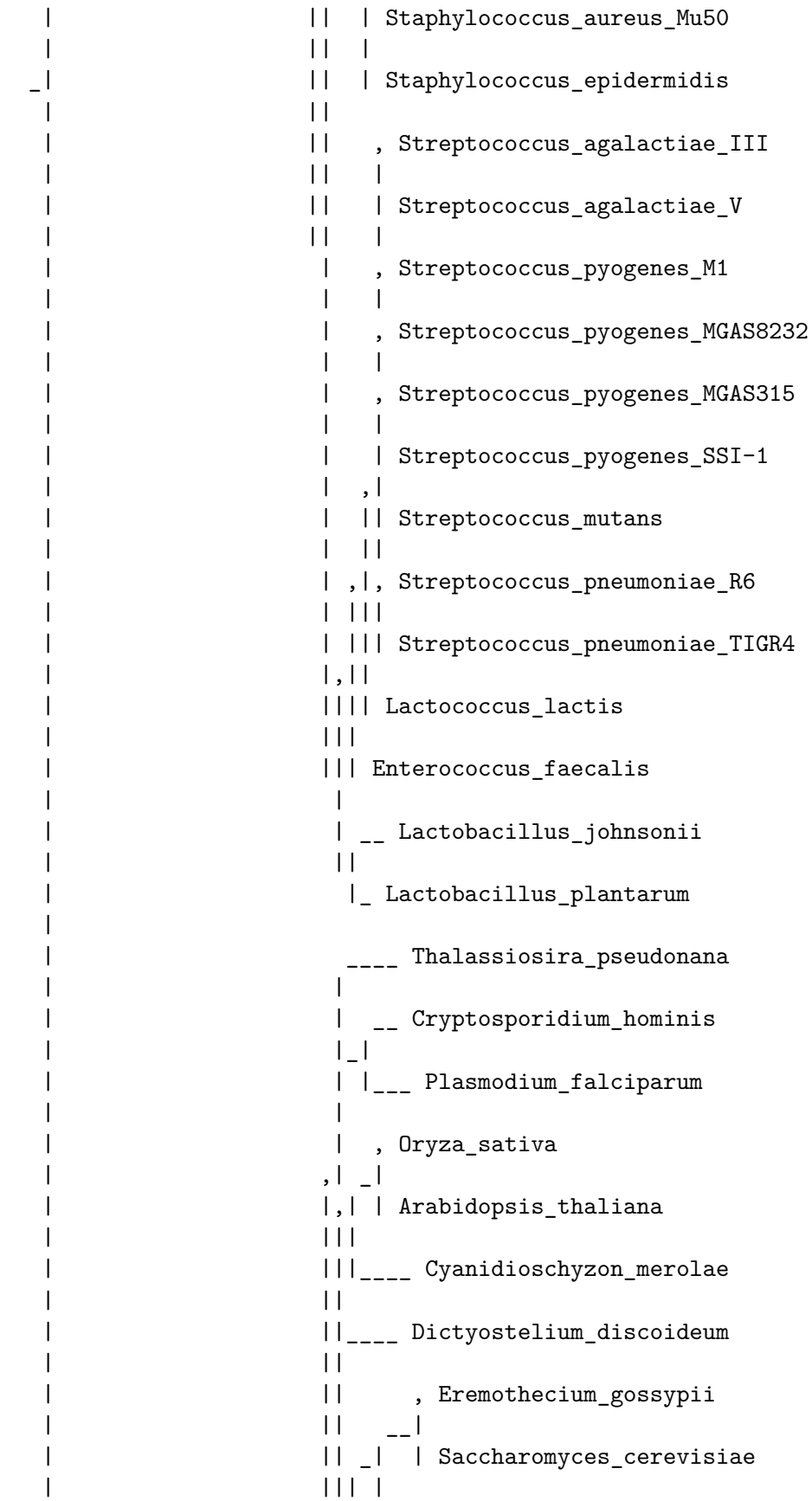

Buchnera_aphidicola_Bp
Buchnera_aphidicola_APS
Buchnera_aphidicola_Sg
Pasteurella_multocida
Haemophilus_influenzae
Haemophilus_ducreyi
Vibrio_vulnificus_YJ016
Vibrio_vulnificus_CMCP6
Vibrio_parahaemolyticus
Vibrio_cholerae
Photobacterium_profundum
Shewanella_oneidensis
Pseudomonas_putida
Pseudomonas_syringae
Pseudomonas_aeruginosa
Xylella_fastidiosa_700964
Xylella_fastidiosa_9a5c
Xanthomonas_axonopodis
Xanthomonas_campestris
Coxiella_burnetii
Neisseria_meningitidis_A
Neisseria_meningitidis_B
Chromobacterium_violaceum
Bordetella_pertussis

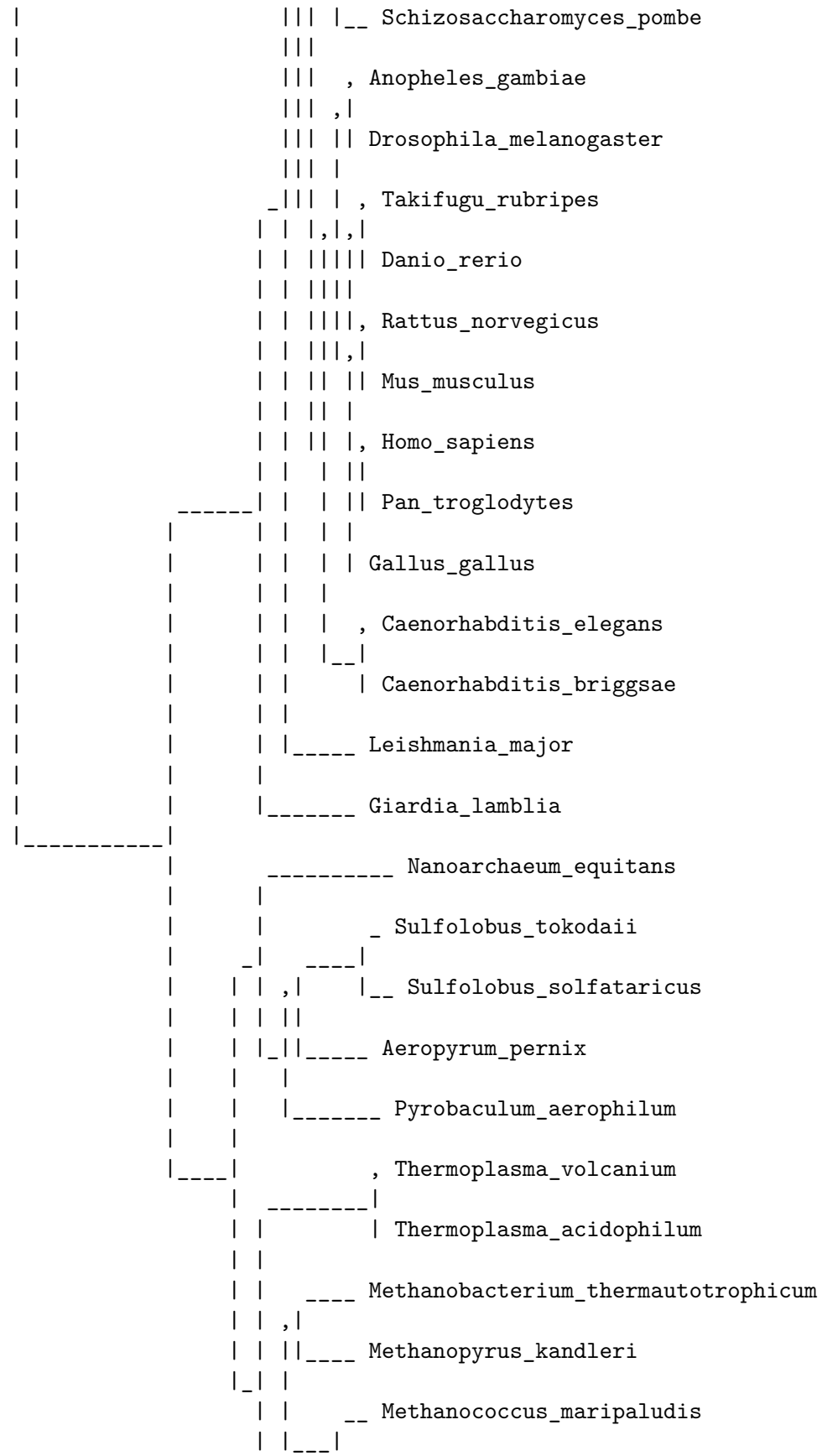

Bordetella_parapertussis
Bordetella_bronchiseptica
Ralstonia_solanacearum
Nitrosomonas_europaea
Agrobacterium_tumefaciens_Cereon
Agrobacterium_tumefaciens_WashU
Rhizobium_meliloti
Brucella_suis
Brucella_melitensis
Rhizobium_loti
Rhodopseudomonas_palustris
Bradyrhizobium_japonicum
Caulobacter_crescentus
Wolbachia_sp._wMel
Rickettsia_prowazekii
Rickettsia_conorii
Helicobacter_pylori_J99
Helicobacter_pylori_26695
Helicobacter_hepaticus
Wolinella_succinogenes
Campylobacter_jejuni
Desulfovibrio_vulgaris
Geobacter_sulfurreducens
Bdellovibrio_bacteriovorus


Acidobacterium_capsulatum
Solibacter_usitatus
Fusobacterium_nucleatum
Aquifex_aeolicus
Thermotoga_maritima
Thermus_thermophilus
Deinococcus_radiodurans
Dehalococcoides_ethenogenes
Nostoc_sp._PCC_7120
Synechocystis_sp._PCC6803
Synechococcus_elongatus
Synechococcus_sp._WH8102
Prochlorococcus_marinus_MIT9313
Prochlorococcus_marinus_SS120
Prochlorococcus_marinus_CCMP1378
Gloeobacter_violaceus
Gemmata_obscuriglobus
Rhodopirellula_baltica
Leptospira_interrogans_L1-130
Leptospira_interrogans_56601
Treponema_pallidum
Treponema_denticola
Borrelia_burgdorferi
Tropheryma_whipplei_TW08/27


Tropheryma_whipplei_Twist
Bifidobacterium_longum
Corynebacterium_glutamicum_13032
Corynebacterium_glutamicum
Corynebacterium_efficiens
Corynebacterium_diphtheriae
Mycobacterium_bovis
Mycobacterium_tuberculosis_CDC1551
Mycobacterium_tuberculosis_H37Rv
Mycobacterium_leprae
Mycobacterium_paratuberculosis
Streptomyces_avermitilis
Streptomyces_coelicolor
Fibrobacter_succinogenes
Chlorobium_tepidum
Porphyromonas_gingivalis
Bacteroides_thetaiotaomicron
Chlamydophila_pneumoniae_TW183
Chlamydia_pneumoniae_J138
Chlamydia_pneumoniae_CWL029
Chlamydia_pneumoniae_AR39
Chlamydophila_caviae
Chlamydia_muridarum
Chlamydia_trachomatis


Thermoanaerobacter_tengcongensis
Clostridium_tetani
Clostridium_perfringens
Clostridium_acetobutylicum
Mycoplasma_mobile
Mycoplasma_pulmonis
Mycoplasma_pneumoniae
Mycoplasma_genitalium
Mycoplasma_gallisepticum
Mycoplasma_penetrans
Ureaplasma_parvum
Mycoplasma_mycoides
Phytoplasma_Onion_yellows
Listeria_monocytogenes_F2365
Listeria_monocytogenes_EGD
Listeria_innocua
Oceanobacillus_iheyensis
Bacillus_halodurans
Bacillus_cereus_ATCC_14579
Bacillus_cereus_ATCC_10987
Bacillus_anthracis
Bacillus_subtilis
Staphylococcus_aureus_MW2
Staphylococcus_aureus_N315

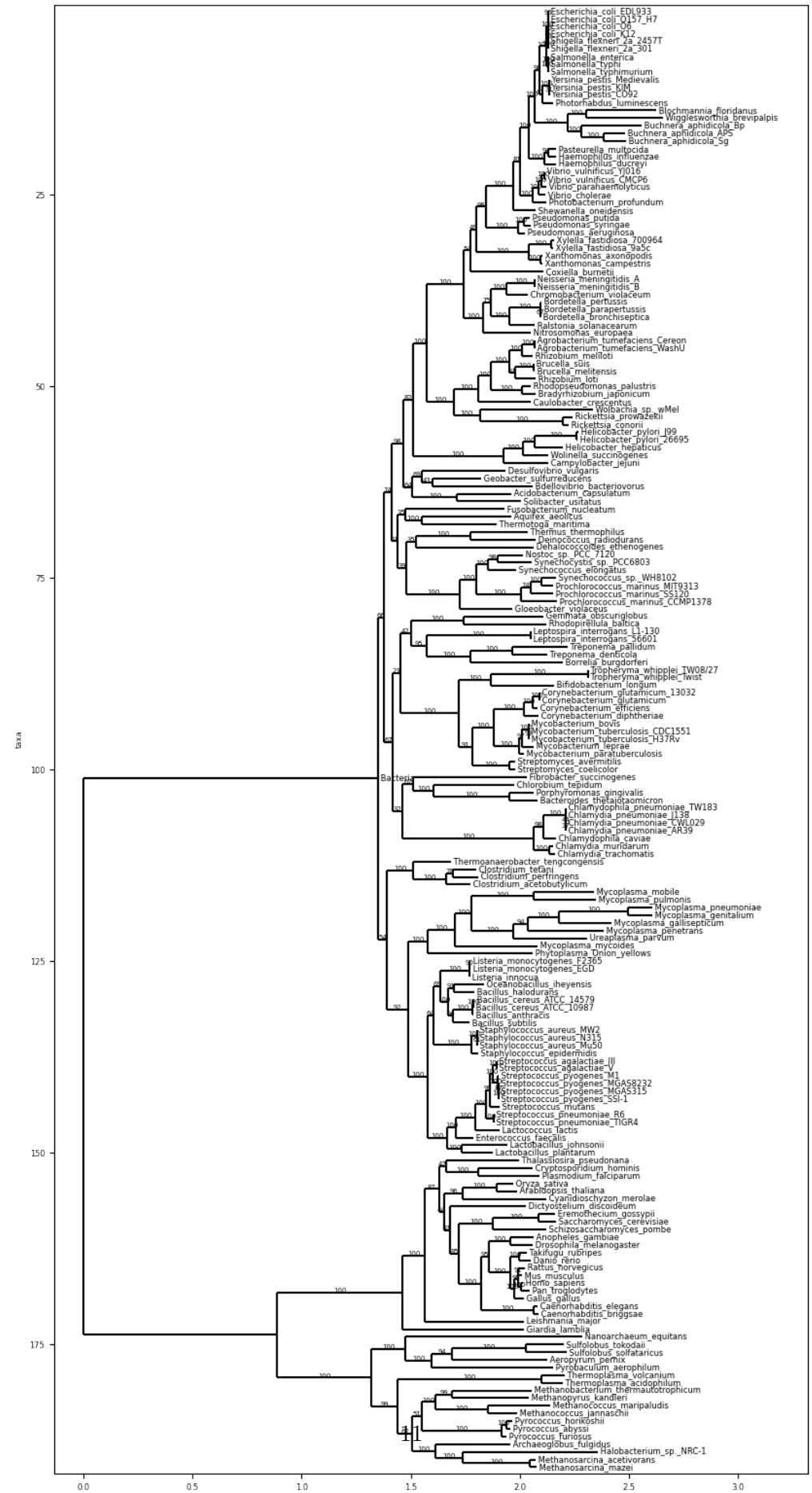

Staphylococcus_aureus_Mu50
Staphylococcus_epidermidis
Streptococcus_agalactiae_III
Streptococcus_agalactiae_V
Streptococcus_pyogenes_M1
Streptococcus_pyogenes_MGAS8232
Streptococcus_pyogenes_MGAS315
Streptococcus_pyogenes_SSI-1
Streptococcus_mutans
Streptococcus_pneumoniae_R6
Streptococcus_pneumoniae_TIGR4
Lactococcus_lactis
Enterococcus_faecalis
Lactobacillus_johnsonii
Lactobacillus_plantarum
Thalassiosira_pseudonana
Cryptosporidium_hominis
Plasmodium_falciparum
Oryza_sativa
Arabidopsis_thaliana
Cyanidioschyzon_merolae
Dictyostelium_discoideum
Eremothecium_gossypii
Saccharomyces_cerevisiae


Schizosaccharomyces_pombe
Anopheles_gambiae
Drosophila_melanogaster
Takifugu_rubripes
Danio_rerio
Rattus_norvegicus
Mus_musculus
Homo_sapiens
Pan_troglodytes
Gallus_gallus
Caenorhabditis_elegans
Caenorhabditis_briggsae
Leishmania_major
Giardia_lamblia
Nanoarchaeum_equitans
Sulfolobus_tokodaii
Sulfolobus_solfataricus
Aeropyrum_pernix
Pyrobaculum_aerophilum
Thermoplasma_volcanium
Thermoplasma_acidophilum
Methanobacterium_thermautotrophicum
Methanopyrus_kandleri
Methanococcus_maripaludis

```
|,|   | Methanococcus_jannaschii
|||
|||    , Pyrococcus_horikoshii
|||   ,|
|||___|| Pyrococcus_abyssi
 |    |
 |    | Pyrococcus_furiosus
 |
 | ____ Archaeoglobus_fulgidus
 ||
  |  _______ Halobacterium_sp._NRC-1
  |_|
    |   , Methanosarcina_acetivorans
    |___|
        | Methanosarcina_mazei
```

0.1.3 1.3 draw the tree with draw

[5]:
```
plot_tree(tree, show=True)
```


Escherichia_coli_EDL933
Escherichia_coli_O157_H7
Escherichia_coli_O6
Escherichia_coli_K12
Shigella_flexneri_2a_2457T
Shigella_flexneri_2a_301
Salmonella_enterica
Salmonella_typhi
Salmonella_typhimurium
Yersinia_pestis_Medievalis
Yersinia_pestis_KIM
Yersinia_pestis_CO92
Photorhabdus_luminescens
Blochmannia_floridanus
Wigglesworthia_brevipalpis
Buchnera_aphidicola_Bp
Buchnera_aphidicola_APS
Buchnera_aphidicola_Sg
Pasteurella_multocida
Haemophilus_influenzae
Haemophilus_ducreyi
Vibrio_vulnificus_YJ016
Vibrio_vulnificus_CMCP6
Vibrio_parahaemolyticus
Vibrio_cholerae
Photobacterium_profundum
Shewanella_oneidensis
Pseudomonas_putida
Pseudomonas_syringae
Pseudomonas_aeruginosa
Xylella_fastidiosa_700964
Xylella_fastidiosa_9a5c
Xanthomonas_axonopodis
Xanthomonas_campestris
Coxiella_burnetii
Neisseria_meningitidis_A
Neisseria_meningitidis_B
Chromobacterium_violaceum
Bordetella_pertussis
Bordetella_parapertussis
Bordetella_bronchiseptica
Ralstonia_solanacearum
Nitrosomonas_europaea
Agrobacterium_tumefaciens_Cereon
Agrobacterium_tumefaciens_WashU
Rhizobium_meliloti
Brucella_suis
Brucella_melitensis
Rhizobium_loti
Rhodopseudomonas_palustris
Bradyrhizobium_japonicum
Caulobacter_crescentus
Wolbachia_sp._wMel
Rickettsia_prowazekii
Rickettsia_conorii
Helicobacter_pylori_J99
Helicobacter_pylori_26695
Helicobacter_hepaticus
Wolinella_succinogenes
Campylobacter_jejuni
Desulfovibrio_vulgaris
Geobacter_sulfurreducens
Bdellovibrio_bacteriovorus
Acidobacterium_capsulatum
Solibacter_usitatus
Fusobacterium_nucleatum
Aquifex_aeolicus
Thermotoga_maritima
Thermus_thermophilus
Deinococcus_radiodurans
Dehalococcoides_ethenogenes
Nostoc_sp._PCC_7120
Synechocystis_sp._PCC6803
Synechococcus_elongatus
Synechococcus_sp._WH8102
Prochlorococcus_marinus_MIT9313
Prochlorococcus_marinus_SS120
Prochlorococcus_marinus_CCMP1378
Gloeobacter_violaceus
Gemmata_obscuriglobus
Rhodopirellula_baltica
Leptospira_interrogans_L1-130
Leptospira_interrogans_56601
Treponema_pallidum
Treponema_denticola
Borrelia_burgdorferi
Tropheryma_whipplei_TW08/27
Tropheryma_whipplei_Twist
Bifidobacterium_longum
Corynebacterium_glutamicum_13032
Corynebacterium_glutamicum
Corynebacterium_efficiens
Corynebacterium_diphtheriae
Mycobacterium_bovis
Mycobacterium_tuberculosis_CDC1551
Mycobacterium_tuberculosis_H37Rv
Mycobacterium_leprae
Mycobacterium_paratuberculosis
Streptomyces_avermitilis
Streptomyces_coelicolor
Bacteria
Fibrobacter_succinogenes
Chlorobium_tepidum
Porphyromonas_gingivalis
Bacteroides_thetaiotaomicron
Chlamydophila_pneumoniae_TW183
Chlamydia_pneumoniae_J138
Chlamydia_pneumoniae_CWL029
Chlamydia_pneumoniae_AR39
Chlamydophila_caviae
Chlamydia_muridarum
Chlamydia_trachomatis
Thermoanaerobacter_tengcongensis
Clostridium_tetani
Clostridium_perfringens
Clostridium_acetobutylicum
Mycoplasma_mobile
Mycoplasma_pulmonis
Mycoplasma_pneumoniae
Mycoplasma_genitalium
Mycoplasma_gallisepticum
Mycoplasma_penetrans
Ureaplasma_parvum
Mycoplasma_mycoides
Phytoplasma_Onion_yellows
Listeria_monocytogenes_F2365
Listeria_monocytogenes_EGD
Listeria_innocua
Oceanobacillus_iheyensis
Bacillus_halodurans
Bacillus_cereus_ATCC_14579
Bacillus_cereus_ATCC_10987
Bacillus_anthracis
Bacillus_subtilis
Staphylococcus_aureus_MW2
Staphylococcus_aureus_N315
Staphylococcus_aureus_Mu50
Staphylococcus_epidermidis
Streptococcus_agalactiae_III
Streptococcus_agalactiae_V
Streptococcus_pyogenes_M1
Streptococcus_pyogenes_MGAS8232
Streptococcus_pyogenes_MGAS315
Streptococcus_pyogenes_SSI-1
Streptococcus_mutans
Streptococcus_pneumoniae_R6
Streptococcus_pneumoniae_TIGR4
Lactococcus_lactis
Enterococcus_faecalis
Lactobacillus_johnsonii
Lactobacillus_plantarum
Thalassiosira_pseudonana
Cryptosporidium_hominis
Plasmodium_falciparum
Oryza_sativa
Arabidopsis_thaliana
Cyanidioschyzon_merolae
Dictyostelium_discoideum
Eremothecium_gossypii
Saccharomyces_cerevisiae
Schizosaccharomyces_pombe
Anopheles_gambiae
Drosophila_melanogaster
Takifugu_rubripes
Danio_rerio
Rattus_norvegicus
Mus_musculus
Homo_sapiens
Pan_troglodytes
Gallus_gallus
Caenorhabditis_elegans
Caenorhabditis_briggsae
Leishmania_major
Giardia_lamblia
Nanoarchaeum_equitans
Sulfolobus_tokodaii
Sulfolobus_solfataricus
Aeropyrum_pernix
Pyrobaculum_aerophilum
Thermoplasma_volcanium
Thermoplasma_acidophilum
Methanobacterium_thermautotrophicum
Methanopyrus_kandleri
Methanococcus_maripaludis
Methanococcus_jannaschii
Pyrococcus_horikoshii
Pyrococcus_abyssi
Pyrococcus_furiosus
Archaeoglobus_fulgidus
Halobacterium_sp._NRC-1
Methanosarcina_acetivorans
Methanosarcina_mazei
taxa
branch length
0.0
0.5
1.0
1.5
2.0
2.5
3.0
25
50
75
100
125
150
175

0.1.4 1.4 saves the tree image in raster format (png) and vector (svg / pdf) (you can use pylab.savefig, for example) (pictures sendagain);

[6]:
```
plot_tree(tree, out_file='tree.png')
plot_tree(tree, out_file='tree.svg')
```

0.1.5 1.5 change the format to phyloxml and writes in a file

[7]:
```
Phylo.write(tree, "tree.xml", "phyloxml")
```

[7]: 1

0.2 PART 1. ETE

0.2.1 1.1 read the same tree using ETE

[8]:
```
tree = get_tree('https://www.jasondavies.com/tree-of-life/life.txt', ete=True)
```

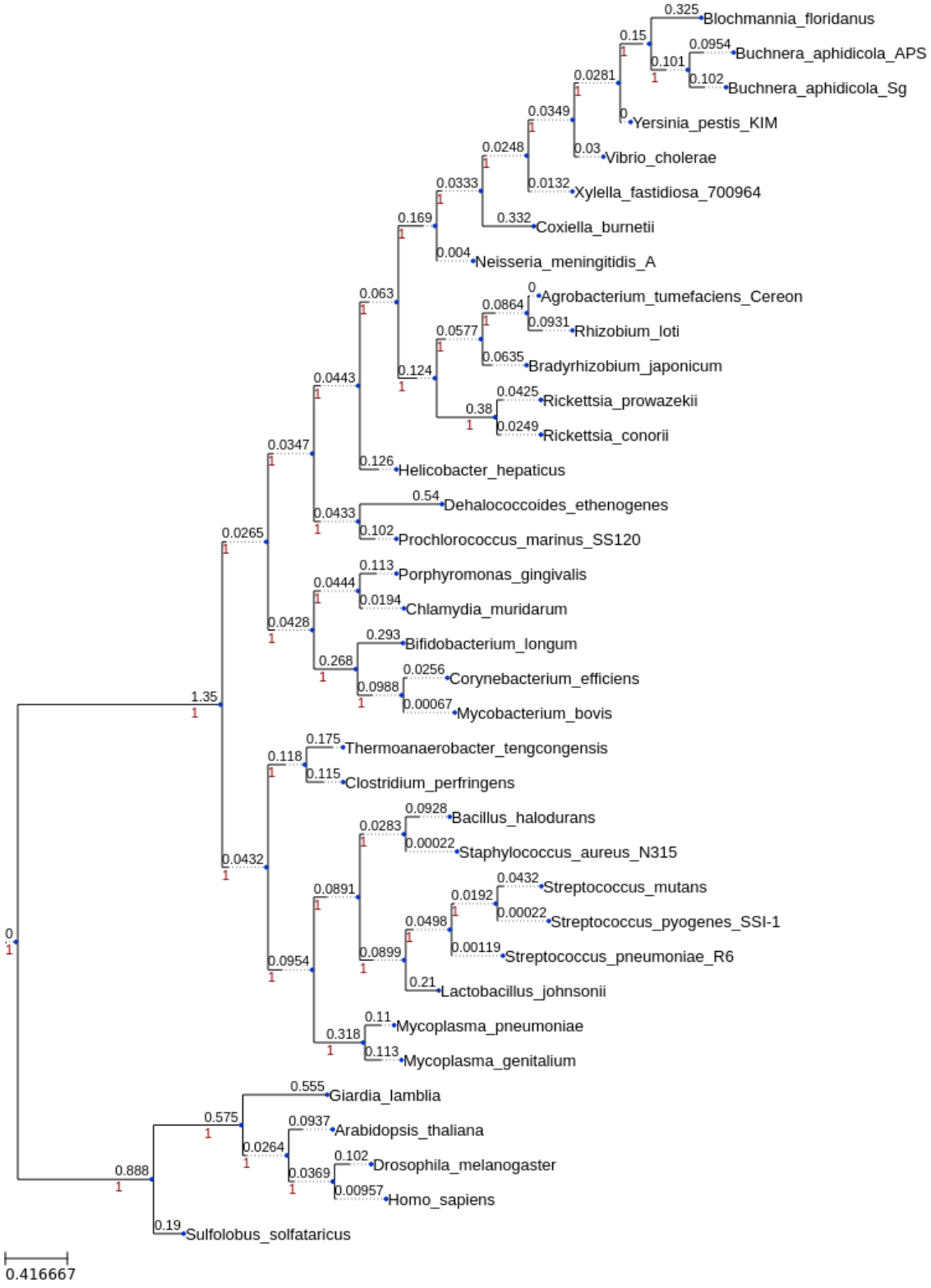
0.2.2 1.2 cut from a tree a random set of 42 (or other number) leaves. Use the "prune" function.

[9]:
```
leaves = tree.get_leaves()
tree.prune([leaves[randint(0, len(leaves) + 1)].name for _ in range(42)])
```

0.2.3 1.3 draw the pruned tree

[10]:
```
ts = TreeStyle()
ts.show_leaf_name = True
ts.show_branch_length = True
ts.show_branch_support = True
ts.scale =  120
tree.render(file_name='%%inline', tree_style=ts)
```

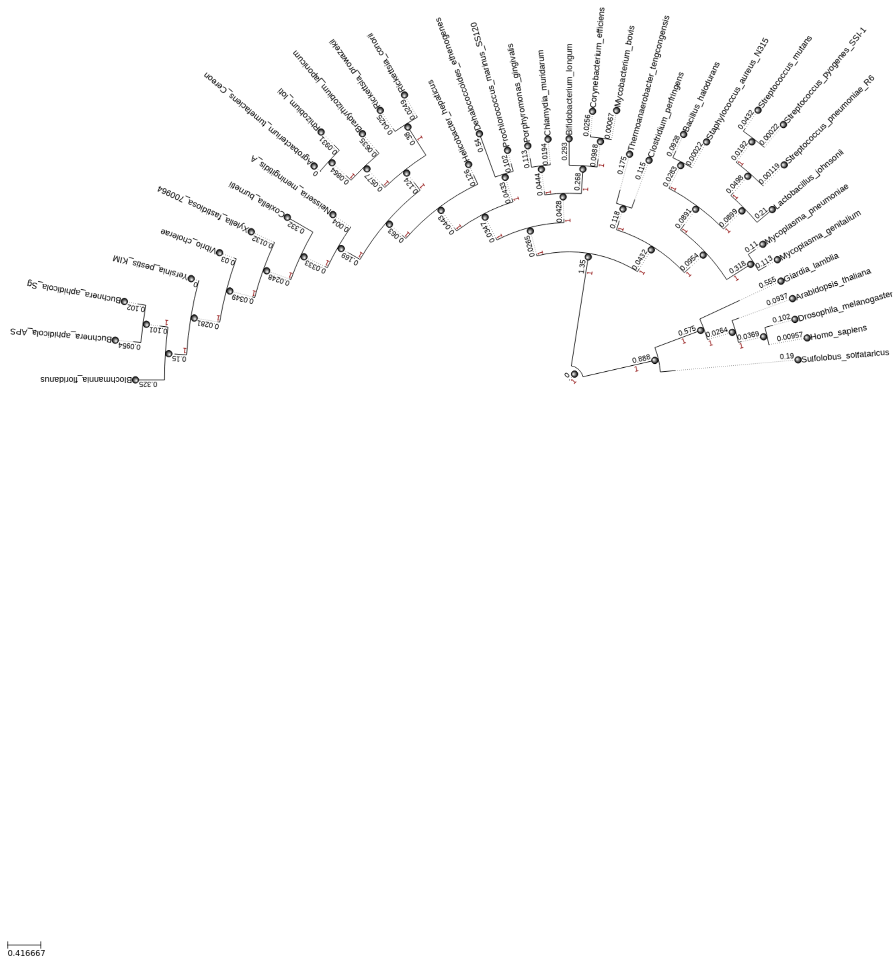
[10]:


Blochmannia_floridanus
Buchnera_aphidicola_APS
Buchnera_aphidicola_Sg
Yersinia_pestis_KIM
Vibrio_cholerae
Xylella_fastidiosa_700964
Coxiella_burnetii
Neisseria_meningitidis_A
Agrobacterium_tumefaciens_Cereon
Rhizobium_loti
Bradyrhizobium_japonicum
Rickettsia_prowazekii
Rickettsia_conorii
Helicobacter_hepaticus
Dehalococcoides_ethenogenes
Prochlorococcus_marinus_SS120
Porphyromonas_gingivalis
Chlamydia_muridarum
Bifidobacterium_longum
Corynebacterium_efficiens
Mycobacterium_bovis
Thermoanaerobacter_tengcongensis
Clostridium_perfringens
Bacillus_halodurans
Staphylococcus_aureus_N315
Streptococcus_mutans
Streptococcus_pyogenes_SSI-1
Streptococcus_pneumoniae_R6
Lactobacillus_johnsonii
Mycoplasma_pneumoniae
Mycoplasma_genitalium
Giardia_lamblia
Arabidopsis_thaliana
Drosophila_melanogaster
Homo_sapiens
Sulfolobus_solfataricus
0.416667

0.2.4 1.4 Draws the pruned tree with additional aesthetic processing

[11]:
```
ts = TreeStyle()
ts.show_leaf_name = True
ts.show_branch_length = True
ts.show_branch_support = True
ts.scale =  120
ts.mode = "c"
ts.arc_start = -180
ts.arc_span = 180
nstyle = NodeStyle()
nstyle["shape"] = "sphere"
nstyle["size"] = 10
nstyle["fgcolor"] = "black"
for n in tree.traverse():
    n.set_style(nstyle)
tree.render(file_name='%%inline', tree_style=ts)
```

[11]:



0.3 PART 2

0.3.1 2.1 queries the base of nucleotide sequences for all sequences according to the name of the gene MKI67 for the organism Homo sapiens and returns xml

[12]:
```
gene = entrez_gene('homo[ORGN] MKI67')
```

```
/home/marina/anaconda3/envs/biopy/lib/python3.8/site-
packages/Bio/Entrez/__init__.py:656: UserWarning:
Email address is not specified.

To make use of NCBI's E-utilities, NCBI requires you to specify your
```

```
email address with each request.  As an example, if your email address
is A.N.Other@example.com, you can specify it as follows:
   from Bio import Entrez
   Entrez.email = 'A.N.Other@example.com'
In case of excessive usage of the E-utilities, NCBI will attempt to contact
a user at the email address provided before blocking access to the
E-utilities.
  warnings.warn(
```

[13]:
```
gene
```

[13]:
```
{'Count': '49', 'RetMax': '20', 'RetStart': '0', 'IdList': ['1890263052',
'1675069958', '1519315735', '1519246095', '1519245506', '1519243472',
'1435213226', '1034568323', '1034568322', '568815596', '568815595', '568815593',
'568815588', '568815586', '568815583', '568815579', '1701945985', '1701108622',
'1700660549', '1026191091'], 'TranslationSet': [{'From': 'homo[ORGN]', 'To':
'"Homo"[Organism]'}], 'TranslationStack': [{'Term': '"Homo"[Organism]', 'Field':
'Organism', 'Count': '27678340', 'Explode': 'Y'}, {'Term': 'MKI67[All Fields]',
'Field': 'All Fields', 'Count': '3226', 'Explode': 'N'}, 'AND'],
'QueryTranslation': '"Homo"[Organism] AND MKI67[All Fields]'}
```

0.3.2 2.2 return a table with UID (in XML this field is called Id), accession number (in XML this field is called Caption), sequence length (Slen);

[14]:
```
f"Let's run this one using gene id {gene['IdList'][0]}"
```

[14]:
```
"Let's run this one using gene id 1890263052"
```

[15]:
```
entrez_summary(gene['IdList'][0])
```

[15]:

	Id	Caption	Length
1890263052	1890263052	NM_001172425	1917

0.3.3 2.3 return the nucleotide sequences in fasta format and writes to the file

[16]:
```
fasta = get_fasta(gene['IdList'][0])
```

```
/home/marina/anaconda3/envs/biopy/lib/python3.8/site-
packages/Bio/Entrez/__init__.py:656: UserWarning:
Email address is not specified.

To make use of NCBI's E-utilities, NCBI requires you to specify your
email address with each request.  As an example, if your email address
is A.N.Other@example.com, you can specify it as follows:
   from Bio import Entrez
   Entrez.email = 'A.N.Other@example.com'
In case of excessive usage of the E-utilities, NCBI will attempt to contact
```

```
a user at the email address provided before blocking access to the
E-utilities.
  warnings.warn(
```

[17]:
```
with open(f"{gene['IdList'][0]}.fasta", 'w') as out_f:
    out_f.write(fasta)
```

0.3.4 2.4 Download all sequences from the paper with a given PMID: 12890024

[18]:
```
seq = Entrez.efetch(db="nucleotide", id="12890024", rettype='fasta',␣
↪retmode="fasta").read()
```

[19]:
```
seq
```

[19]:
```
'>AZ769660.1 1M0570P06R Mouse 10kb plasmid UUGC1M library Mus musculus genomic
clone UUGC1M0570P06 R, genomic survey sequence\nTCTGGCTCGTTCCTCTGAAAACAAGGATTGCA
CAGAGTCATTTTTAAAGAATCTATTCATTTTTGAATTT\nTCCCTCCAATAACACCTTCAGTTCTCTCTGTACCATTTCC
CACAGNAGGAAGAAAATAGTATGTATTTGT\nCCCATTCTTCTGTGCTGTGCTCATGTGCTATGAACATGTGTGCACATA
CATGTGGAGGTGTCAGGACTCA\nGCCTCCGCCACTCTTCTAGCTTATTTAGTGAGGCAGGGTCTTCCTGCAAACCCTAG
AGCTCACCAATACA\nGCTCGTCTTGCCAGCCAGCTTGCTCTGGGAATTTCCTGTCTCTGCCTTC\n\n'
```