BIOLAWEB

Workshop

www.biolaweb.com

Acronym:	BIOLAWEB Boosting Institute of Chemistry, Technology and Metallurgy in Water Biomonitoring
Grant No:	101079234
Type of action:	HORIZON Coordination and Support Actions (HORIZON - CSA)
Starting Date:	01/10/2022
Duration:	36 months

Workshop, Belgrade, October 2023

BIOLAWEB
presentation

Funded by
the European Union

Use of phylogenies in ecology

F. Rimet

UMR Carrtel, INRAE, Thonon France

Schedule

0. Including phylogenies into ecological studies

0.1 Phylogenetic diversity

0.2 Measuring phylogenetic signal

0.3 Measuring and testing community phylogenetic structure

1. Theory of sequence alignment

1.1 What is an alignment?

1.2 Objective of aligning sequences

1.3 Edit distance

1.4 Local alignments

1.5 Global alignments

2. Phylogenies

2.1 Choose appropriate model of sequence evolution

2.2 Phylogeny inference

2.3 Graphical representation of phylogenies

3. Phylogenetic placement

3.1 Definition

3.2 Why using this

3.3 Example with RaxML in R

4. Ecophylogenetic training in R (R studio)

PD

NRI NTI

Ressources to download at:

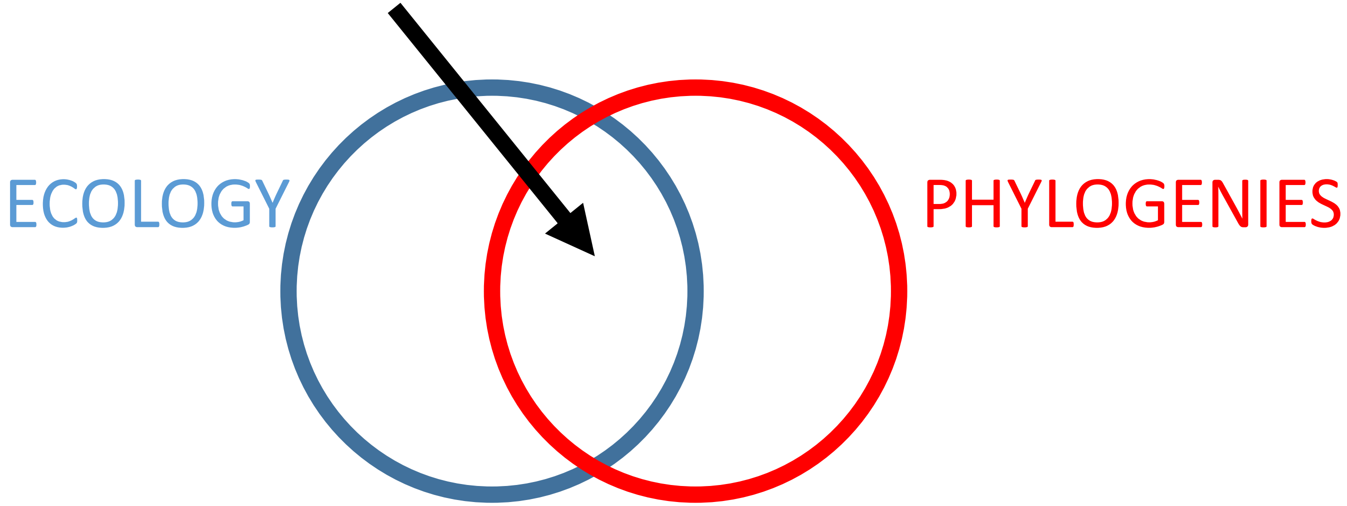
https://filesender.renater.fr/?s=download&token=2bf3be86-0b5c-4fde-8ace-852d4ae56853

0. Including phylogenies into ecological studies

- Central question in Community Assembly and Species Coexistence
  - Why do species occur at particular places?
  - Why do some pairs of species coexist while others not?
- There are 2 main predictions:
  - **Environmental filtering**: Ecologically similar species should coexist in ecologically similar environments.
  - **Limiting similarity**: Ecologically dissimilar species should coexist because too similar species competing for the same resources cannot stably coexist.

0. Including phylogenies into ecological studies

- Including phylogeny into ecological thinking represents an opportunity for biologists because:
  - Species distributions are shaped by evolutionary and ecological processes
  - These 2 processes are intimately related
  - So, it is important to study them together
- “Ecophylogenetic” Mouquet et al. 2012 (Biological Reviews)



0. Including phylogenies into ecological studies

Examples of ecophylogenetic analyses through different types of measures:

0.1 Measure of phylogenetic diversity

e.g. Phylogenetic diversity and ecosystem functioning (Faith 1992, Cadotte et al. 2008)

0.2 Measure of phylogenetic signal

e.g. Phylogenetic signal and measure of niche conservatism (Bloomerg et al. 2003, Pagel et al. 1999…)

0.3 Measure and test of community phylogenetic structure

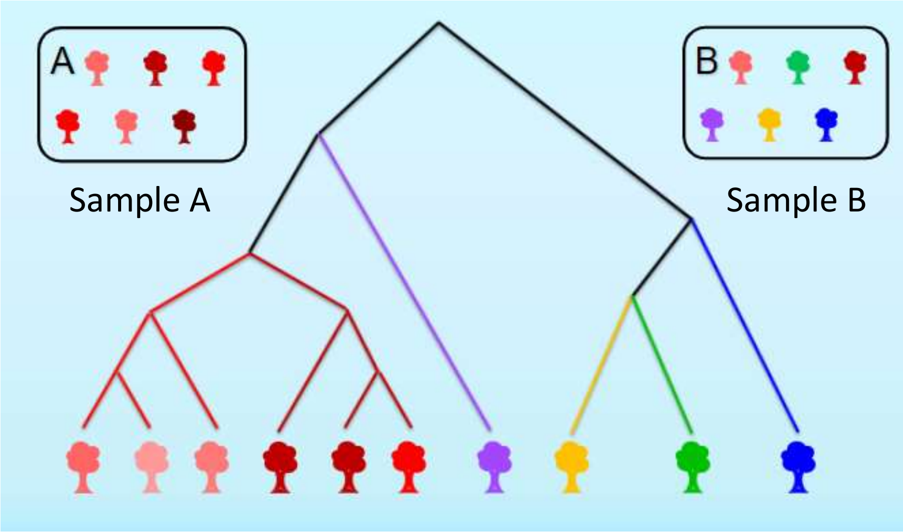
e.g. Assembly rules (environmental filtering vs competition): NTI, NRI indices (Webb et al. 2000)

0.1 Phylogenetic diversity

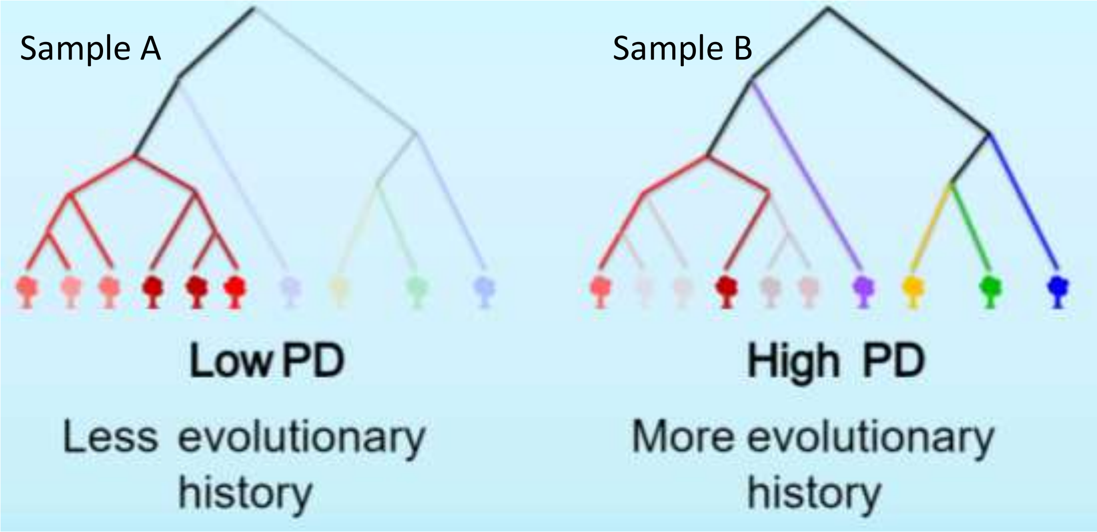
- What is phylogenetic diversity (PD)?

PD is a measure of diversity based on phylogeny

1st step: reconstruction of the phylogeny of the clade



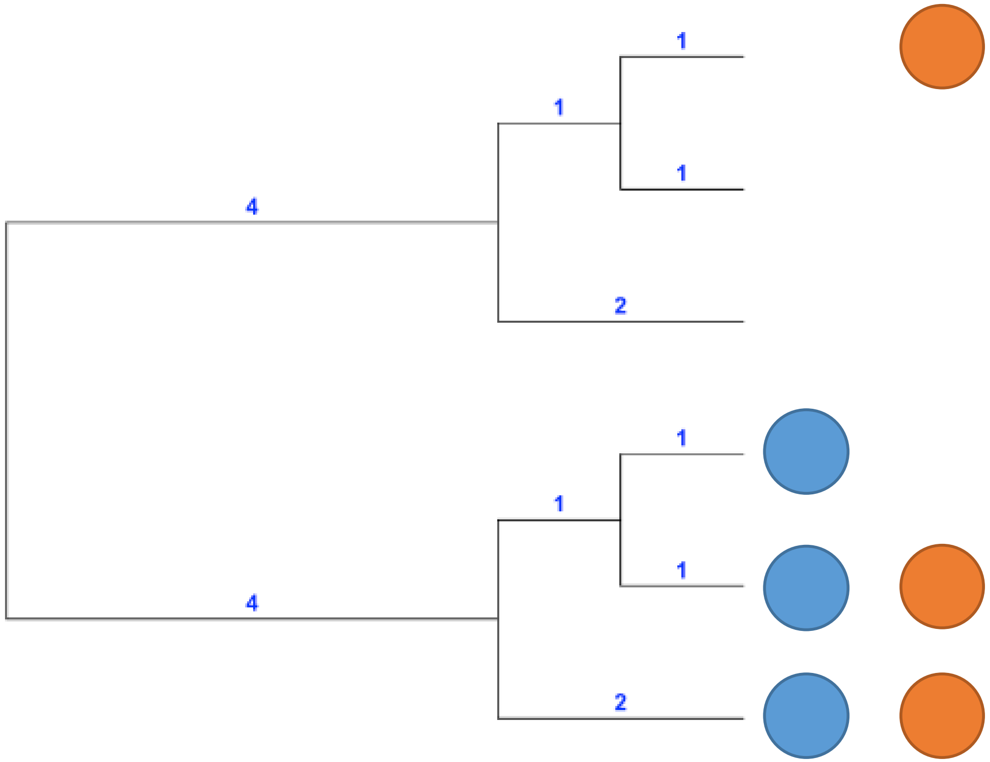
2nd step: use of the phylogenetic distances between organisms to weight the diversity metric



0.1 Phylogenetic diversity

- Faith’s index (PD)

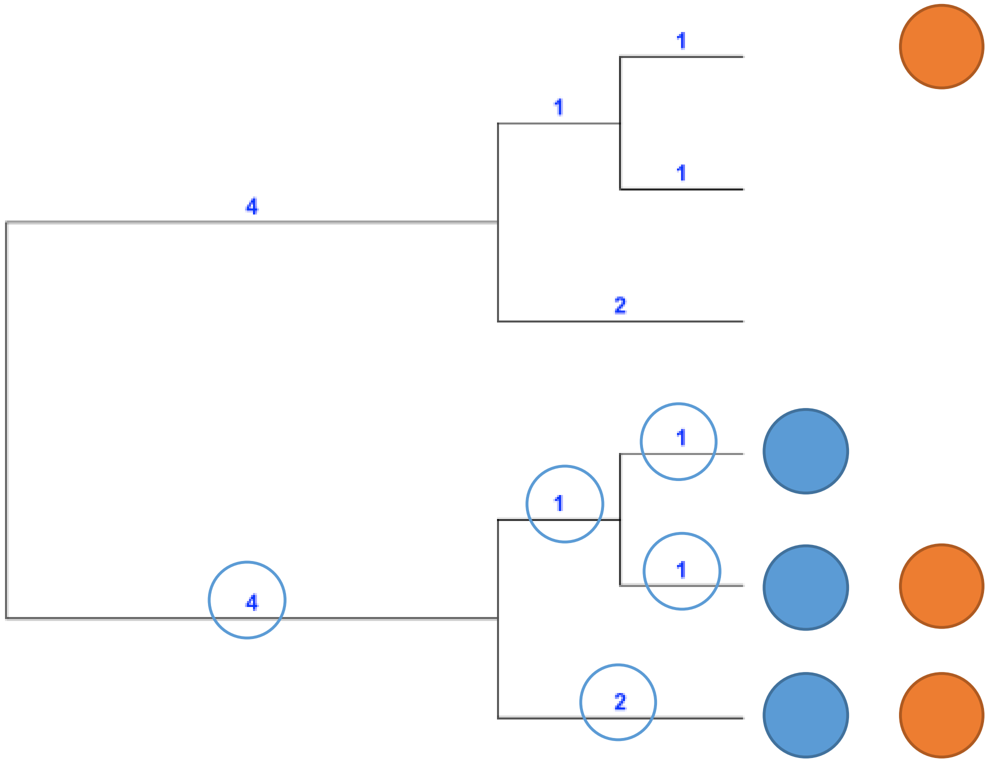
•PD = sum of the lengths of branches where species are occuring



0.1 Phylogenetic diversity

- Faith’s index (PD)

•PD = sum of the lengths of branches where species are occuring

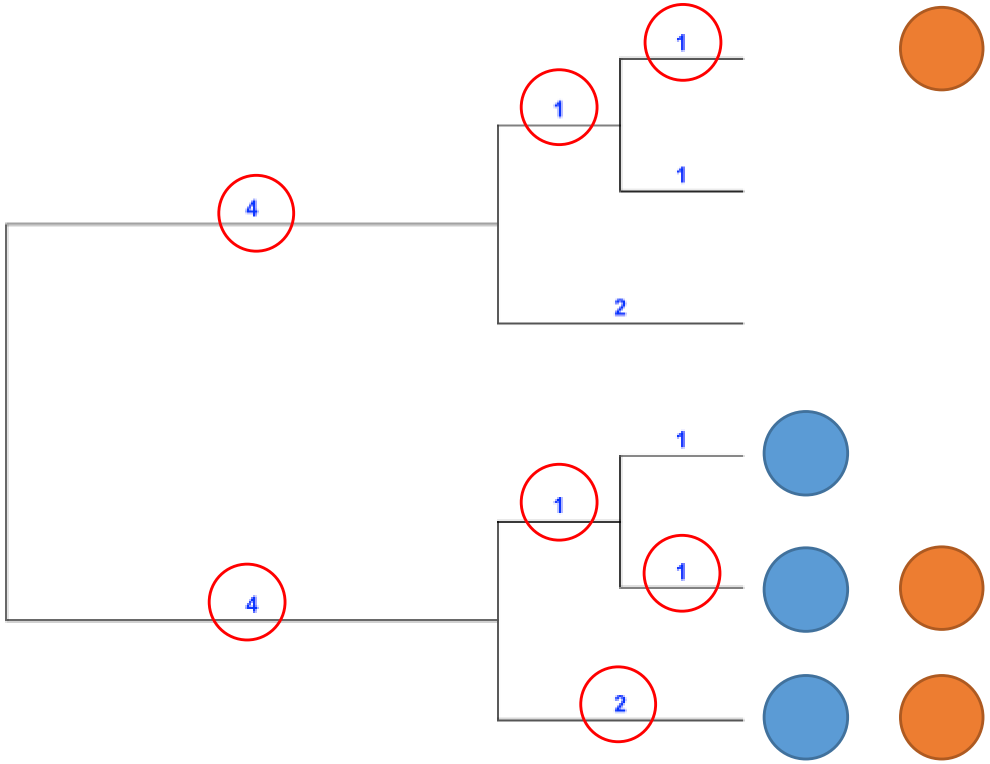


● PD = 9

0.1 Phylogenetic diversity

- Faith's index (PD)

•PD = sum of the lengths of branches where species are occuring



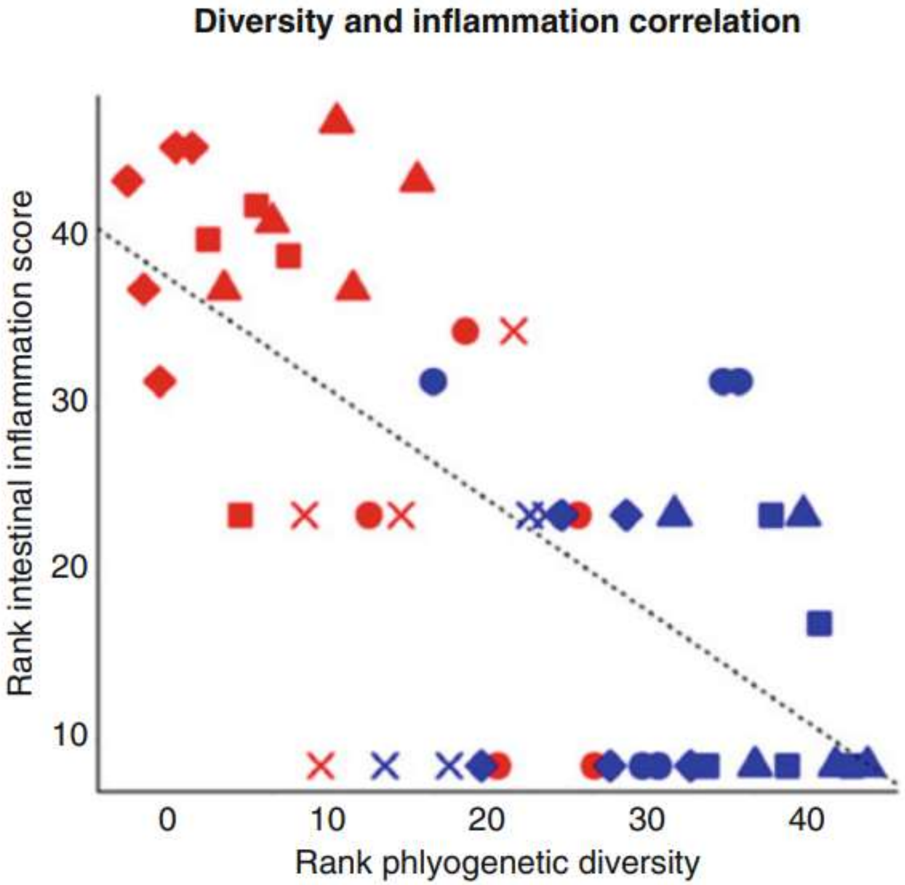
🔵 PD = 9

🟠 PD = 14

0.1 Phylogenetic diversity

- 1^{st} example: gut microbial diversity

Reduced microbial PD in the human body may indicate reduced resilience, and it is now associated with many human diseases

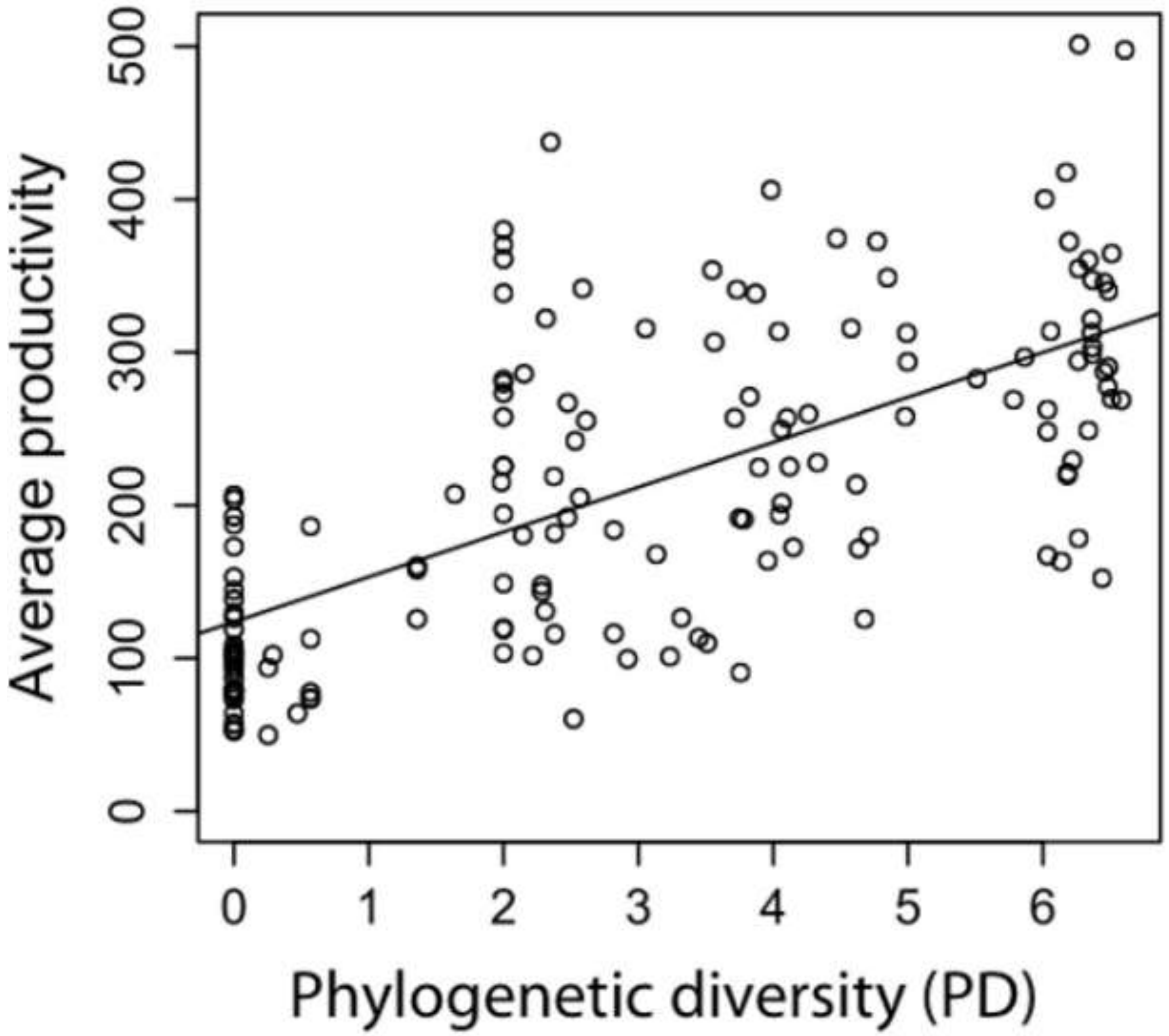


Bassett SA, Young W, Barnett MPG, Cookson AL, McNabb WC, Roy NC (2015) Changes in composition of caecal microbiota associated with increased colon inflammation in interleukin10 gene-deficient mice inoculated with Enterococcus species. Nutrients 7:1798–1816

Fig. 1.1 X axis is PD amounts and Y axis is inflammation rating. Blue points indicate less susceptible individuals and red points indicate more susceptible individuals. Shapes of the points indicate treatment groups. Overall, the plot shows that increased inflammation was associated with a decrease in caecal microbial PD. For further information, see Bassett et al. (2015). Figure reproduced from Bassett et al. (2015)

0.1 Phylogenetic diversity

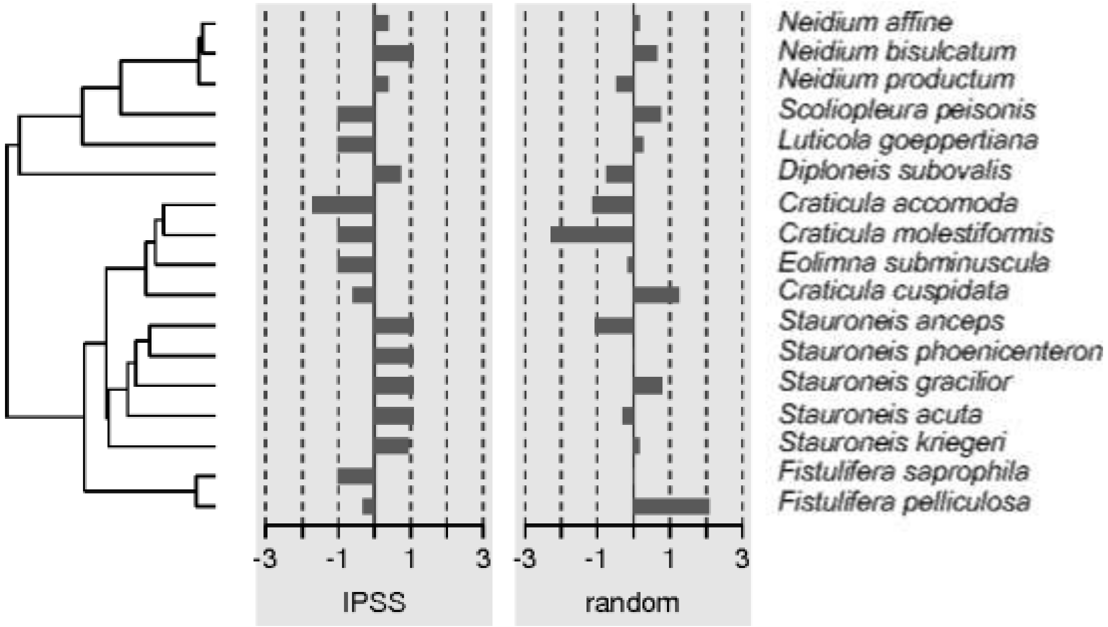
- 2nd example: ecosystem productivity and PD
- Study on a 20 years grassland monitoring (flora)
- Evolutionary relationships among species appear to explain patterns of grassland productivity.



Cadotte MW, Cavender-Bares J, Tilman D, Oakley TH (2009) Using Phylogenetic, Functional and Trait Diversity to Understand Patterns of Plant Community Productivity. PLOS ONE 4(5): e5695. https://doi.org/10.1371/journal.pone.0005695 https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0005695

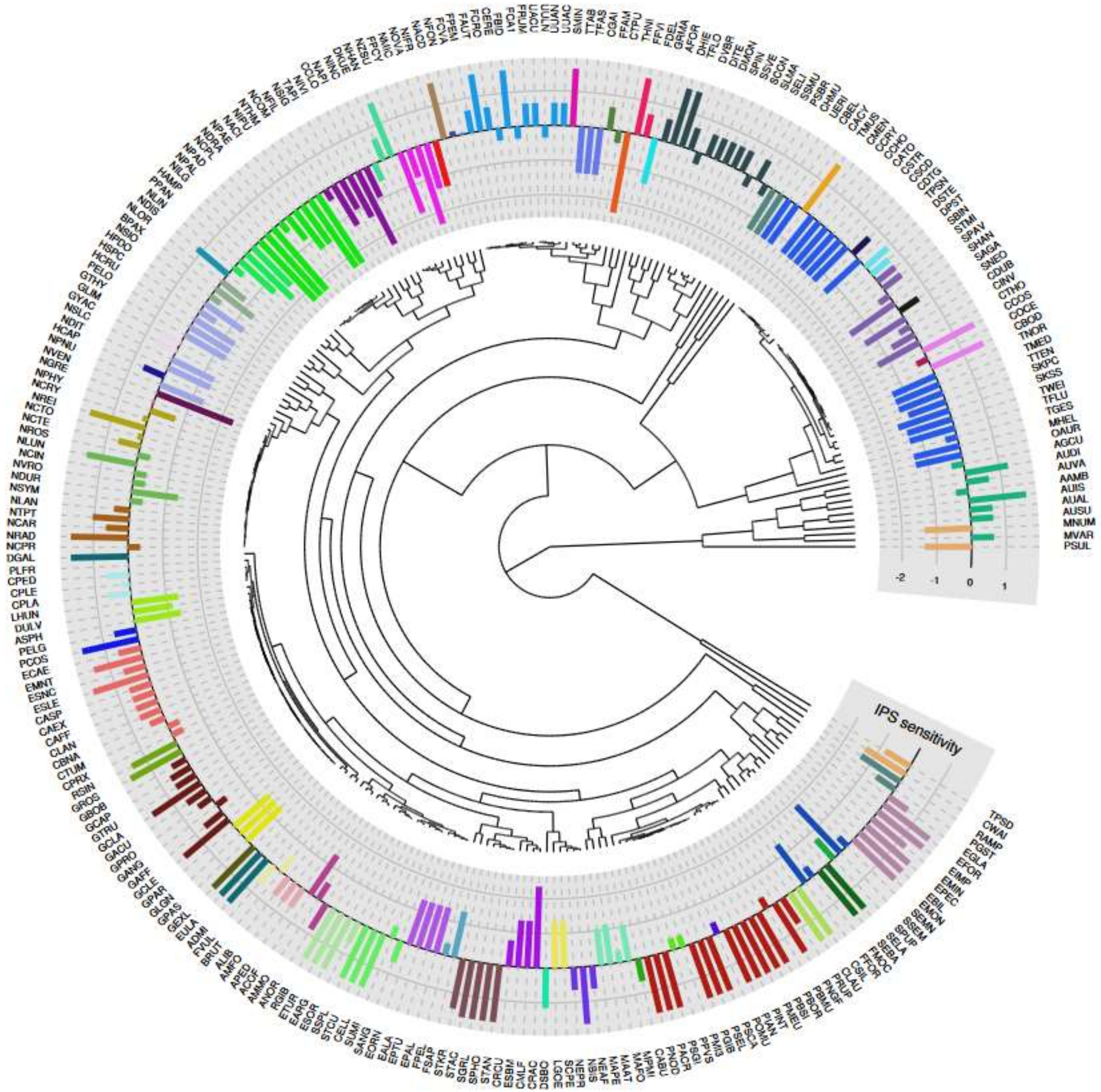
0.2 Measuring phylogenetic signal

- What is phylogenetic signal?
- It is the tendency of related species in a tree to resemble each other more than species taken randomly from the same tree. This pattern is of considerable interest in ecological and evolutionary studies (Münkemüller et al. 2012 Meth. Ecol. Evol.).



Keck, F., et al. 2016. phylosignal: an R package to measure, test, and explore the phylogenetic signal. Ecology and Evolution 6, 2774–2780. https://doi.org/10.1002/ece3.2051

- Various indices can quantifying it: Abouheif’s Cmean, Pagel’s λ, Moran’s I, Blomberg’s K.



Phylogenetic tree of 262 diatoms species and their respective IPS sensitivity value (s).

The colors delineate 68 clusters based on t = 0.6 and p = 0.1. Diatoms names are reported using 4-letter codes (Lecointe et al., 1993, see Appendix B, Section B.2.1 for corresponding Linnaean names).

Keck, F., 2016. Evaluation des liens entre phylogenie et traits écologiques chez les diatomées : pistes d’utilisation pour la bioindication des milieux aquatiques. Thèse. Université Grenoble Alpes.

0.3 Measuring and testing community phylogenetic structure

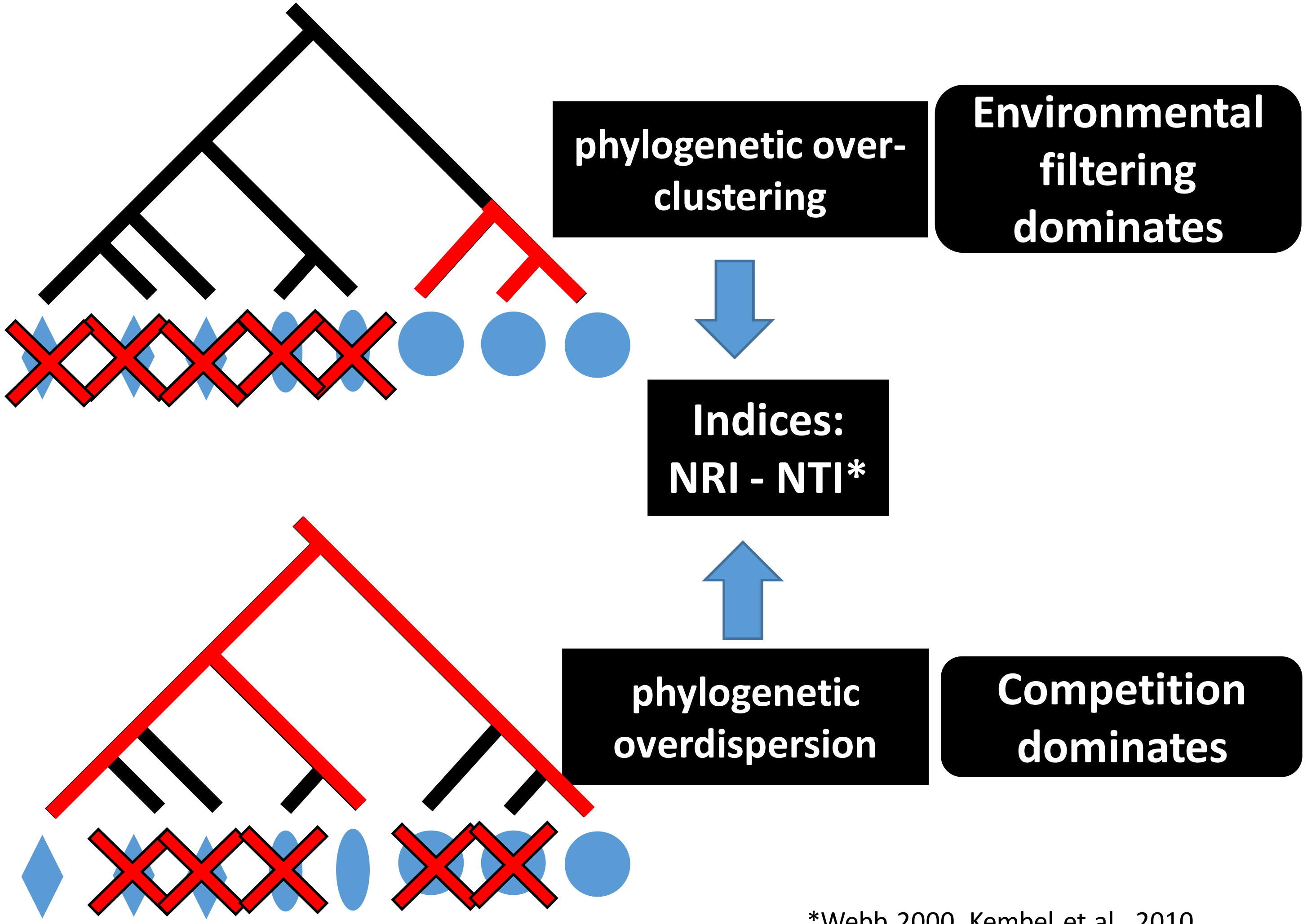
If there is a niche conservatism in the evolution,

then phylogenetic structure of samples can be interpreted in terms of ecological processes

Environmental filtering

vs

Competitive exclusion



phylogenetic over-clustering

Environmental filtering dominates

Indices: NRI - NTI*

phylogenetic overdispersion

Competition dominates

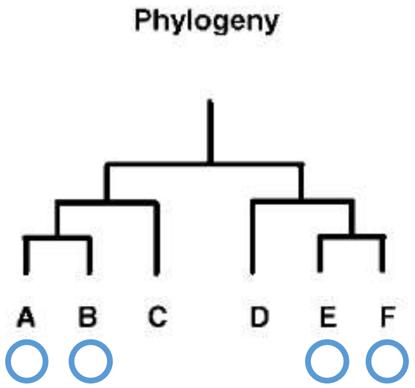
*Webb 2000, Kembel et al., 2010

NRI: Net relatedness index

Measure the mean pairwise phylogenetic distance

- Example with a 4 species community
- First: define the community with the highest pairwise distance: ABEF
- Somme of distances/nb of nodes : 22/6 = 3,66

$$\underbrace{1+5+5+5+5+1=22}_{\text{6 nodes}}$$



Greatest possible mean *pairwise* nodal distance for a community of 4 taxa (given this phylogeny) = 3.66 nodes (for A, B, E, F)

Greatest possible mean *nearest* nodal distance for a community of 4 taxa (given this phylogeny) = 2.00 nodes (for A, C, D, F)

Distance matrix

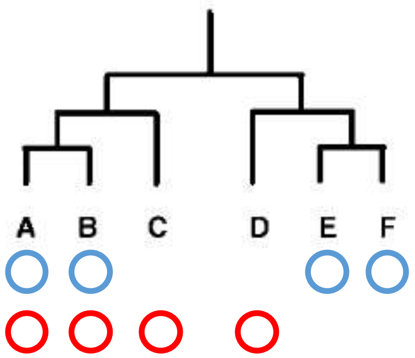
	A	B	C	D	E	F
A		1	2	4	5	5
B	1		2	4	5	5
C	2	2		3	4	4
D	4	4	3		2	2
E	5	5	4	2		1
F	5	5	4	2	1	

NRI: Net relatedness index

Measure the mean pairwise phylogenetic distance

- Example with a 4 species community
- First: define the community with the highest pairwise distance: ABEF
- Somme of distances/nb of nodes : 22/6 = 3,66
- Compare 2 communities
  - ABCD :
    - Mean pairwise distance : (1+2+2+4+4+3)/6 nodes = 2,66
    - NRI (Net index) : 1- 2,66/3,66 = 0,273
  - ABEF
    - Mean pairwise distance : (1 + 5 + 5 + 5 + 5 + 1) / 6 = 3.66
    - NRI (Net index) : 1 - (3.66 / 3.66) = 0

Phylogeny



Greatest possible mean *pairwise* nodal distance for a community of 4 taxa (given this phylogeny) = 3.66 nodes (for A, B, E, F)

Greatest possible mean *nearest* nodal distance for a community of 4 taxa (given this phylogeny) = 2.00 nodes (for A, C, D, F)

Distance matrix

	A	B	C	D	E	F
A		1	2	4	5	5
B	1		2	4	5	5
C	2	2		3	4	4
D	4	4	3		2	2
E	5	5	4	2		1
F	5	5	4	2	1	

> ABCD is less dispersed that ABEF in terms of average distances

NTI: Nearest Taxon index

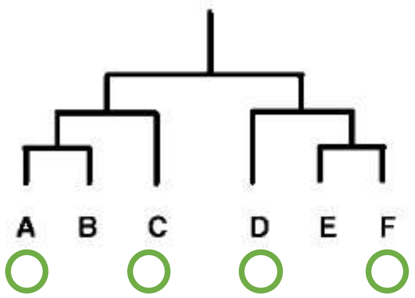
Measure the mean nearest phylogenetic neighbor

- Example with a 4 species community
- First: define the community with the greatest possible mean nearest nodal distances: ACDF

(A->C, C->A, D->F, F->D)

- Somme of distances/nb of nodes : 8/4 = 2

Phylogeny



Greatest possible mean *pairwise* nodal distance for a community of 4 taxa (given this phylogeny) = 3.66 nodes (for A, B, E, F)

Greatest possible mean *nearest* nodal distance for a community of 4 taxa (given this phylogeny) = 2.00 nodes (for A, C, D, F)

Distance matrix

	A	B	C	D	E	F
A		1	2	4	5	5
B	1		2	4	5	5
C	2	2		3	4	4
D	4	4	3		2	2
E	5	5	4	2		1
F	5	5	4	2	1	

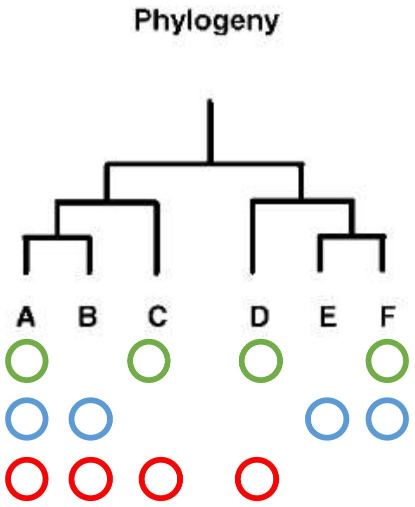
NTI: Nearest Taxon index

Measure the mean nearest phylogenetic neighbor

- Example with 4 species
- First: define the community with the greatest possible mean nearest nodal distances: ACDF

(A->C, C->A, D->F, F->D)

- Somme of distances/nb of nodes : 8/4 = 2
- Compare 2 communities
  - ABCD :
    - Mean nearest nodal distance : (1+1+2+3)/4 nodes = 1,75
    - NTI (Net index) : 1- 1,75/2 = 0,125
  - ABEF
    - -
    - -
    - -



Greatest possible mean *pairwise* nodal distance for a community of 4 taxa (given this phylogeny) = 3.66 nodes (for A, B, E, F)

Greatest possible mean *nearest* nodal distance for a community of 4 taxa (given this phylogeny) = 2.00 nodes (for A, C, D, F)

Distance matrix

	A	B	C	D	E	F
A		1	2	4	5	5
B	1		2	4	5	5
C	2	2		3	4	4
D	4	4	3		2	2
E	5	5	4	2		1
F	5	5	4	2	1	

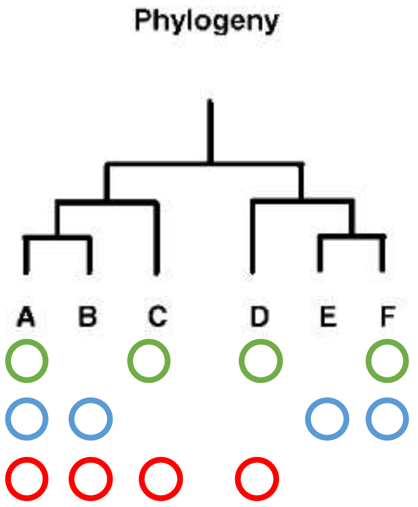
NTI: Nearest Taxon index

Measure the mean nearest phylogenetic neighbor

- Example with 4 species
- First: define the community with the greatest possible mean nearest nodal distances: ACDF

(A->C, C->A, D->F, F->D)

- Somme of distances/nb of nodes : 8/4 = 2
- Compare 2 communities
  - ABCD :
    - Mean nearest nodal distance : (1+1+2+3)/4 nodes = 1,75
    - NTI (Net index) : 1- 1,75/2 = 0,125
  - ABEF
    - Mean nearest nodal distance : (1+1+1+1)/4 nodes = 1
    - NTI (Net index) : 1- 1/2 = 0,5



Greatest possible mean *pairwise* nodal distance for a community of 4 taxa (given this phylogeny) = 3.66 nodes (for A, B, E, F)

Greatest possible mean *nearest* nodal distance for a community of 4 taxa (given this phylogeny) = 2.00 nodes (for A, C, D, F)

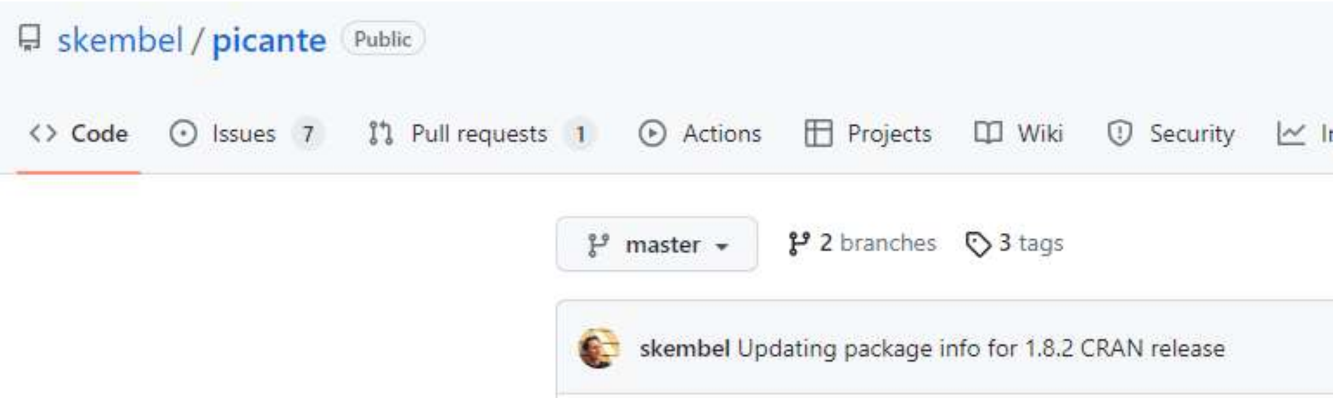
Distance matrix

	A	B	C	D	E	F
A		1	2	4	5	5
B	1		2	4	5	5
C	2	2		3	4	4
D	4	4	3		2	2
E	5	5	4	2		1
F	5	5	4	2	1	

> ABCD is more dispersed that ABEF in terms of average nearest neighbor

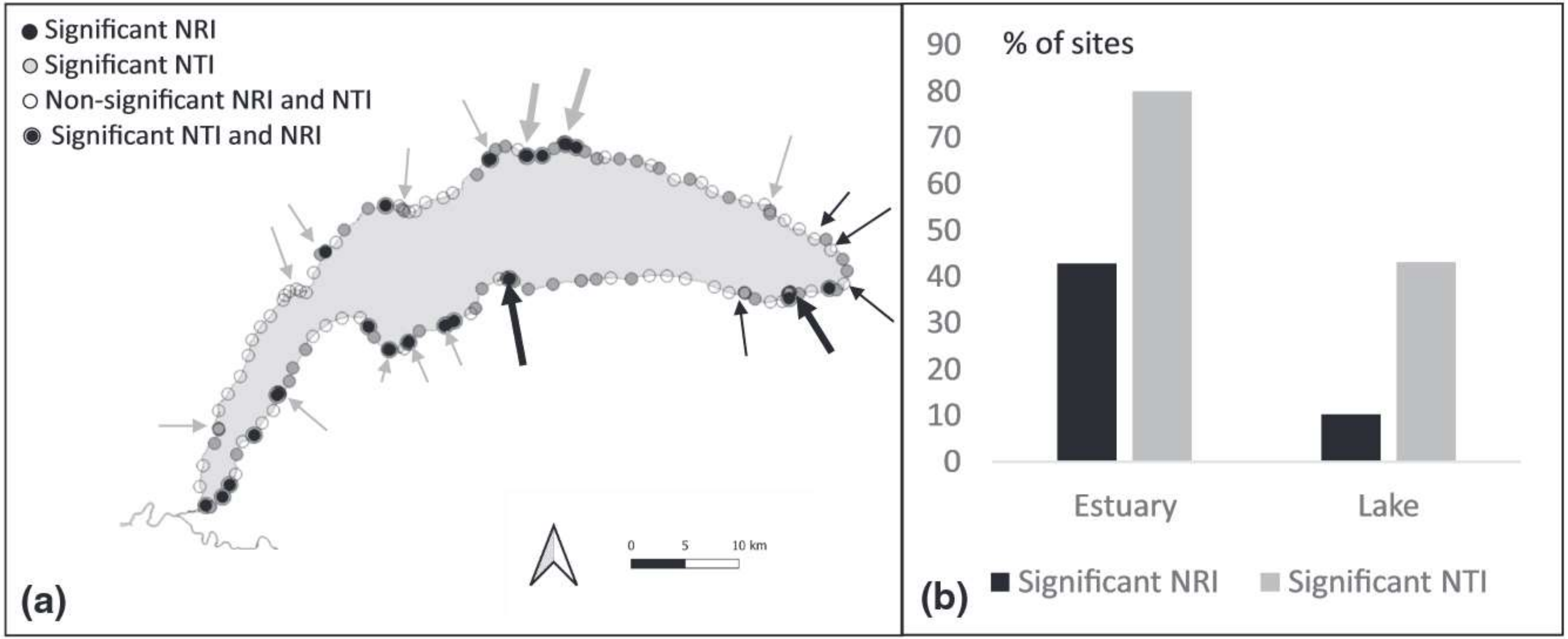
NRI and NTI calculation

- Calculations are carried out using the picante package
- The NRI and NTI values of each sample are compared to a null model (randomisation process) and a p-value is associated to NRI and NTI values



skembel / picante Public

skembel Updating package info for 1.8.2 CRAN release



- Assessment of environmental filtering vs competition in diatom communities of lake Geneva: only environmental filtering (over-clustering)

- Rimet, F., Canino, A., Chonova, T., Guéguen, J., Bouchez, A., 2023. Environmental filtering and mass effect are two important processes driving lake benthic diatoms: Results of a DNA metabarcoding study in a large lake. Molecular Ecology 32, 124–137. https://doi.org/10.1111/mec.16737

Schedule

BIOLAWEB

0. Including phylogenies into ecological studies

0.1 Phylogenetic diversity

0.2 Measuring phylogenetic signal

0.3 Measuring and testing community phylogenetic structure

1. Theory of sequence alignment

1.1 What is an alignment?

1.2 Objective of aligning sequences

1.3 Edit distance

1.4 Local alignments

1.5 Global alignments

2. Phylogenies

2.1 Choose appropriate model of sequence evolution

2.2 Phylogeny inference

2.3 Graphical representation of phylogenies

3. Phylogenetic placement

3.1 Definition

3.2 Why using this

3.3 Example with RaxML in R

4. Ecophylogenetic training in R (R studio)

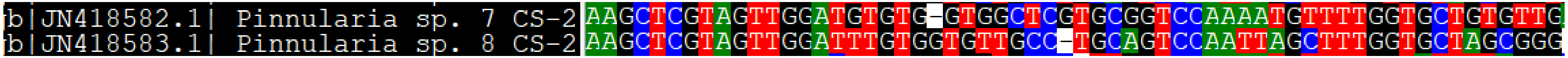
PD

NRI NTI

1.1 What is an alignment?

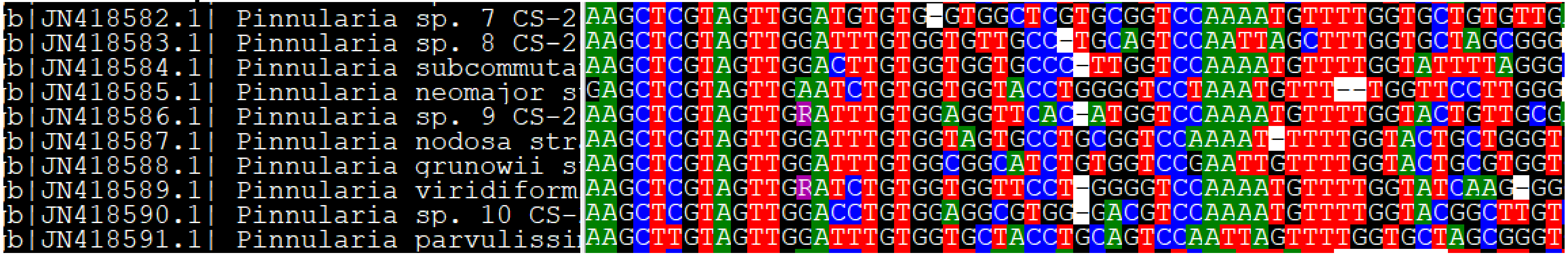
- Sequence alignment is the procedure of comparing 2 (pairwise alignment) or several sequences by searching series of individual characters or patterns that are in the same order in the sequences.

- Pairwise



```
b|JN418582.1| Pinnularia sp. 7 CS-2 AAGCTCGTAGTTGGATGTGTG-GTGGCTCGTGCGGTCCAAAATGTTTTGGTGCTGTGTTG
b|JN418583.1| Pinnularia sp. 8 CS-2 AAGCTCGTAGTTGGATTTGTGGTGTTGCC-TGCAGTCCAATTAGCTTTGGTGCTAGCGGG
```

- Multiple



```
b|JN418582.1| Pinnularia sp. 7 CS-2 AAGCTCGTAGTTGGATGTGTG-GTGGCTCGTGCGGTCCAAAATGTTTTGGTGCTGTGTTG
b|JN418583.1| Pinnularia sp. 8 CS-2 AAGCTCGTAGTTGGATTTGTGGTGTTGCC-TGCAGTCCAATTAGCTTTGGTGCTAGCGGG
b|JN418584.1| Pinnularia subcommuta AAGCTCGTAGTTGGACTTGTGGTGGTGCCC-TTGGTCCAAAATGTTTTGGTATTTTAGGG
b|JN418585.1| Pinnularia neomajor s GAGCTCGTAGTTGAATCTGTGGTGGTACCTGGGGTCCTAAATGTTT--TGGTTCCTTGGG
b|JN418586.1| Pinnularia sp. 9 CS-2 AAGCTCGTAGTTGRATTTGTGGAGGTTCAC-ATGGTCCAAAATGTTTTGGTACTGTTGCG
b|JN418587.1| Pinnularia nodosa str AAGCTCGTAGTTGGATTTGTGGTAGTGCCTGCGGTCCAAAAT-TTTTGGTACTGCTGGGT
b|JN418588.1| Pinnularia grunowii s AAGCTCGTAGTTGGATTTGTGGCGGCATCTGTGGTCCGAATTGTTTTGGTACTGCGTGGT
b|JN418589.1| Pinnularia viridiform AAGCTCGTAGTTGRATCTGTGGTGGTTCCT-GGGGTCCAAAATGTTTTGGTATCAAG-GG
b|JN418590.1| Pinnularia sp. 10 CS- AAGCTCGTAGTTGGACCTGTGGAGGCGTGG-GACGTCCAAAATGTTTTGGTACGGCTTGT
b|JN418591.1| Pinnularia parvulissi AAGCTTGTAGTTGGATTTGTGGTGCTACCTGCAGTCCAATTAGTTTTGGTGCTAGCGGGT
```

1.2 Objective of aligning sequences

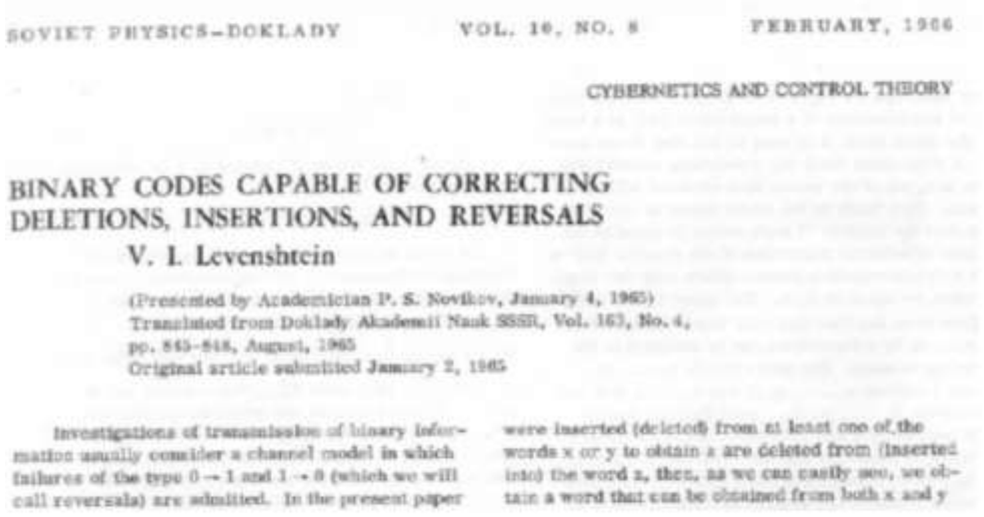
- Major objective: compare sequences between each other

- Applied objectives:
  - Find evolutionary relationships
  - To search databases (eg reference barcoding libraries) -> pairwise alignment
  - Prediction of protein structure and function (if same sequences -> same 3D structure -> same function)

1.3 Edit distance (Levenshtein distance)

- How do we measure distance between strings?
- The edit distance between 2 strings is defined as the minimum number of edits needed to transform one string into the other, with the following edit operations:
    - Insertion
    - Deletion
    - Substitution

of a single character



SOVIET PHYSICS–DOKLADY VOL. 10, NO. 8 FEBRUARY, 1966

CYBERNETICS AND CONTROL THEORY

BINARY CODES CAPABLE OF CORRECTING DELETIONS, INSERTIONS, AND REVERSALS

V. I. Levenshtein

(Presented by Academician P. S. Novikov, January 4, 1965)
Translated from Doklady Akademii Nauk SSSR, Vol. 163, No. 4,
pp. 845-848, August, 1965
Original article submitted January 2, 1965

Investigations of transmission of binary information usually consider a channel model in which failures of the type 0 → 1 and 1 → 0 (which we will call reversals) are admitted. In the present paper

were inserted (deleted) from at least one of the words x or y to obtain z are deleted from (inserted into) the word z, then, as we can easily see, we obtain a word that can be obtained from both x and y

1.3 Edit distance (Levenshtein distance)

- How do we measure distance between strings?
- The edit distance between 2 strings is defined as the minimum number of edits needed to transform one string into the other, with the following edit operations :
    - Insertion: helo -> hello
    - Deletion: helo -> he-o
    - Substitution: helo -> help

of a single character

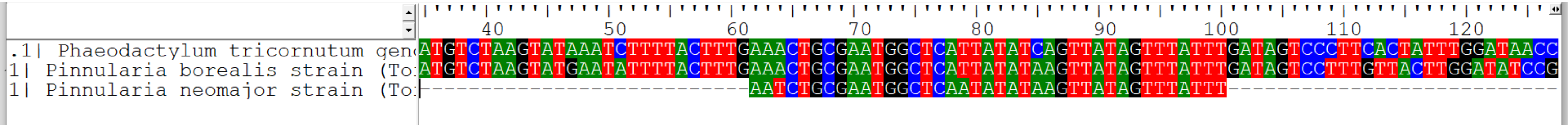
- Try with:

kitten > sitting

kitten > sitten > sittin > sitting → distance = 3

1.4 Local alignment

- Local alignment is to try to find the regions with highest density of matches.



```
                                         40        50        60        70        80        90        100       110       120
.1| Phaeodactylum tricornutum gen ATGTCTAAGTATAAATCTTTTACTTTGAAACTGCGAATGGCTCATTATATCAGTTATAGTTTATTTGATAGTCCCTTCACTATTTGGATAACC
1| Pinnularia borealis strain (To ATGTCTAAGTATGAATATTTTACTTTGAAACTGCGAATGGCTCATTATATAAGTTATAGTTTATTTGATAGTCCTTTGTTACTTGGATATCCG
1| Pinnularia neomajor strain (To ---------------------------AATCTGCGAATGGCTCAATATATAAGTTATAGTTTATTT---------------------------
```

- Local alignment is based on Smith-Waterman: Focuses on the region of greatest similarity between two sequences
- Suitable for aligning more divergent sequences. Used for performing searches on large databases

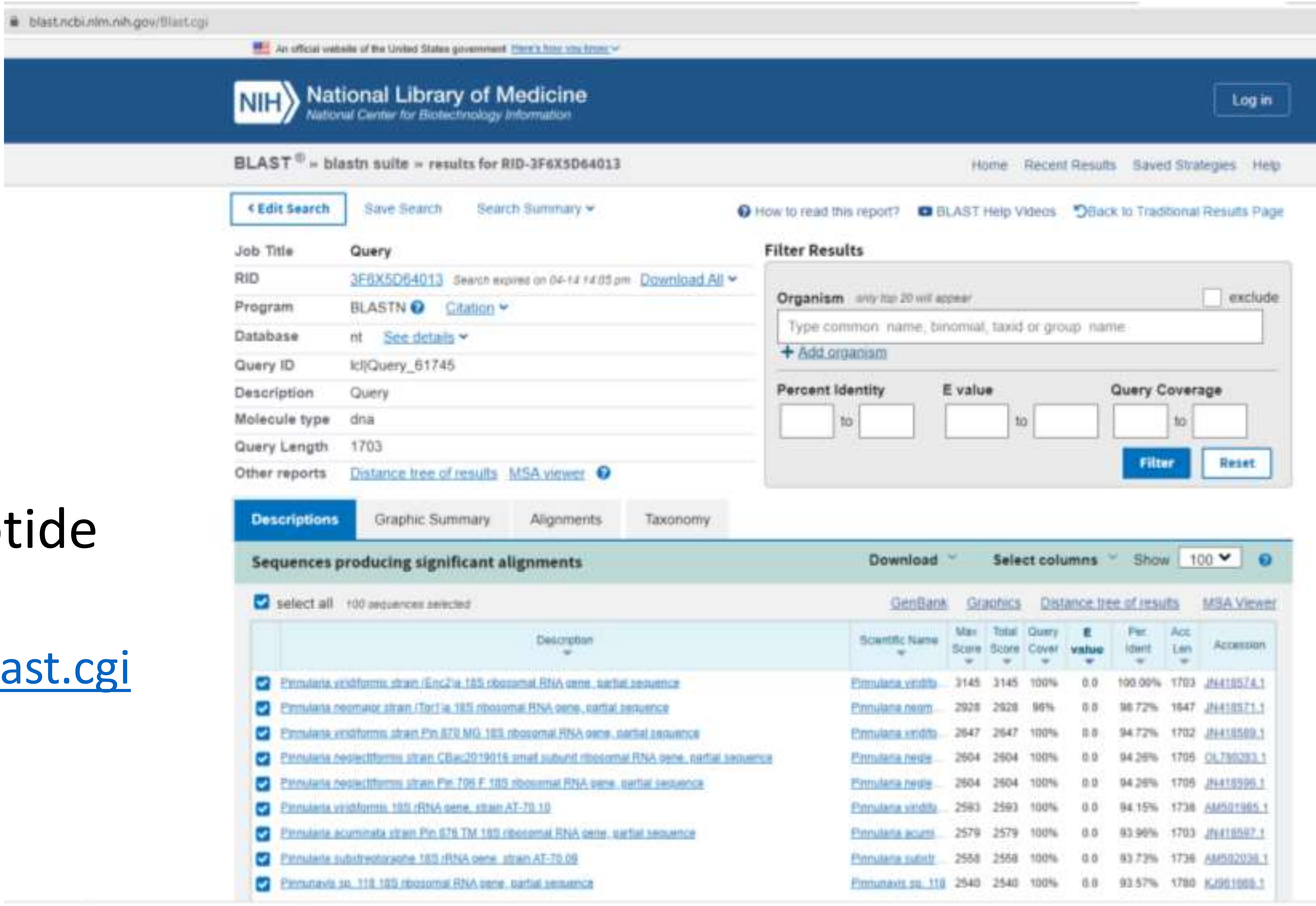
1.4 Local alignment

Use the following file:

Query.fasta

and blast it on NCBI nucleotide

https://blast.ncbi.nlm.nih.gov/Blast.cgi


blast.ncbi.nlm.nih.gov/Blast.cgi
An official website of the United States government Here's how you know
NIH National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST ® » blastn suite » results for RID-3F6X5D64013
Home
Recent Results
Saved Strategies
Help
< Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page

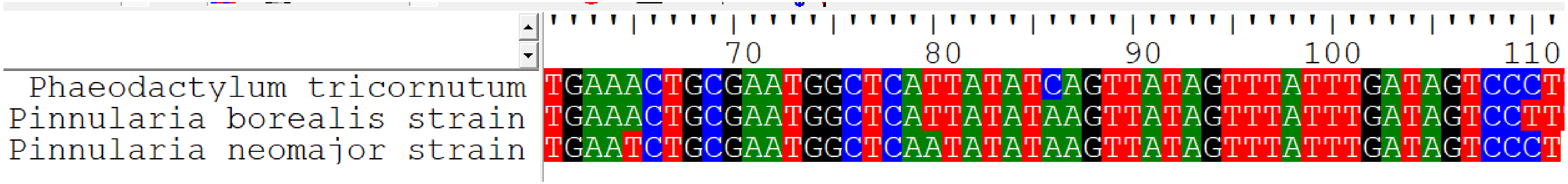
Job Title	Query
RID	3F6X5D64013 Search expires on 04-14 14:05 pm Download All
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_61745
Description	Query
Molecule type	dna
Query Length	1703
Other reports	Distance tree of results MSA viewer

Filter Results
Organism only top 20 will appear
□ exclude
Type common name, binomial, taxid or group name
+ Add organism
Percent Identity
E value
Query Coverage
to
Filter
Reset
Descriptions
Graphic Summary
Alignments
Taxonomy
Sequences producing significant alignments
Download
Select columns
Show 100
☑ select all
100 sequences selected
GenBank
Graphics
Distance tree of results
MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Pinnularia viridiformis strain (Enc2)a 18S ribosomal RNA gene, partial sequence	Pinnularia viridifo...	3145	3145	100%	0.0	100.00%	1703	JN418574.1
☑	Pinnularia neomajor strain (Tor1)a 18S ribosomal RNA gene, partial sequence	Pinnularia neom...	2928	2928	96%	0.0	98.72%	1647	JN418571.1
☑	Pinnularia viridiformis strain Pin 870 MG 18S ribosomal RNA gene, partial sequence	Pinnularia viridifo...	2647	2647	100%	0.0	94.72%	1702	JN418589.1
☑	Pinnularia neglectiformis strain CBac2019016 small subunit ribosomal RNA gene, partial sequence	Pinnularia negle...	2604	2604	100%	0.0	94.26%	1705	OL780283.1
☑	Pinnularia neglectiformis strain Pin 706 F 18S ribosomal RNA gene, partial sequence	Pinnularia negle...	2604	2604	100%	0.0	94.26%	1705	JN418596.1
☑	Pinnularia viridiformis 18S rRNA gene, strain AT-70.10	Pinnularia viridifo...	2593	2593	100%	0.0	94.15%	1738	AM501985.1
☑	Pinnularia acuminata strain Pin 876 TM 18S ribosomal RNA gene, partial sequence	Pinnularia acumi...	2579	2579	100%	0.0	93.96%	1703	JN418597.1
☑	Pinnularia substreptoraphe 18S rRNA gene, strain AT-70.09	Pinnularia substr...	2558	2558	100%	0.0	93.73%	1736	AM502036.1
☑	Pinnunavis sp. 118 18S ribosomal RNA gene, partial sequence	Pinnunavis sp. 118	2540	2540	100%	0.0	93.57%	1780	KJ961669.1

1.5 Global alignment

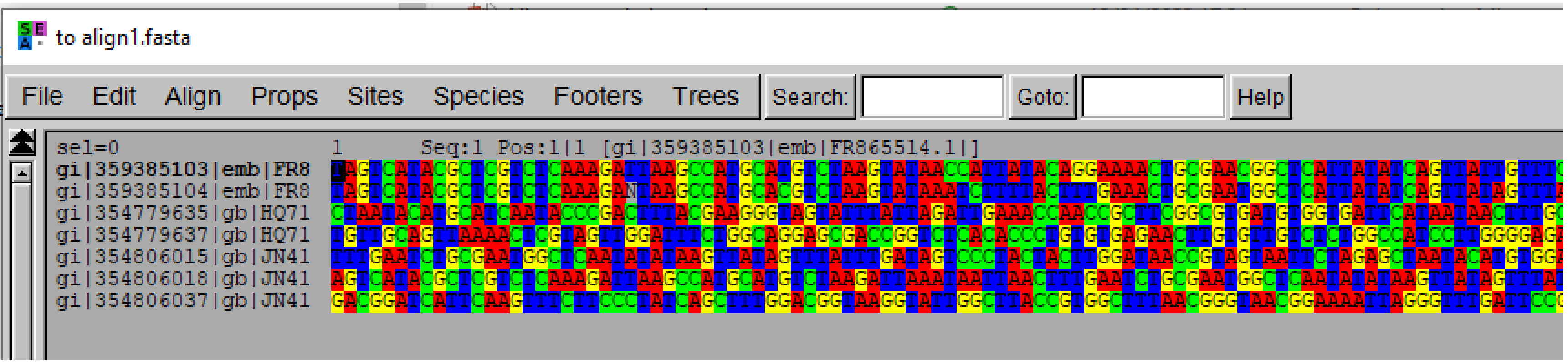
- A global alignment is attempting to match as much of the sequence as possible.
- Global alignment is based on Needleman-Wunsch algorithm.
- Suitable for aligning two closely related sequences, homologous genes (=gene inherited in two species from a common ancestor)



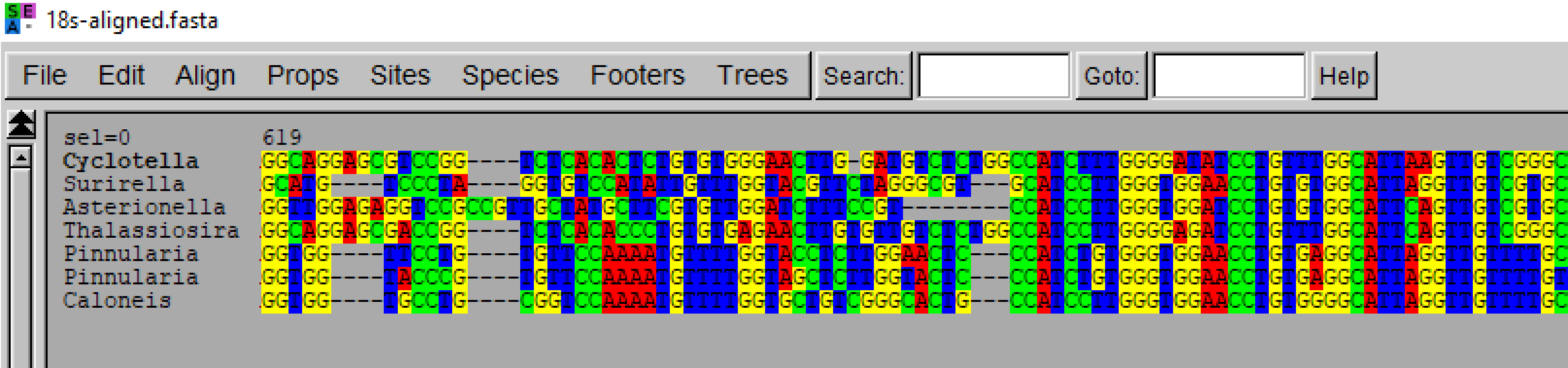
```
                                  70        80        90       100       110
  Phaeodactylum tricornutum TGAAACTGCGAATGGCTCATTATATCAGTTATAGTTTATTTGATAGTCCCT
Pinnularia borealis strain TGAAACTGCGAATGGCTCATTATATAAGTTATAGTTTATTTGATAGTCCTT
Pinnularia neomajor strain TGAATCTGCGAATGGCTCAATATATAAGTTATAGTTTATTTGATAGTCCCT
```

1.5 Global alignment

- 1st example, use file « 18s-to align.fasta »
- Use SeaView : https://doua.prabi.fr/software/seaview



```
sel=0              1        Seq:1 Pos:1|1 [gi|359385103|emb|FR865514.1|]
gi|359385103|emb|FR8  TAGTCATACGCTCGTCTCAAAGATTAAGCCATGCATGTCTAAGTATAACCATTATACAGGAAAACTGCGAACGGCTCATTATATCAGTTATTGTTTG
gi|359385104|emb|FR8  TAGTCATACGCTCGTCTCAAAGANTAAGCCATGCACGTCTAAGTATAAATCTTTTACTTTGAAACTGCGAATGGCTCATTATATCAGTTATAGTTTA
gi|354779635|gb|HQ71  CTAATACATGCATCAATACCCGACTTTACGAAGGGTAGTATTTATTAGATTGAAACCAACCGCTTCGGCGTGATGTGGTGATTCATAATAACTTTGC
gi|354779637|gb|HQ71  TGTTGCAGTTAAAACTCGTAGTTGGATTTCTGGCAGGAGCGACCGGTCTCACACCCTGTGTGAGAACTTGTGTTGTCTCTGGCCATCCTTGGGGAGA
gi|354806015|gb|JN41  TTTGAATCTGCGAATGGCTCAATATATAAGTTATAGTTTATTTGATAGTCCCTACTACTTGGATAACCGTAGTAATTCTAGAGCTAATACATGTGGA
gi|354806018|gb|JN41  AGTCATACGCTCGTCTCAAAGATTAAGCCATGCATGTCTAAGATTAAATAATTAACTTTGAATCTGCGAATGGCTCAATATATAAGTTATAGTTTA
gi|354806037|gb|JN41  GACGGATCATTCAAGTTTCTTCCCTATCAGCTTTGGACGGTAAGGTATTGGCTTACCGTGGCTTTAACGGGTAACGGAAAATTAGGGTTTGATTCCG
```

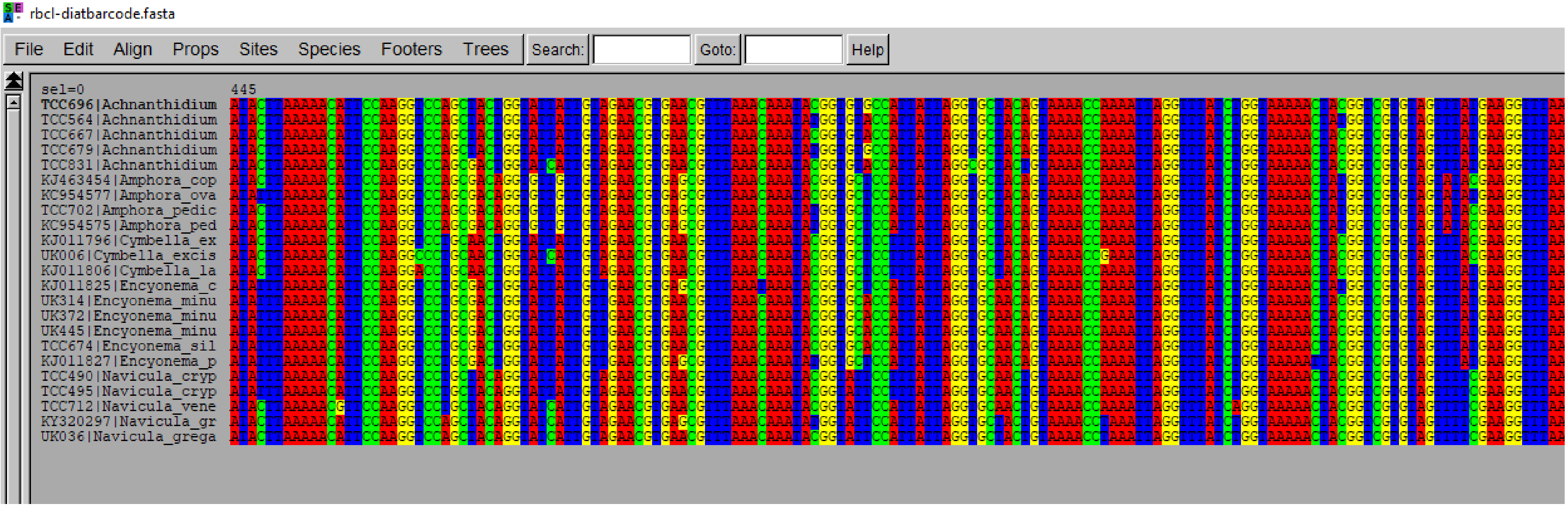


```
sel=0          619
Cyclotella     GGCAGGAGCGTCCGG----TCTCACACTCTGTGGGAACTTG-GATGTCTCTGGCCATCTTTGGGGATATCCTGTTTGGCATTAAGTTGTCGGGC
Surirella      GCATG----TCCCTA----GGTGTCCATATTGTTTGGTACGTTCTAGGGCGT---GCATCCTTGGGTGGAACCTGTGTGGCATTAGGTTGTCGTGC
Asterionella   GGTTGGAGAGGTCCGCCGTTGCTATGCTTCGTGTTGGATCTTTCCGT--------CCATCCTTGGGTGGATCCTGTGTGGCATTCAGTTGTCGTGC
Thalassiosira  GGCAGGAGCGACCGG----TCTCACACCCTGTGTGAGAACTTGTGTTGTCTCTGGCCATCCTTGGGGAGATCCTGTTTGGCATTCAGTTGTCGGGC
Pinnularia     GGTGG----TTCCTG----TGTTCCAAAATGTTTTGGTACCTCTTGGAACTC---CCATCTGTGGGTGGAACCTGTGAGGCATTAGGTTGTTTTGC
Pinnularia     GGTGG----TACCCG----TGTTCCAAAATGTTTTGGTAGCTCTTGGTACTC---CCATCTGTGGGTGGAACCTGTGAGGCATTAGGTTGTTTTGT
Caloneis       GGTGG----TGCCTG----CGGTCCAAAATGTTTTGGTGCTGTCGGGCACTG---CCATCCTTGGGTGGAACCTGTGGGGCATTAGGTTGTTTTGC
```

Presence of insertion/deletions/substitution

1.5 Global alignment

- 2nd example, use file: « rbcl-diatbarcode.fasta »
- Use SeaView to open « rbcl-diatbarcode.fasta » (don’t align it, it’s already done)



```
sel=0                    445
TCC696|Achnanthidium
TCC564|Achnanthidium
TCC667|Achnanthidium
TCC679|Achnanthidium
TCC831|Achnanthidium
KJ463454|Amphora_cop
KC954577|Amphora_ova
TCC702|Amphora_pedic
KC954575|Amphora_ped
KJ011796|Cymbella_ex
UK006|Cymbella_excis
KJ011806|Cymbella_la
KJ011825|Encyonema_c
UK314|Encyonema_minu
UK372|Encyonema_minu
UK445|Encyonema_minu
TCC674|Encyonema_sil
KJ011827|Encyonema_p
TCC490|Navicula_cryp
TCC495|Navicula_cryp
TCC712|Navicula_vene
KY320297|Navicula_gr
UK036|Navicula_grega
```

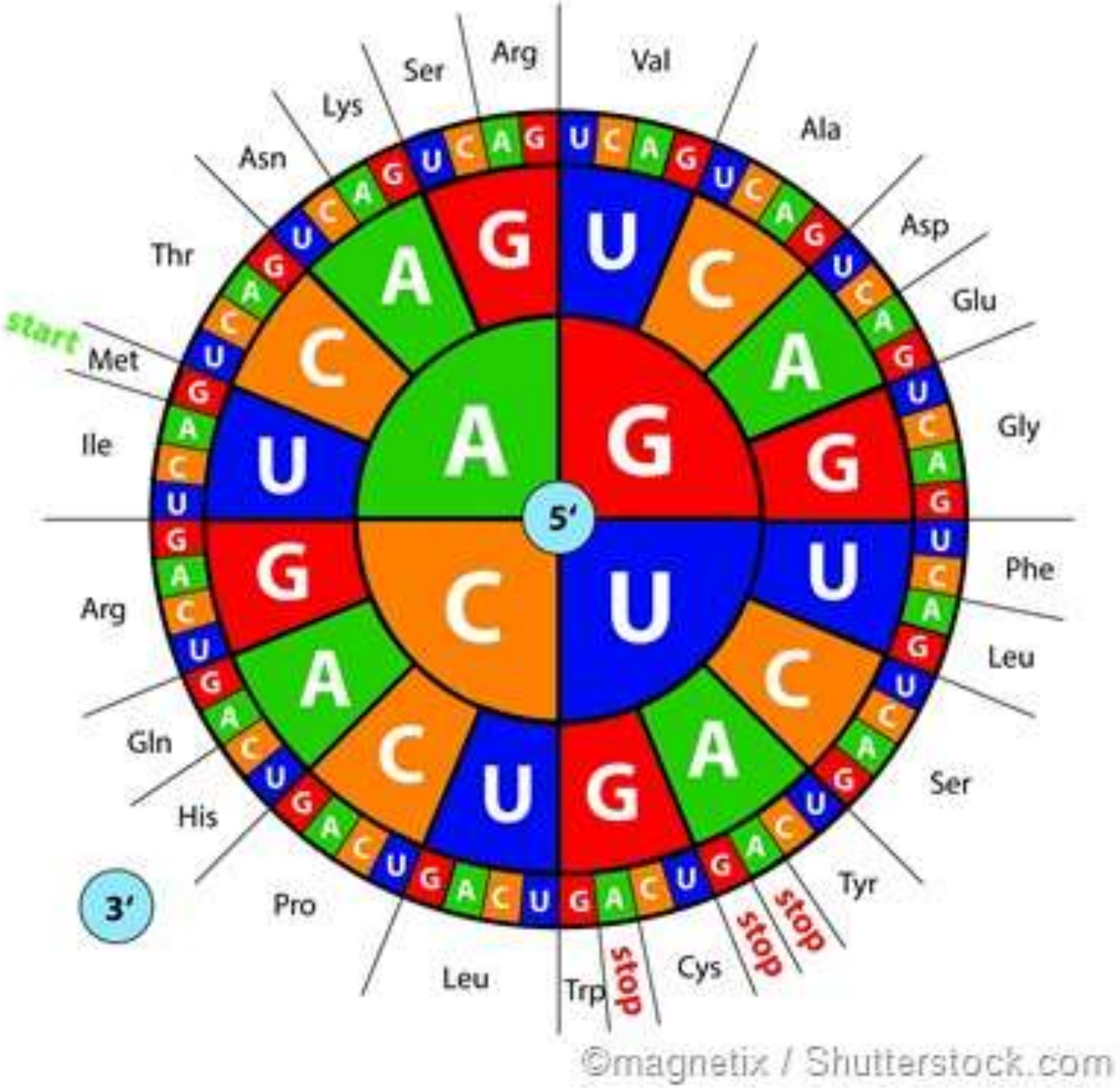
Absence of insertion/deletions: it is a coding marker

- 3 nucleotides

= 1 codon

= 1 amino acid

or a stop codon



Schedule

0. Including phylogenies into ecological studies

0.1 Phylogenetic diversity

0.2 Measuring phylogenetic signal

0.3 Measuring and testing community phylogenetic structure

1. Theory of sequence alignment

1.1 What is an alignment?

1.2 Objective of aligning sequences

1.3 Edit distance

1.4 Local alignments

1.5 Global alignments

2. Phylogenies

2.1 Choose appropriate model of sequence evolution

2.2 Phylogeny inference

2.3 Graphical representation of phylogenies

3. Phylogenetic placement

3.1 Definition

3.2 Why using this

3.3 Example with RaxML in R

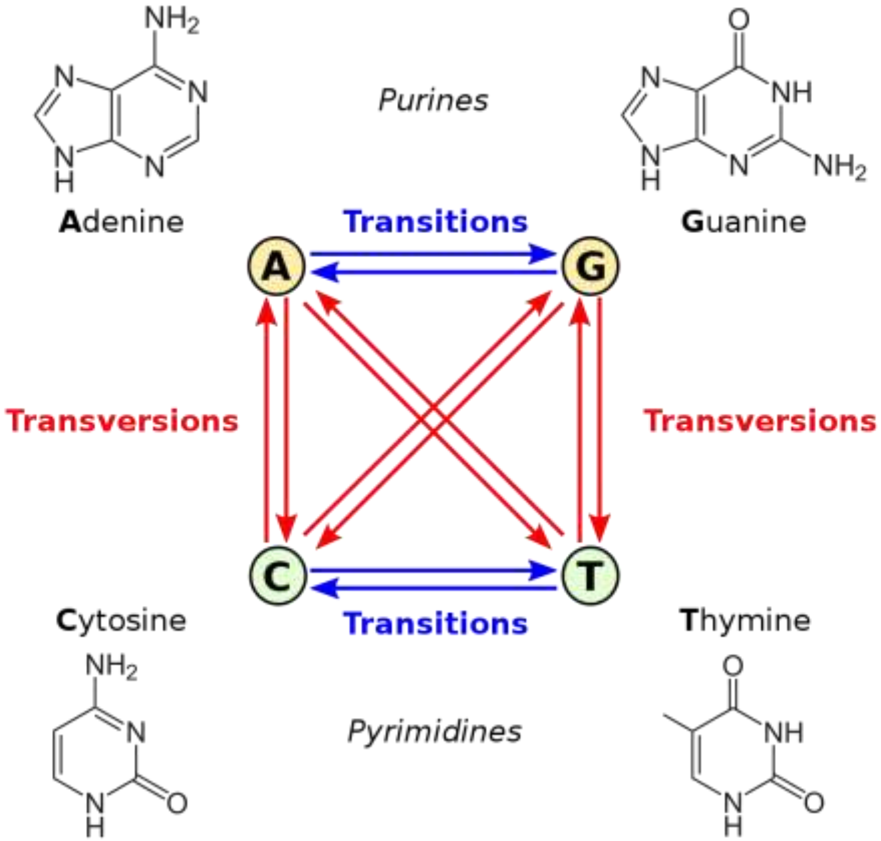
4. Ecophylogenetic training in R (R studio)

PD

NRI NTI

2.1 Choose appropriate models of sequence evolution

- 4 nucleotides
  - A=T -> 2 hydrogen bonds
  - CΞG -> 3 hydrogen bonds
  - A, G: double ring structure
  - C, T: single ring structure
- 2 kinds of mutations :
  - Transitions
  - Transversions
- Transitions are more frequent than Transversions (easier to change from a single ring structure to a single ring structure, than from a single to double ring structure)

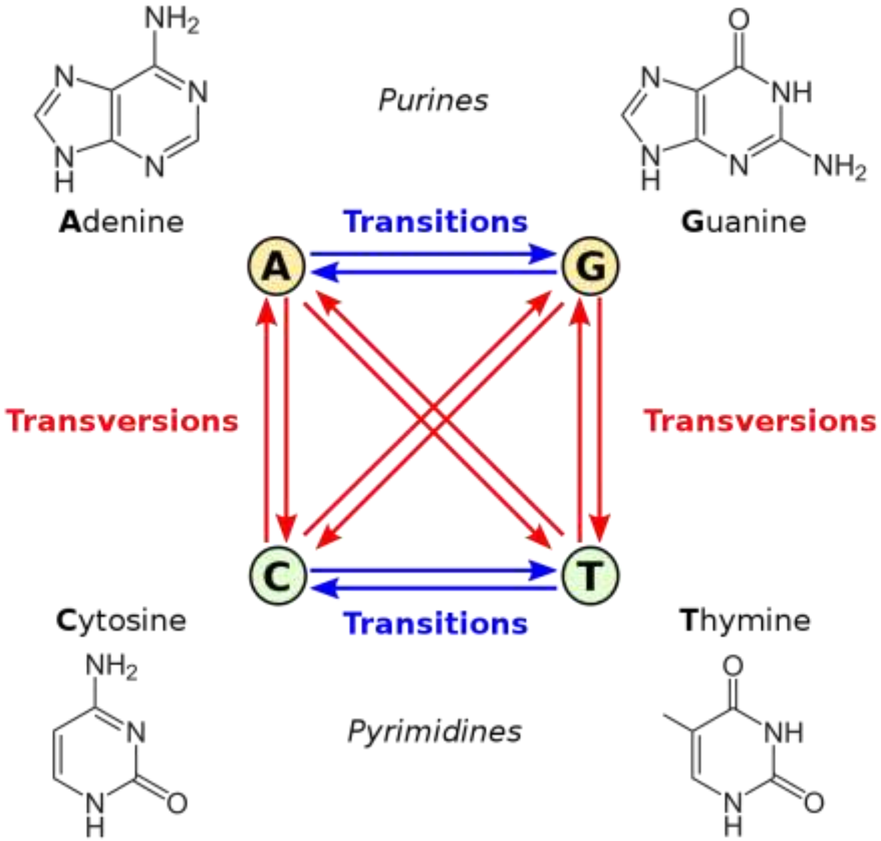


2.1 Choose appropriate models of sequence evolution

- to model the substitution process in DNA sequences, the corresponding transition matrices will look like the following matrix

$$P(t) = \begin{pmatrix} p_{\mathrm{AA}}(t) & p_{\mathrm{AG}}(t) & p_{\mathrm{AC}}(t) & p_{\mathrm{AT}}(t) \\ p_{\mathrm{GA}}(t) & p_{\mathrm{GG}}(t) & p_{\mathrm{GC}}(t) & p_{\mathrm{GT}}(t) \\ p_{\mathrm{CA}}(t) & p_{\mathrm{CG}}(t) & p_{\mathrm{CC}}(t) & p_{\mathrm{CT}}(t) \\ p_{\mathrm{TA}}(t) & p_{\mathrm{TG}}(t) & p_{\mathrm{TC}}(t) & p_{\mathrm{TT}}(t) \end{pmatrix}$$

Where p(t) probability to change from a nucleotide to another in a time t

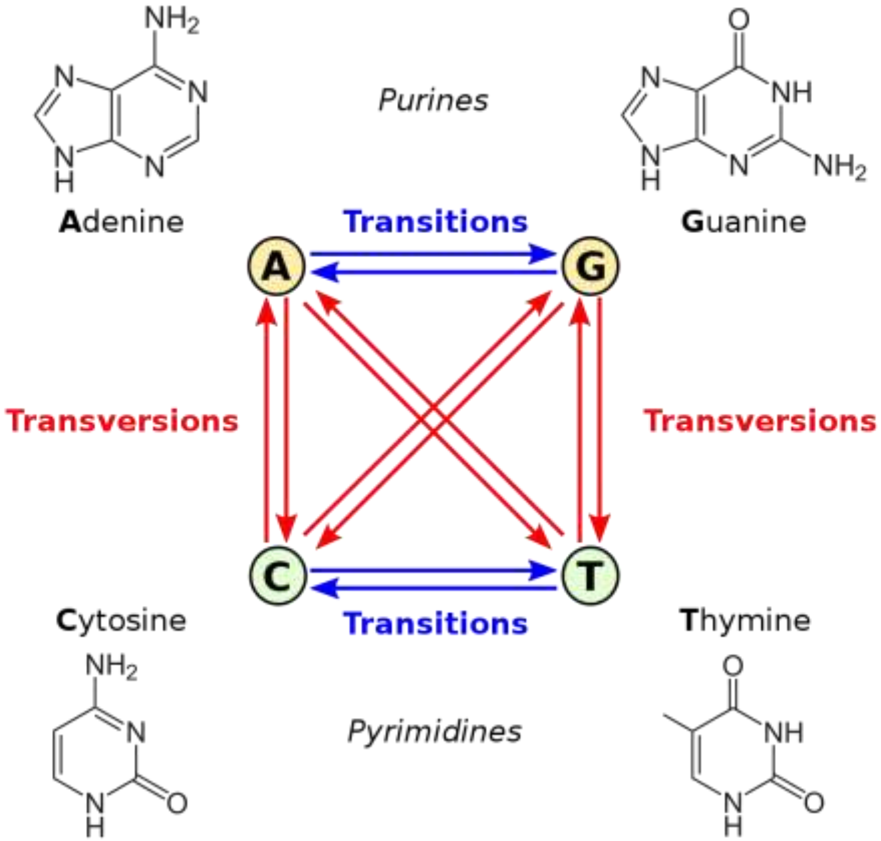


2.1 Choose appropriate models of sequence evolution

- Also given in the rate matrix (Q matrix),

μ = mutation rate

$$Q = \begin{pmatrix} -\mu_A & \mu_{AG} & \mu_{AC} & \mu_{AT} \\ \mu_{GA} & -\mu_G & \mu_{GC} & \mu_{GT} \\ \mu_{CA} & \mu_{CG} & -\mu_C & \mu_{CT} \\ \mu_{TA} & \mu_{TG} & \mu_{TC} & -\mu_T \end{pmatrix}$$



$\mu_A = \mu_{AG} + \mu_{AC} + \mu_{AT}$

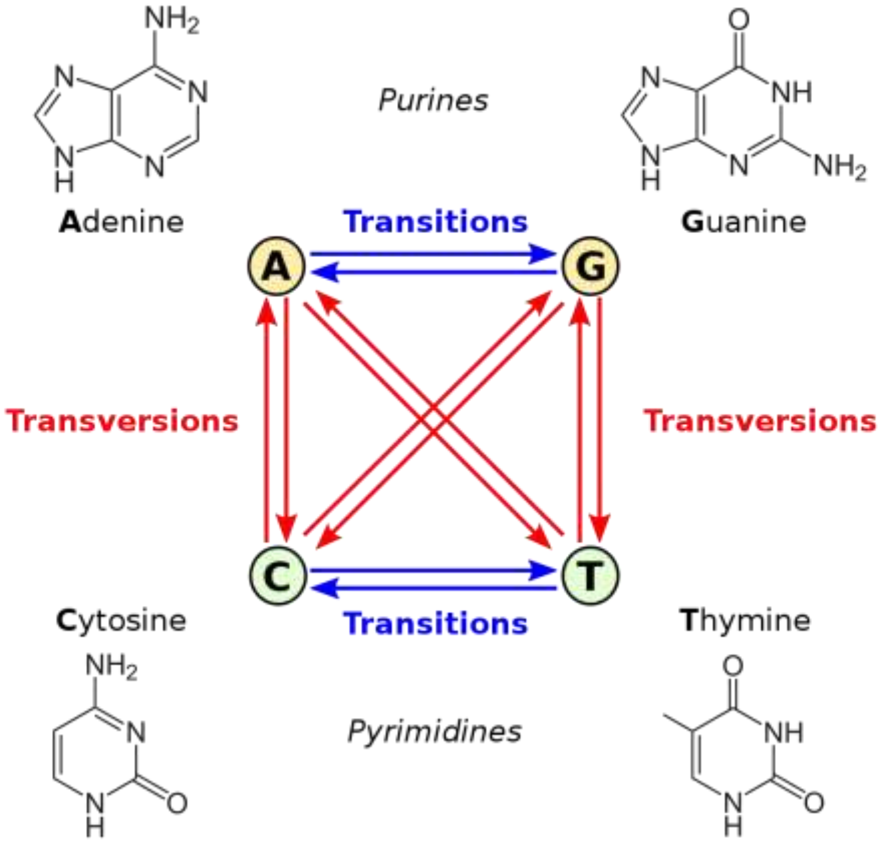
Sum of entries of Q equals 0

2.1 Choose appropriate models of sequence evolution

- The simplest model: Juke & Cantor 1969

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

- Strong hypothesis:
  - Equal base frequencies (1/4)
  - Equal mutation rates

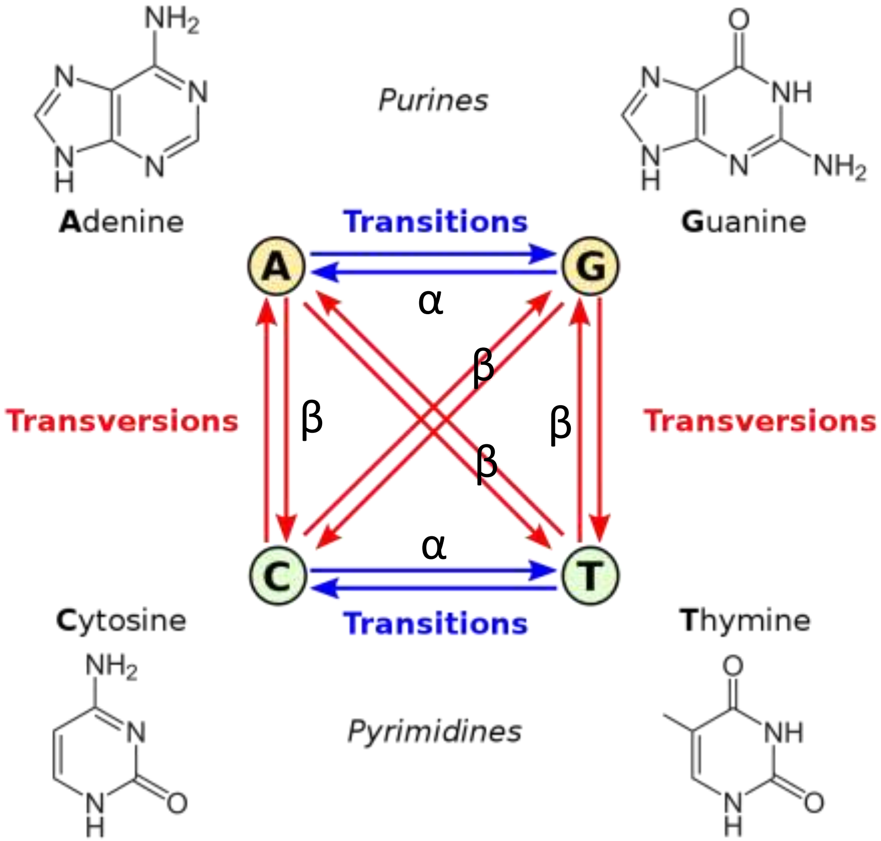


2.1 Choose appropriate models of sequence evolution

- 2 parameters: Kimura 1980

$$Q = \begin{pmatrix} * & \alpha & \beta & \beta \\ \alpha & * & \beta & \beta \\ \beta & \beta & * & \alpha \\ \beta & \beta & \alpha & * \end{pmatrix}$$

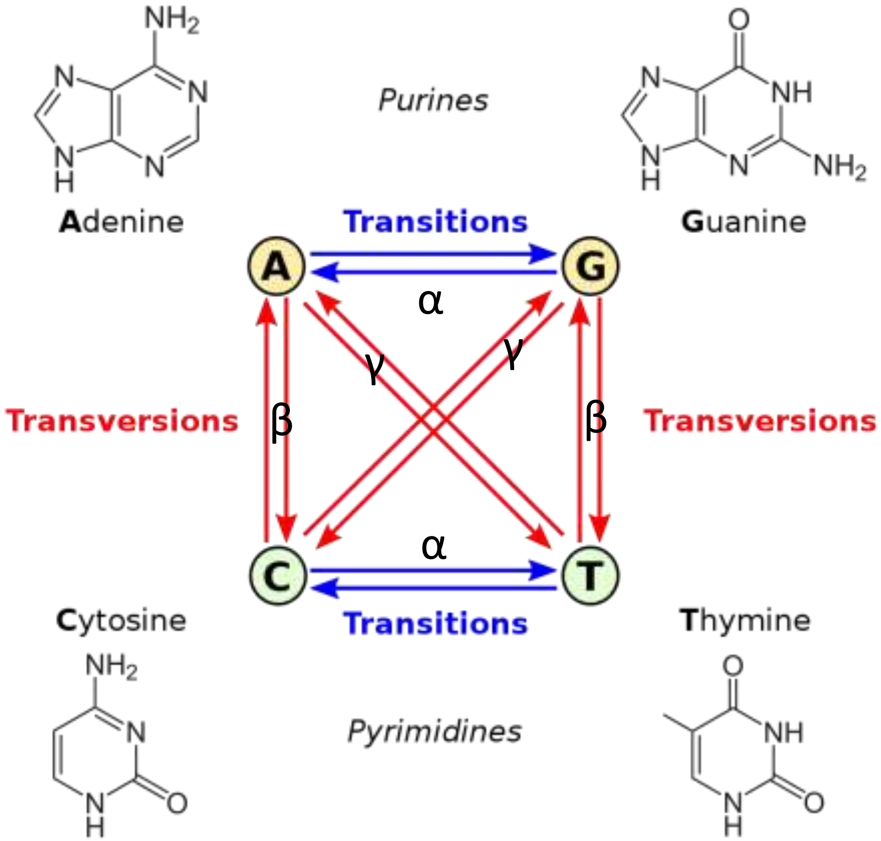
- Hypotheses:
  - Equal base frequencies (1/4)
  - has distinct rates for transitions (α) and transversions (β)



2.1 Choose appropriate models of sequence evolution

- 3 parameters: Kimura 1981

$$Q = \begin{pmatrix} * & \alpha & \beta & \gamma \\ \alpha & * & \gamma & \beta \\ \beta & \gamma & * & \alpha \\ \gamma & \beta & \alpha & * \end{pmatrix}$$

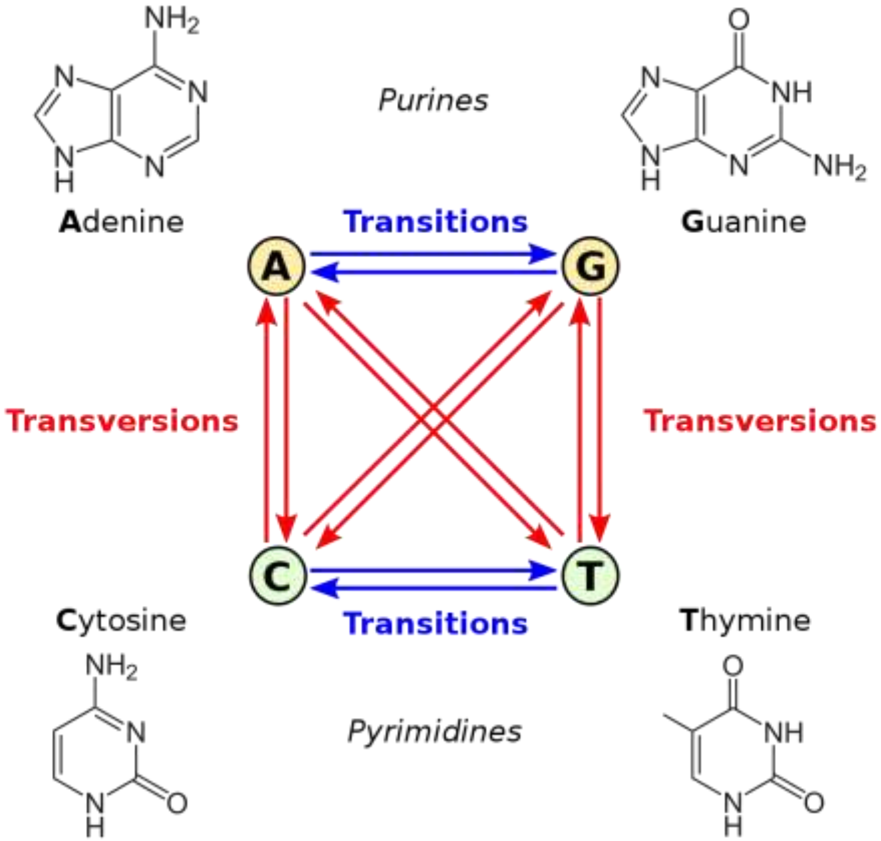


- Hypotheses:
  - Equal base frequencies (1/4)
  - has distinct rates for transitions (α) and two distinct types of transversions (β, γ)

2.1 Choose appropriate models of sequence evolution

- F81 model (Felsenstein 1981)

$$Q = \begin{pmatrix} * & \pi_G & \pi_C & \pi_T \\ \pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \pi_T \\ \pi_A & \pi_G & \pi_C & * \end{pmatrix}$$

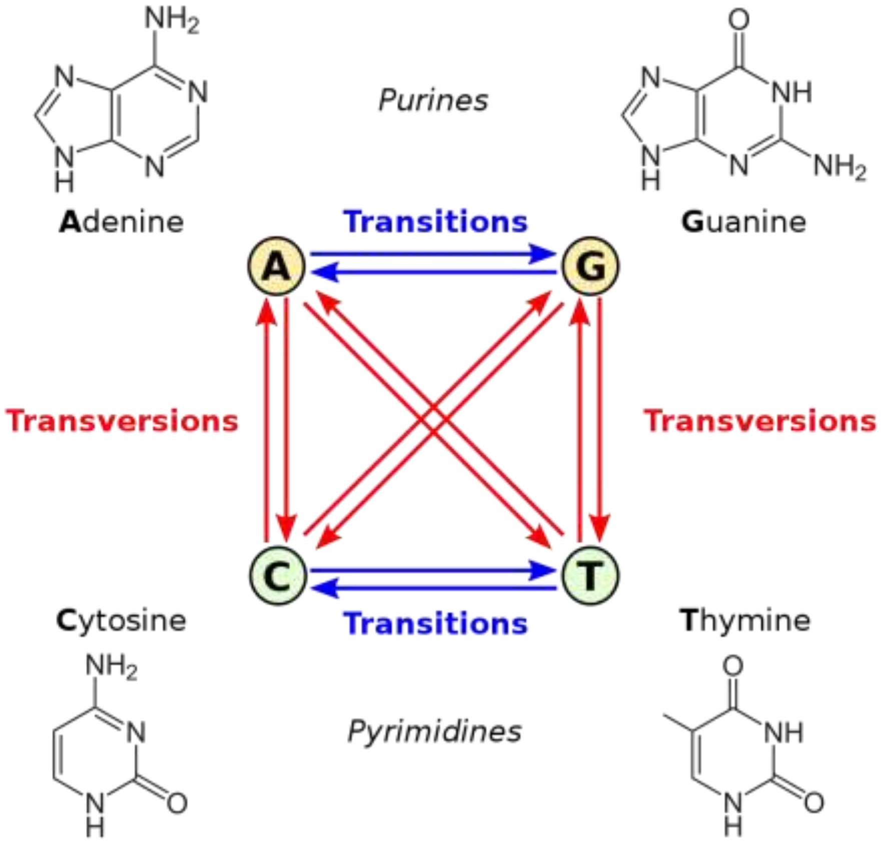


- Hypotheses:
  - Base frequencies are different (≠1/4, $\pi_A \neq \pi_T \neq \pi_C \neq \pi_G$)

2.1 Choose appropriate models of sequence evolution

- Generalised time-reversible model (Tavaré 1986) - GTR

$$Q = \begin{pmatrix} -(\alpha\pi_G+\beta\pi_C+\gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A+\delta\pi_C+\epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A+\delta\pi_G+\eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & -(\gamma\pi_A+\epsilon\pi_G+\eta\pi_C) \end{pmatrix}$$



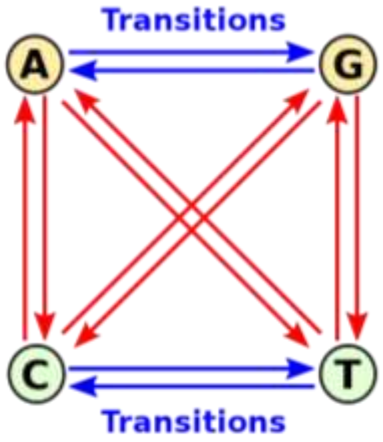
- Hypotheses:
    - Base frequencies are different (≠1/4, $\pi_A \neq \pi_T \neq \pi_C \neq \pi_G$)
    - All mutation are different (α, β, γ, δ, ε, η)

2.1 Choose appropriate models of sequence evolution

- We need to choose the correct model to weight the nucleotide differences between sequences.

- Juke&Cantor? Kimura81? F81? GTR?

?



2.1 Choose appropriate models of sequence evolution

- Test of the model: in MEGA-X
  - Open : “rbcl-diatbarcode.fasta”
  - Click: “Analyze”
  - Click: “nucleotide sequence”
  - Protein coding nucleotide sequence? “Yes”
  - Select a genetic code: “standard”
  - Analysis > Model > “Find best DNA model” > “Ok”

MEGA Molecular Evolutionary Genetics Analysis

Download:
https://www.megasoftware.net/index.html

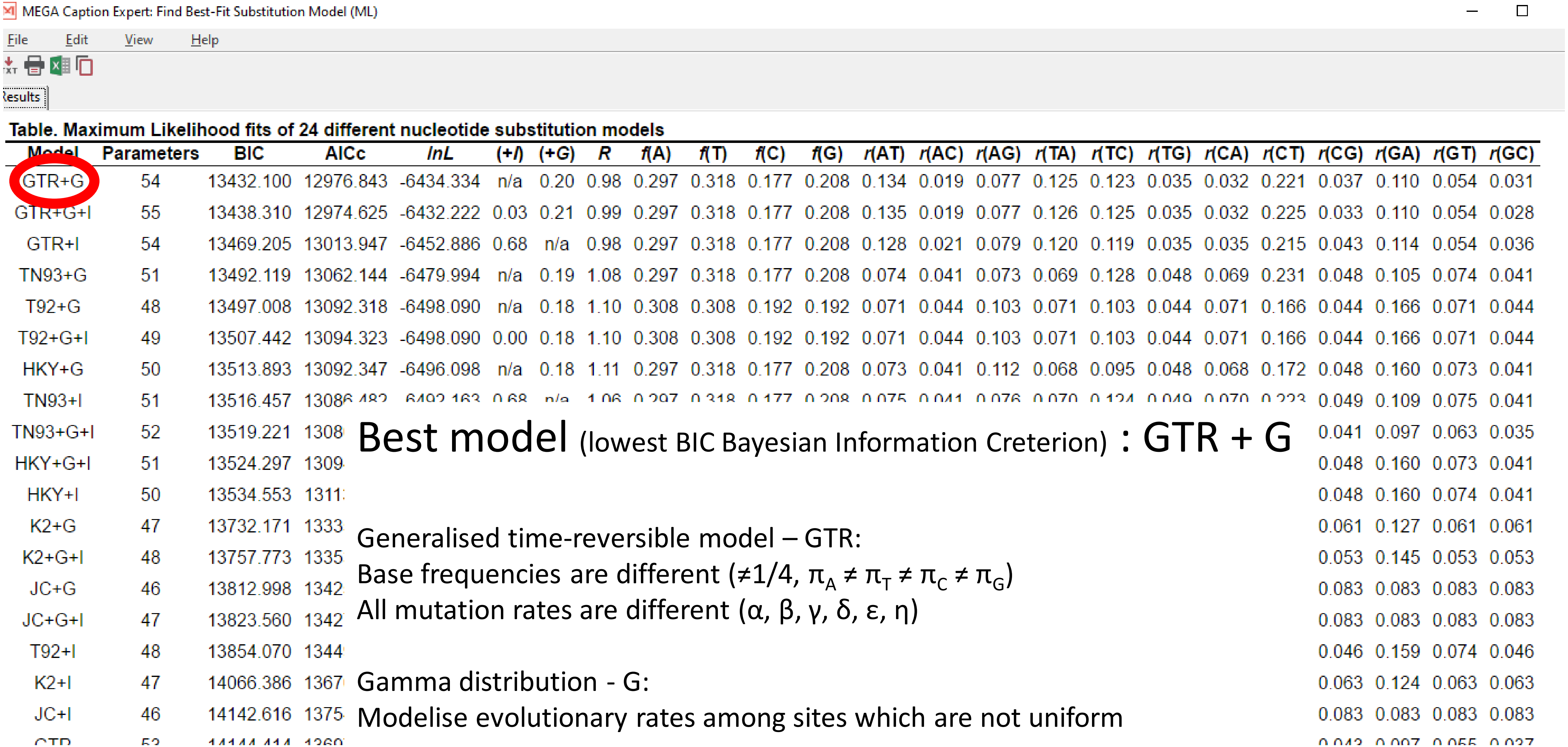


Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)	*r*(CT)	*r*(CG)	*r*(GA)	*r*(GT)	*r*(GC)
GTR+G	54	13432.100	12976.843	-6434.334	n/a	0.20	0.98	0.297	0.318	0.177	0.208	0.134	0.019	0.077	0.125	0.123	0.035	0.032	0.221	0.037	0.110	0.054	0.031
GTR+G+I	55	13438.310	12974.625	-6432.222	0.03	0.21	0.99	0.297	0.318	0.177	0.208	0.135	0.019	0.077	0.126	0.125	0.035	0.032	0.225	0.033	0.110	0.054	0.028
GTR+I	54	13469.205	13013.947	-6452.886	0.68	n/a	0.98	0.297	0.318	0.177	0.208	0.128	0.021	0.079	0.120	0.119	0.035	0.035	0.215	0.043	0.114	0.054	0.036
TN93+G	51	13492.119	13062.144	-6479.994	n/a	0.19	1.08	0.297	0.318	0.177	0.208	0.074	0.041	0.073	0.069	0.128	0.048	0.069	0.231	0.048	0.105	0.074	0.041
T92+G	48	13497.008	13092.318	-6498.090	n/a	0.18	1.10	0.308	0.308	0.192	0.192	0.071	0.044	0.103	0.071	0.103	0.044	0.071	0.166	0.044	0.166	0.071	0.044
T92+G+I	49	13507.442	13094.323	-6498.090	0.00	0.18	1.10	0.308	0.308	0.192	0.192	0.071	0.044	0.103	0.071	0.103	0.044	0.071	0.166	0.044	0.166	0.071	0.044
HKY+G	50	13513.893	13092.347	-6496.098	n/a	0.18	1.11	0.297	0.318	0.177	0.208	0.073	0.041	0.112	0.068	0.095	0.048	0.068	0.172	0.048	0.160	0.073	0.041
TN93+I	51	13516.457	13086.482	-6492.163	0.68	n/a	1.06	0.297	0.318	0.177	0.208	0.075	0.041	0.076	0.070	0.124	0.049	0.070	0.223	0.049	0.109	0.075	0.041
TN93+G+I	52	13519.221	1308																	0.041	0.097	0.063	0.035
HKY+G+I	51	13524.297	1309																	0.048	0.160	0.073	0.041
HKY+I	50	13534.553	1311																	0.048	0.160	0.074	0.041
K2+G	47	13732.171	1333																	0.061	0.127	0.061	0.061
K2+G+I	48	13757.773	1335																	0.053	0.145	0.053	0.053
JC+G	46	13812.998	1342																	0.083	0.083	0.083	0.083
JC+G+I	47	13823.560	1342																	0.083	0.083	0.083	0.083
T92+I	48	13854.070	1344																	0.046	0.159	0.074	0.046
K2+I	47	14066.386	1367																	0.063	0.124	0.063	0.063
JC+I	46	14142.616	1375																	0.083	0.083	0.083	0.083
GTR	53	14144.414	1360																	0.043	0.097	0.055	0.037

Best model (lowest BIC Bayesian Information Creterion) : GTR + G

Generalised time-reversible model – GTR:
Base frequencies are different (≠1/4, $\pi_A \neq \pi_T \neq \pi_C \neq \pi_G$)
All mutation rates are different (α, β, γ, δ, ε, η)

Gamma distribution - G:
Modelise evolutionary rates among sites which are not uniform

2.2 Phylogeny inference

- There are different methods to construct a phylogeny:
  - Distance methods
  - Parsimony methods
  - Likelihood methods

Depending on the method used, they can give different results

2.2 Phylogeny inference

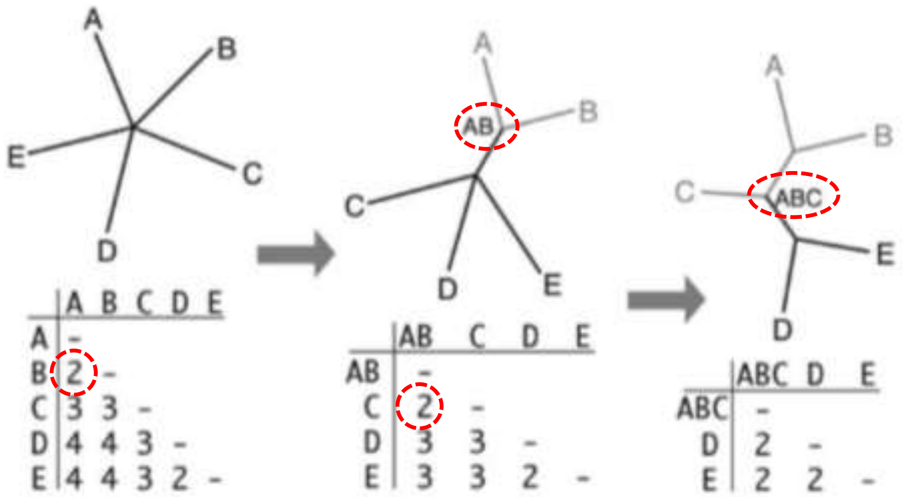
- Distance methods:
  - find a tree such that branch lengths of paths between sequences fit the matrix of pairwise distances
  - An example of distance method:

  Neighbor Joining:

  based on the principle of minimal evolution. Assumes that the best tree is the tree of smallest length.

  It starts with a tree star, then there is an iterative process to reach the smallest tree.

  - There are other methods (ex: UPGMA)



	A	B	C	D	E
A	-				
B	2	-			
C	3	3	-		
D	4	4	3	-	
E	4	4	3	2	-

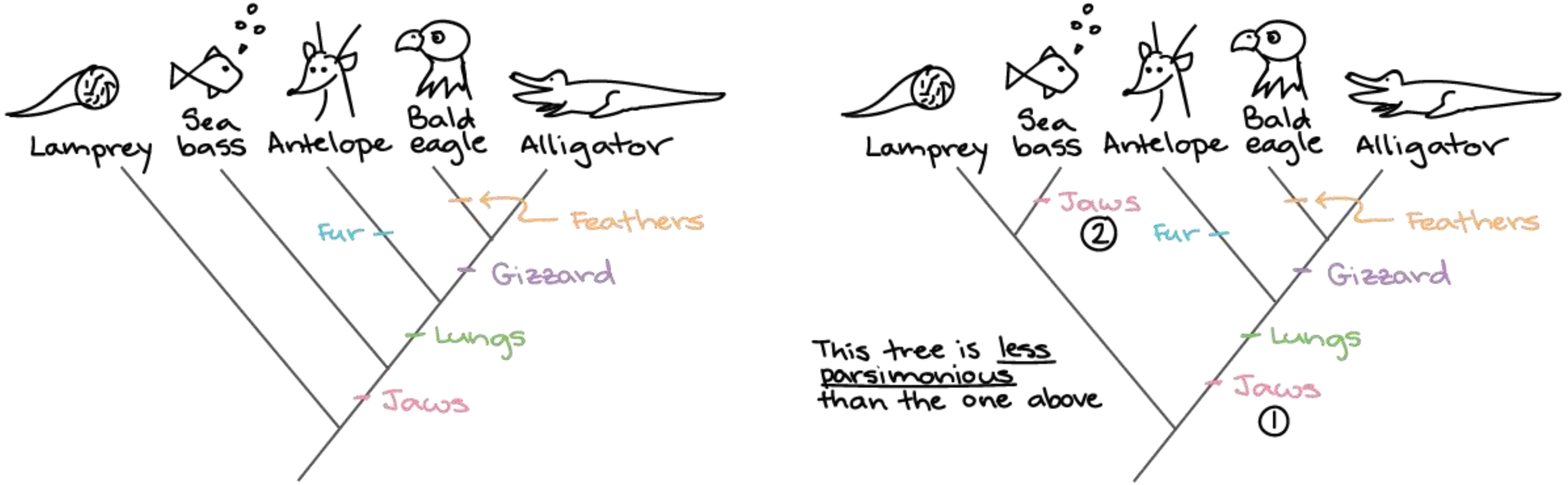
	AB	C	D	E
AB	-			
C	2	-		
D	3	3	-	
E	3	3	2	-

	ABC	D	E
ABC	-		
D	2	-	
E	2	2	-

Figure 8. Neighbor Joining algorithm: start from a star phylogeny (left); find the nearest pair of nodes (according to the distance matrix, either of A-B or D-E) (middle); recalculate the distance matrix using the new node (AB); repeat until the tree is fully resolved (right).

2.2 Phylogeny inference

- Maximum Parsimony methods:
  - The assumption is that the true evolutionary story is the one that involves the <u>fewest evolutionary events</u>
  - The objective is to identify the phylogenetic tree that requires the smallest total number of evolutionary events. There is an iterative process, the best tree is the one with the <u>maximum parsimony</u>



This tree is <u>less parsimonious</u> than the one above

Robert Bear et al., CC BY 4.0

2.2 Phylogeny inference

Maximum Likelihood method:

- What is likelihood ?
- Example with coin tossing:
  - p = proba of landing on head - H
  - 1-p = proba of landing on tail – T
  - p=0.5
  - 2 tossings: HH
    - $p^2.(1-p)^0$
    - $0.5^2.(1-0.5)^0=0.25$
  - 5 tossings: HHTTH
    - $p^3.(1-p)^2$
    - $0.5^3.(1-0.5)^2=0.03125$

This probability defines de likelihood function:

$L(p) = p^h.(1-p)^{n-h}$

with n nb of tossings, h nb of heads

- If we don't know p, some values of p will generate the observed data (ex. HHTTH) with higher probabilities. The highest probability will be obtained with p=0.5.
- How can we find p to maximize L(p)?

The solution is: p= h/n

- This is the maximum likelihood estimate (MLE)

- In evolution, point mutations are considered chance events, just like tossing a coin. Therefore, the probability of finding a mutation along one branch in a phylogenetic tree can be calculated by using the same maximum likelihood framework.

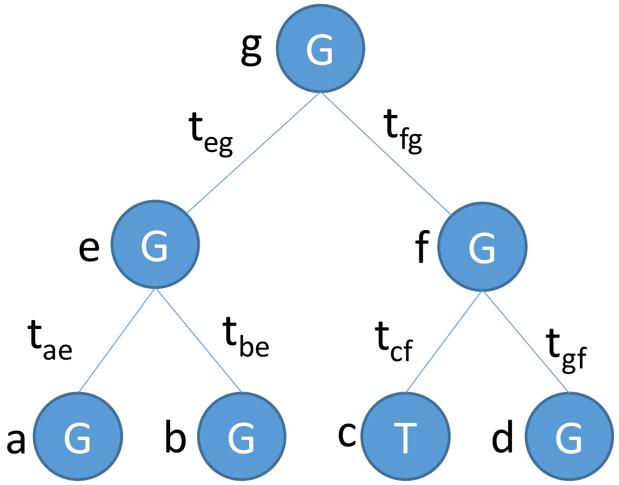
2.2 Phylogeny inference

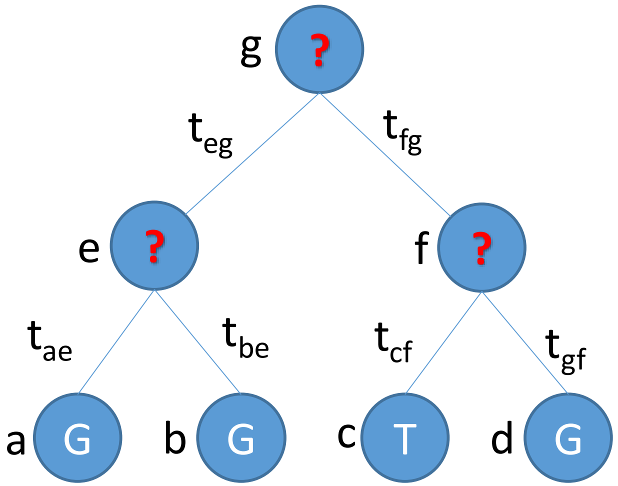
- Maximum Likelihood method:
  - This method compares phylogenetic trees on the basis of their ability to predict the observed data. The tree that has the highest probability of producing the observed sequences is preferred.
  - More in details:
    - the nucleotides of all sequences at each site are considered separately
    - the likelihood of having these bases are computed for a given topology by using the same evolutionary model (ex. GTR+G).

Suppose we have:

- A fixed topology
- Observed data (a:G, b:G, c:T, d: G)
- Ancestral states
- Mutation rates t

Likelihood = Pr(g:G).
Pr(e:GIg:G, t_{eg}). Pr(f:GIg:G, t_{fg})
Pr(a:GIe:G, t_{ae}). Pr(b:GIe:G, t_{be})
Pr(c:GIf:G, t_{cf}). Pr(d:GIf:G, t_{df})
-conditional probabilities are used from (GTR+G) model





Other tree topology ?

2.2 Phylogeny inference

- Maximum Likelihood method:
  - This method compares phylogenetic trees on the basis of their ability to predict the observed data. The tree that has the highest probability of producing the observed sequences is preferred.
  - More in details:
    - the nucleotides of all sequences at each site are considered separately
    - the likelihood of having these bases are computed for a given topology by using the same evolutionary model (ex. GTR+G).
    - This likelihood is added for all sites, and the sum of the likelihood is maximized to estimate the branch length of the tree.
    - This procedure is repeated for all possible topologies, and the topology that shows the highest likelihood is chosen as the final tree.
    - Number of topologies is factorial of the number n of sequences: (2n – 5)!!

- Problem: long to compute (need to calculate for all tree topologies possible), but very robust (no assumptions behind).
- Solution: some heuristics (simplifications) are needed especially when large trees are inferred (for instance we can set the initial tree from a neighbor joining tree)

  >RaxML (Randomized Axelerated Maximum Likelihood)

MEGA Molecular Evolutionary Genetics Analysis

2.2 Phylogeny inference

- Let’s infer a ML phylogeny in MEGA-X software
- Open « rbcl-diatbarcode.fasta »
- Analysis > Phylogeny > Construct test Maximum Likelihood tree

Use in the substitution model « GTR »

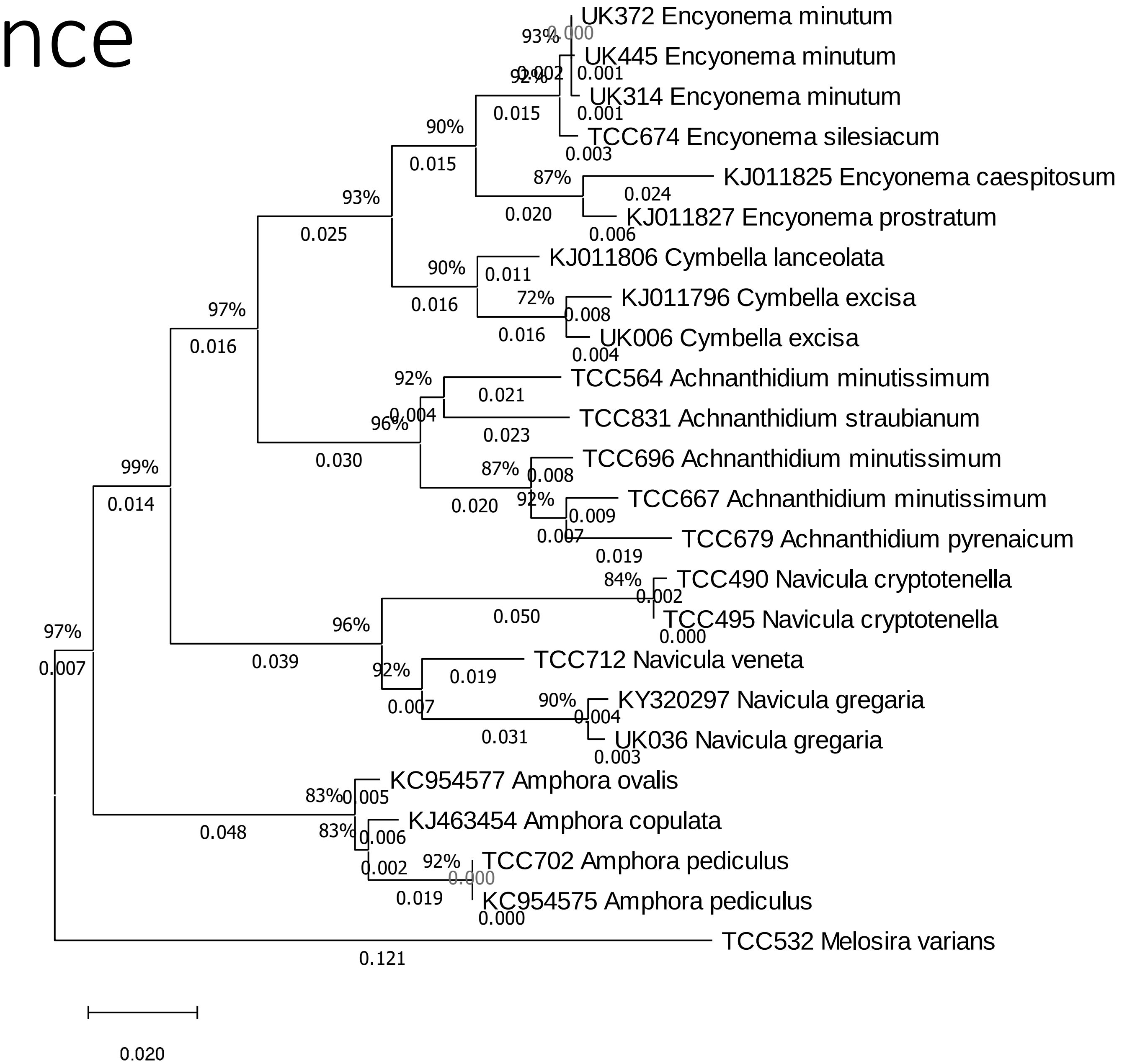
Use in the rates and patterns
« Gamma distributed»

Bootstrap = 50 (this is quite low, usually, use 100 and even more)

Result in Mega-X:

Use “original tree” (=best tree).

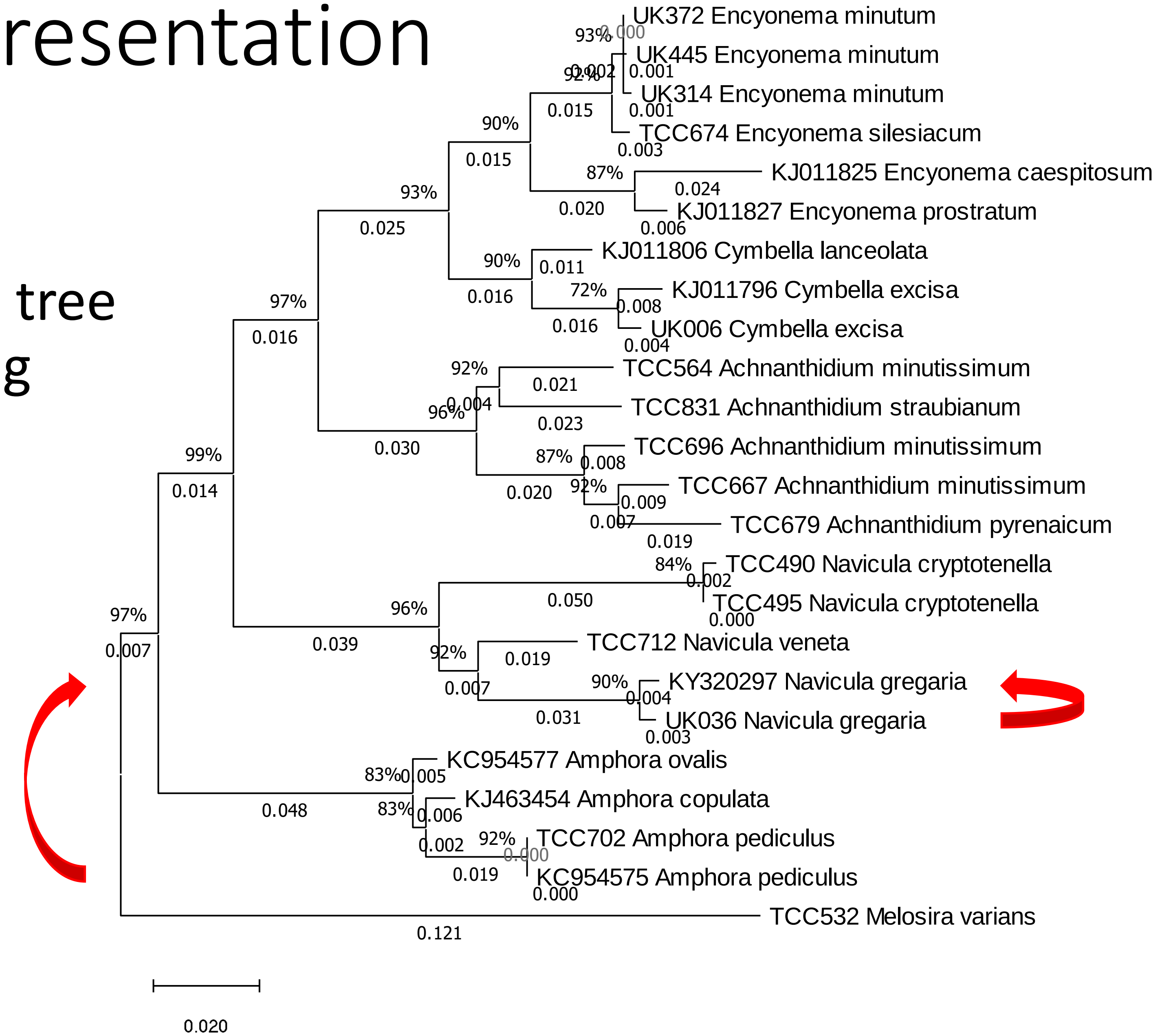
Don’t use the “Bootstrat consensus tree” (= not all trees have the same topology, the consensus tree summarize the most frequently observed topologies)



2.3 Graphical representation

- Interpret tree topology

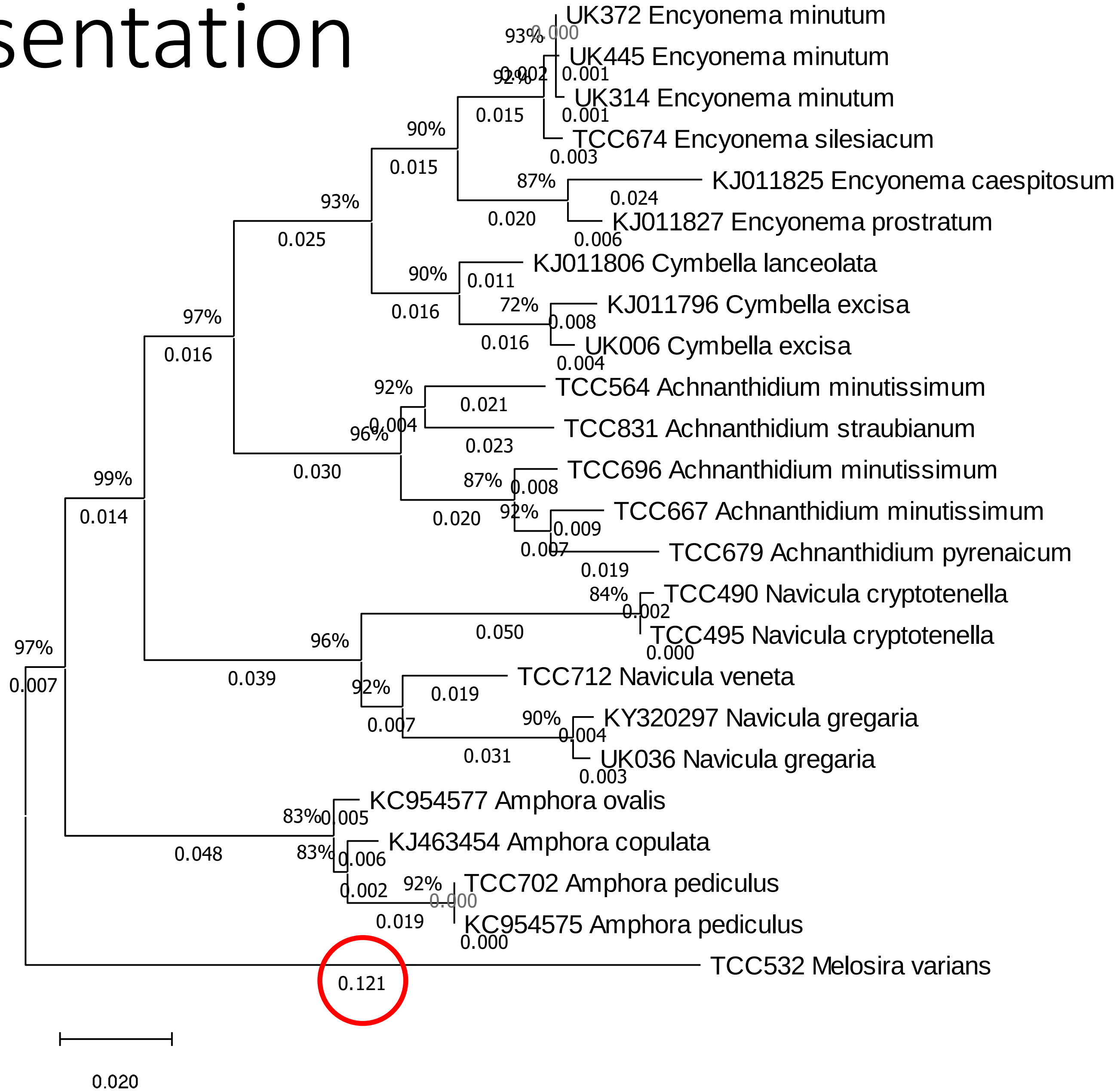
You can flip the leaves, the tree will have the same meaning



2.3 Graphical representation

- Interpret tree topology

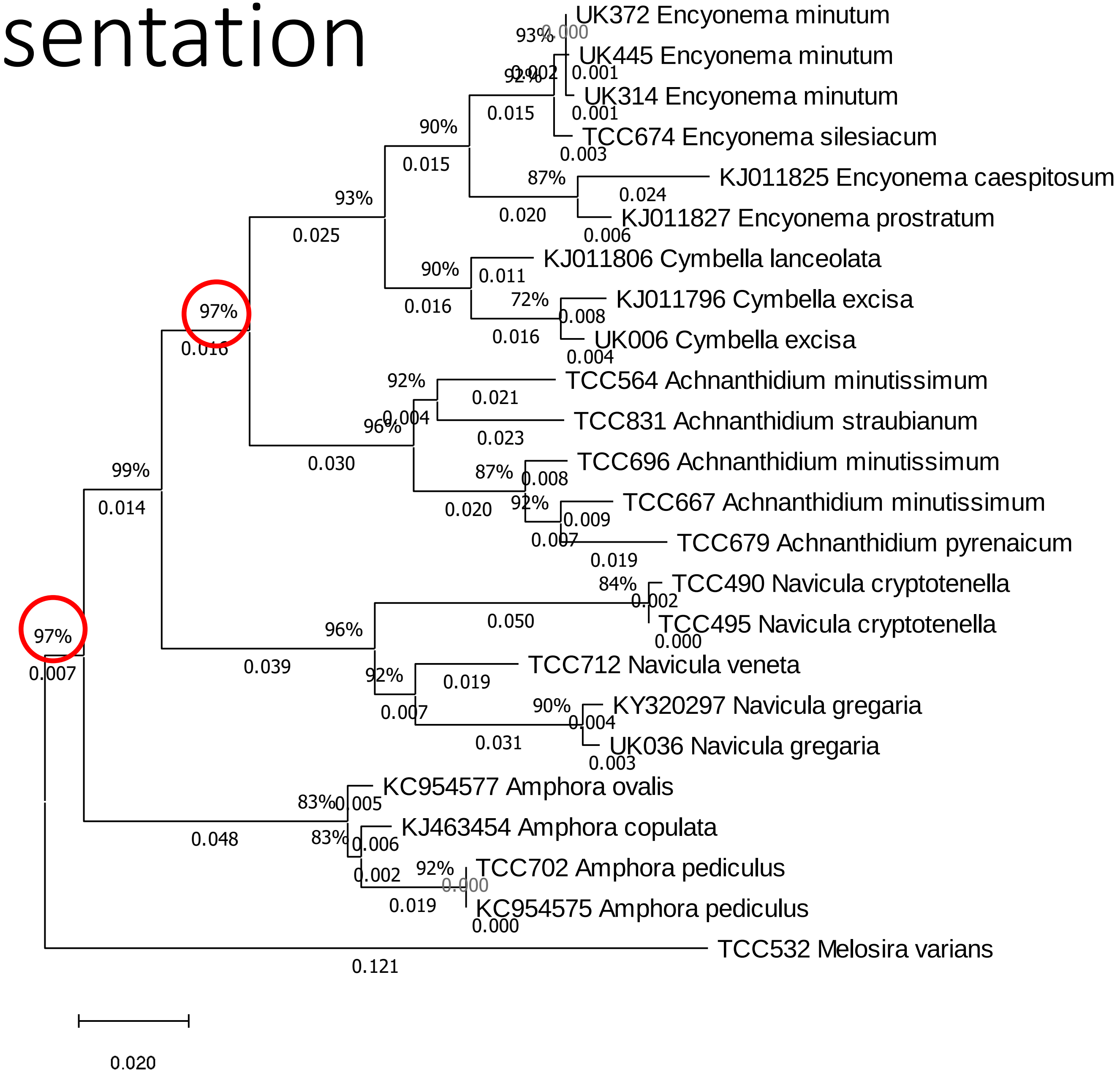
Branch length indicate genetic change : 0,1 = 0,1 substitution/site if we use the simplest substitution model (equal prob. of substitutions)



2.3 Graphical representation

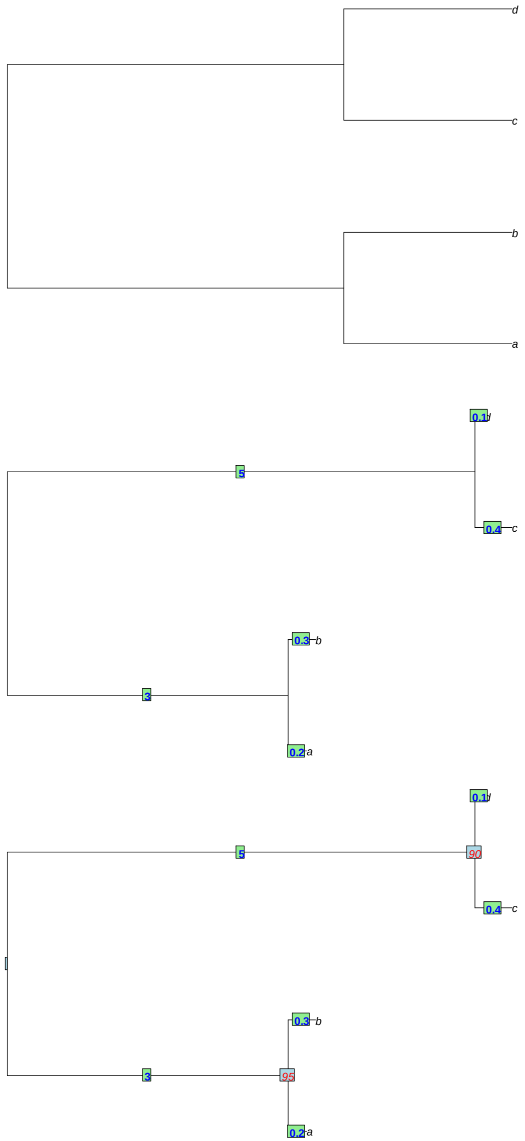
- Interpret tree topology

Bootstrap value: number of time this node was found during the iterations



2.3 Graphical representation

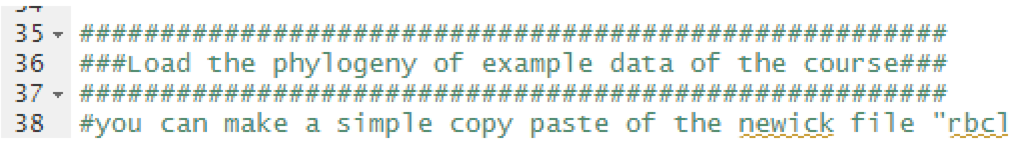
- in R studio
- Open “phylogeny.R”
- Discover Newick format
- All trees follow the newick standard
  - Simple tree: ((a,b),(c,d));
  - With branch lengths: ((a:0.2,b:0.3):3,(c:0.4,d:0.1):5);
  - With bootstrap: ((a:0.2,b:0.3)95:3,(c:0.4,d:0.1)90:5);



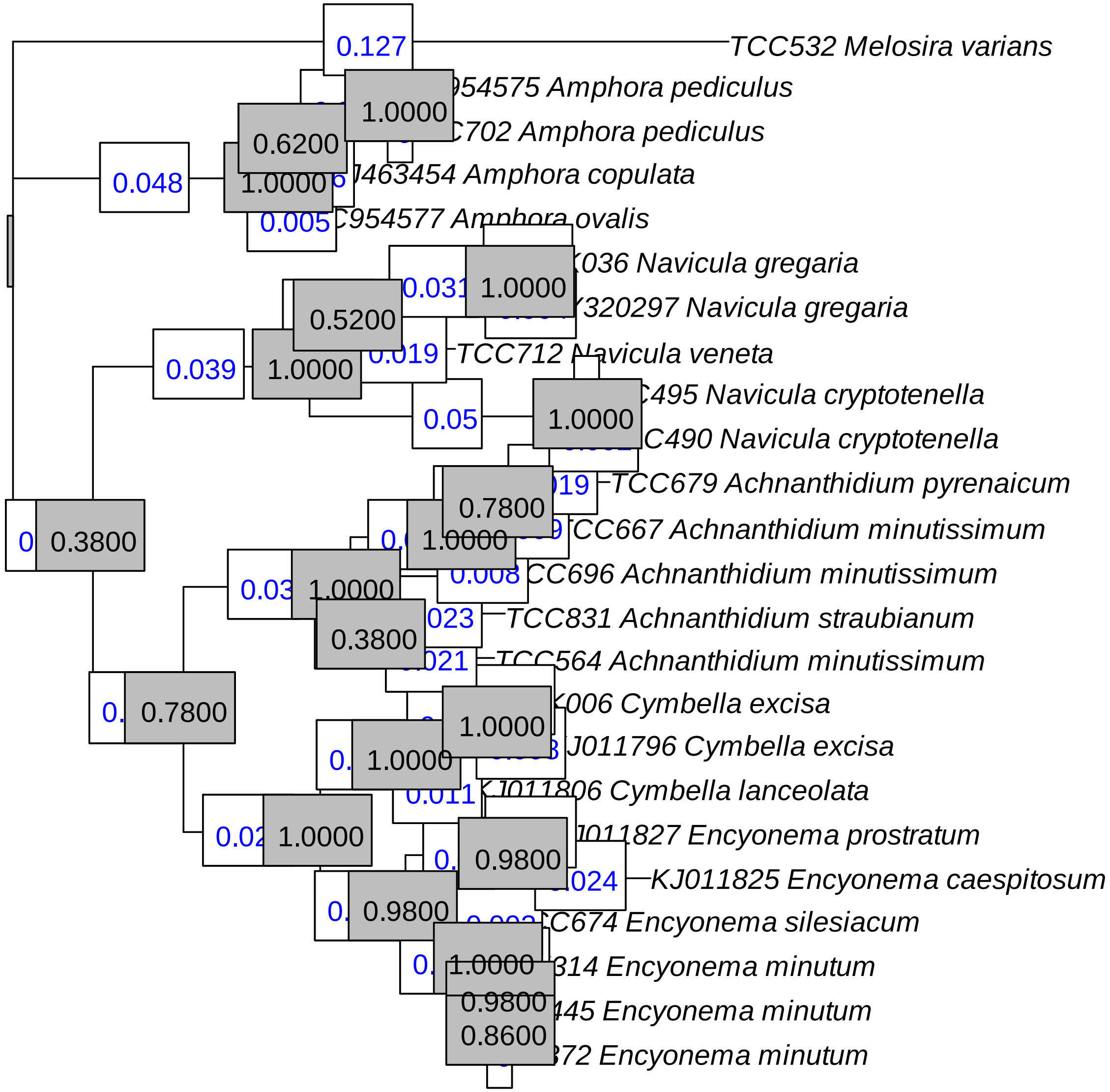
```
####################################
###   Discover newick format    #####
####################################
```

2.3 Graphical representation

- Load the phylogeny inferred in MEGA-X with R

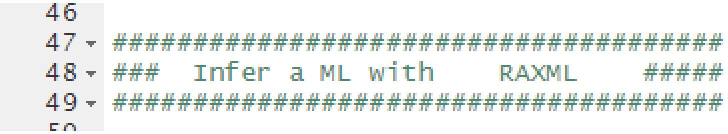


```
######################################################
###Load the phylogeny of example data of the course###
######################################################
#you can make a simple copy paste of the newick file "rbcl
```

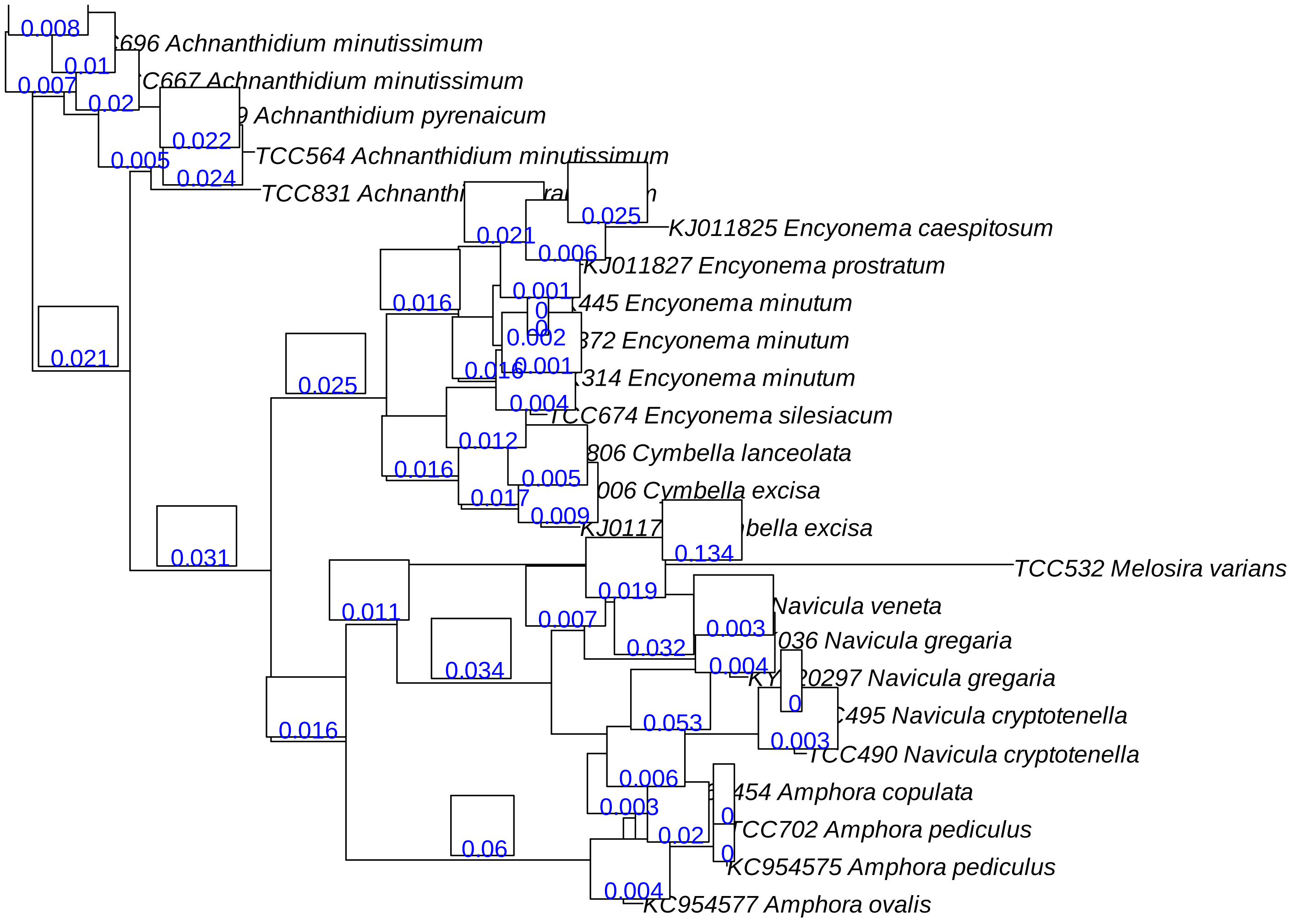


RaxML inference in R

- Infer a ML with RaxML



```
#####################################
###  Infer a ML with    RAXML   #####
#####################################
```



Schedule

BIOLAWEB

0. Including phylogenies into ecological studies

- 0.1 Phylogenetic diversity
- 0.2 Measuring phylogenetic signal
- 0.3 Measuring and testing community phylogenetic structure

1. Theory of sequence alignment

- 1.1 What is an alignment?
- 1.2 Objective of aligning sequences
- 1.3 Edit distance
- 1.4 Local alignments
- 1.5 Global alignments

2. Phylogenies

- 2.1 Choose appropriate model of sequence evolution
- 2.2 Phylogeny inference
- 2.3 Graphical representation of phylogenies

3. Phylogenetic placement

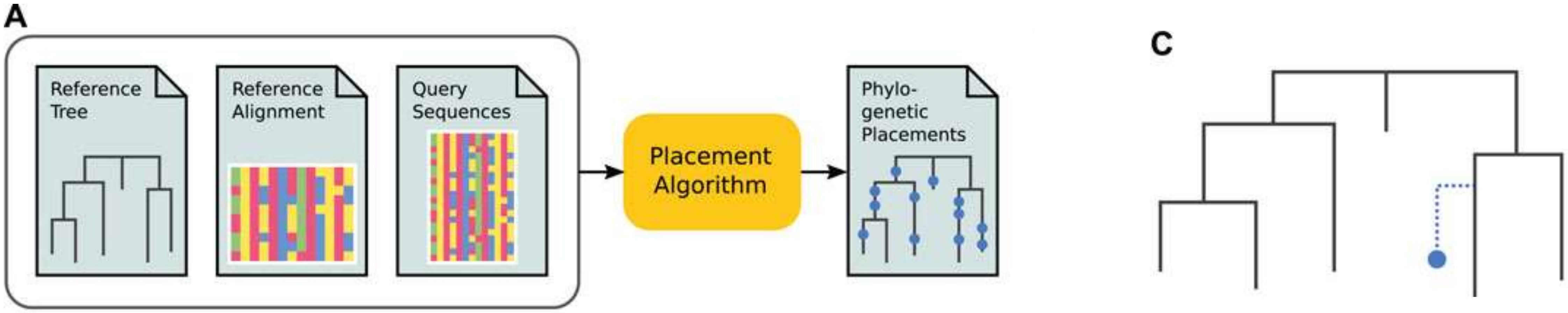
- 3.1 Definition
- 3.2 Why using this
- 3.3 Example with RaxML in R

4. Ecophylogenetic training in R (R studio)

- PD
- NRI NTI

3.1 Definition

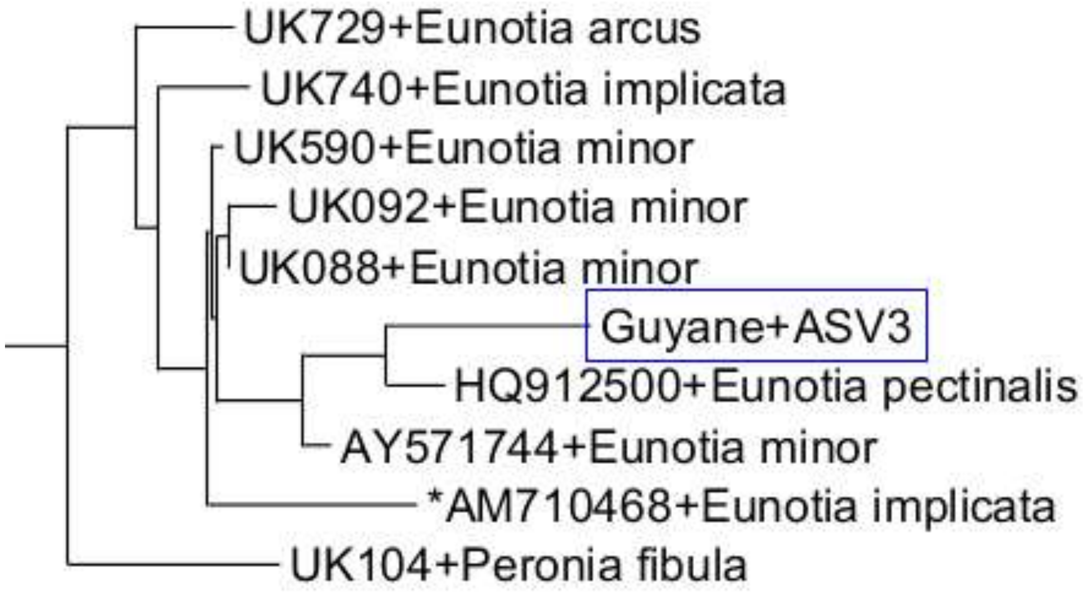
- Phylogenetic placement: A family of methods to place query sequences onto the branches of a reference tree
  - Query sequence: a sequence to be placed in a tree. Typically: short sequences from metabarcoding
  - Reference tree: the phylogenetic tree used to place the queries, inferred with ML and (usually) long sequences



Czech, L., Stamatakis, A., Dunthorn, M., Barbera, P., 2022. Metagenomic Analysis Using Phylogenetic Placement—A Review of the First Decade. Frontiers in Bioinformatics 2.

3.2 Why using phylogenetic placement?

- To investigate the taxonomic composition of samples:
    - can be an alternative to DADA2 taxonomic assignation

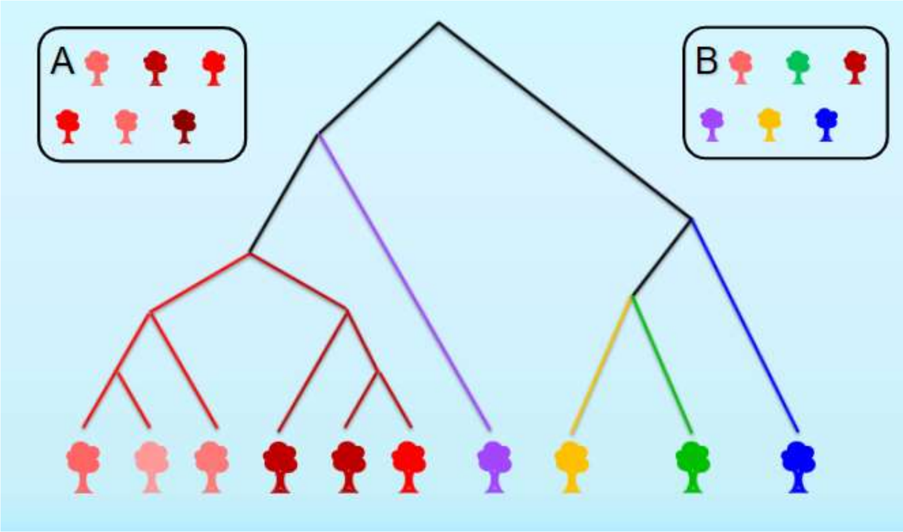


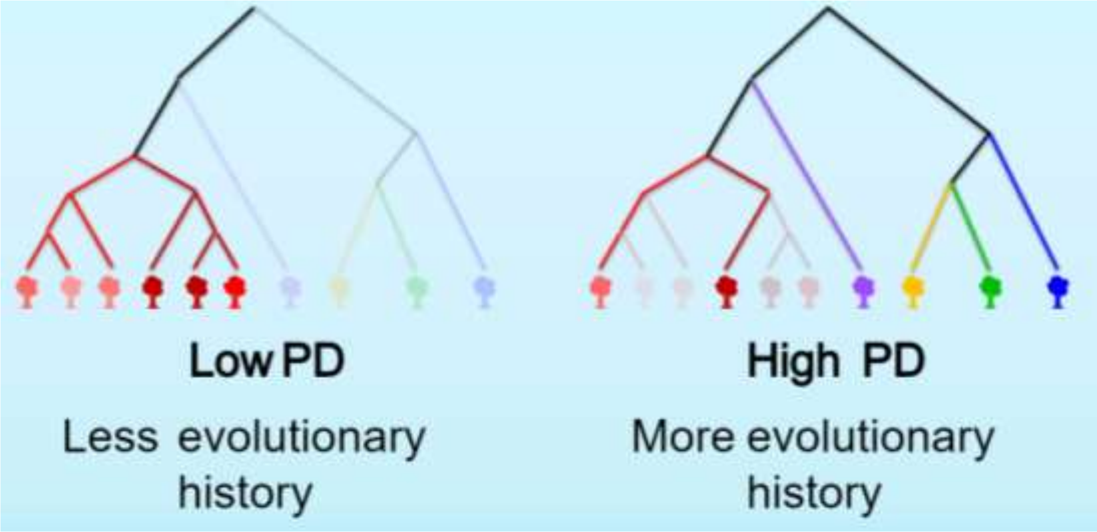
3.2 Why using phylogenetic placement?

- For ecological studies

  Phylogenetic diversity

  > integration of the phylogenetic dimension in diversity metrics





3.2 Why using phylogenetic placement?

- For ecological studies

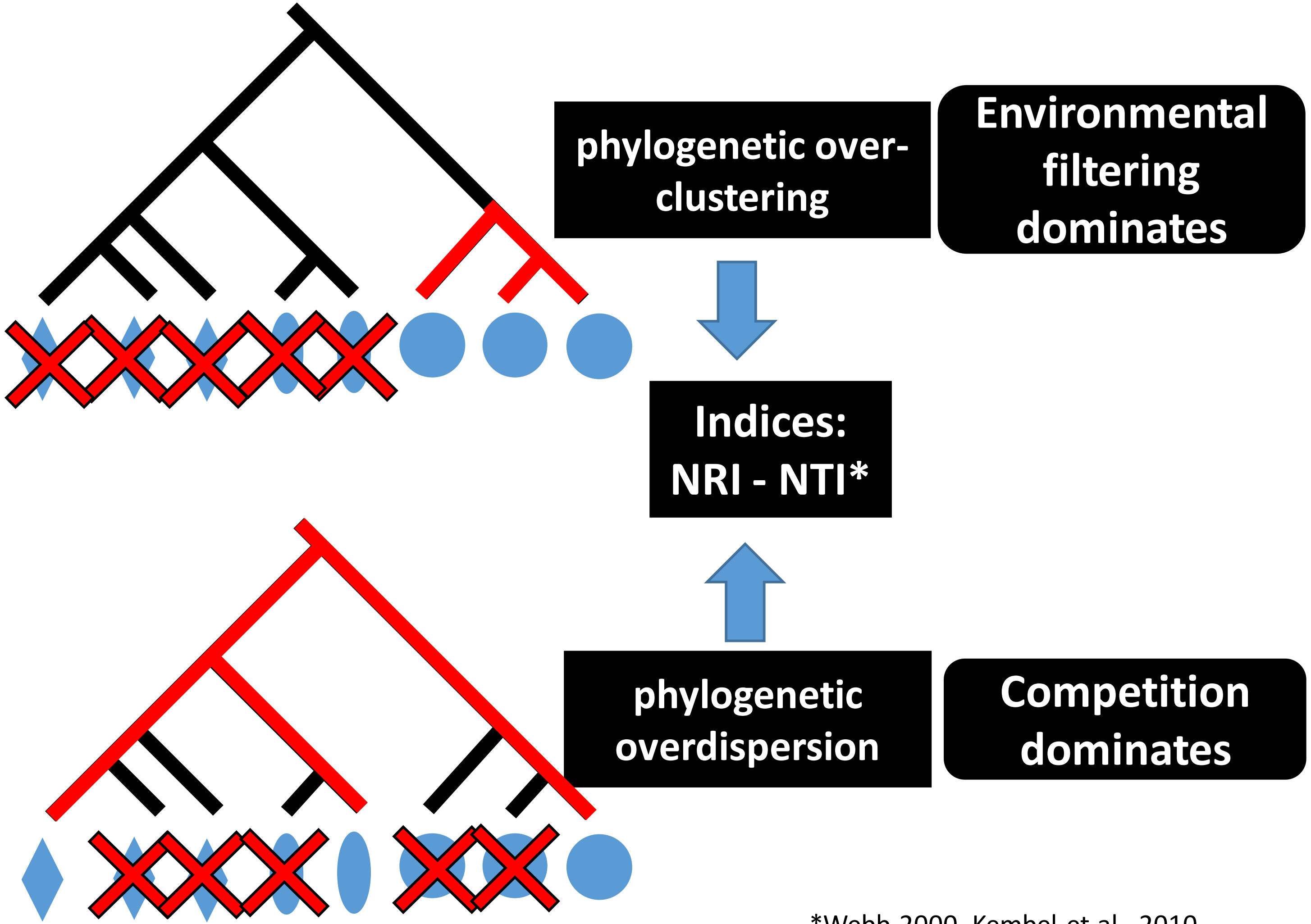
> If there is a niche conservatism in the evolution,

> Phylogenetic structure of samples can be interpreted in terms of ecological processes

Environmental filtering

vs

Competitive exclusion



phylogenetic over-clustering

Environmental filtering dominates

Indices: NRI - NTI*

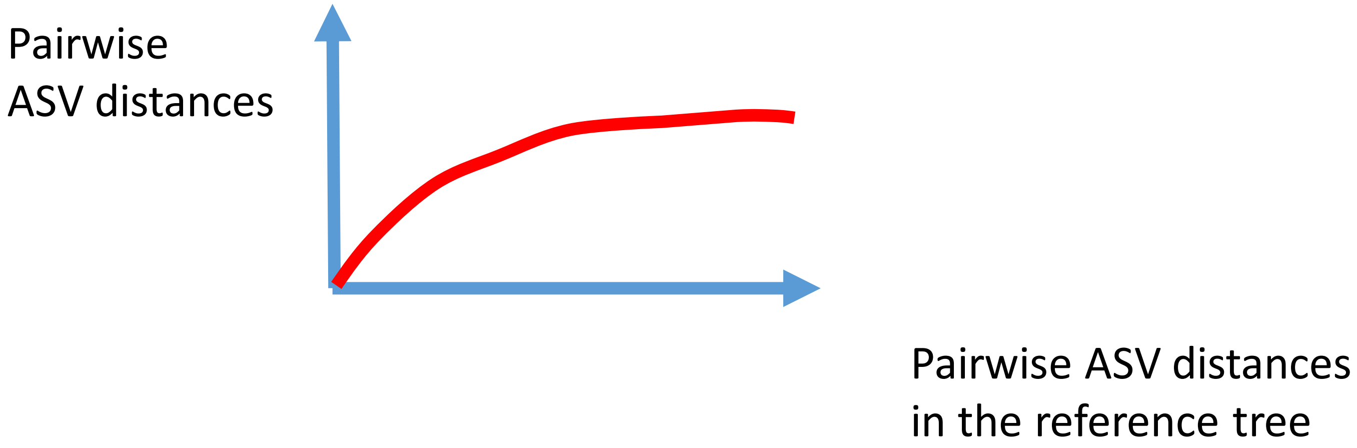
phylogenetic overdispersion

Competition dominates

*Webb 2000, Kembel et al., 2010

3.2 Why using phylogenetic placement?

- For metabarcoding studies, we need to place our ASV in a reference tree and extract their pairwise phylogenetic distances
- It is not possible to calculate directly the distances between ASV because they are too short >> underestimation of the distances for phylogenetically distant ASV.



Which algorithms?

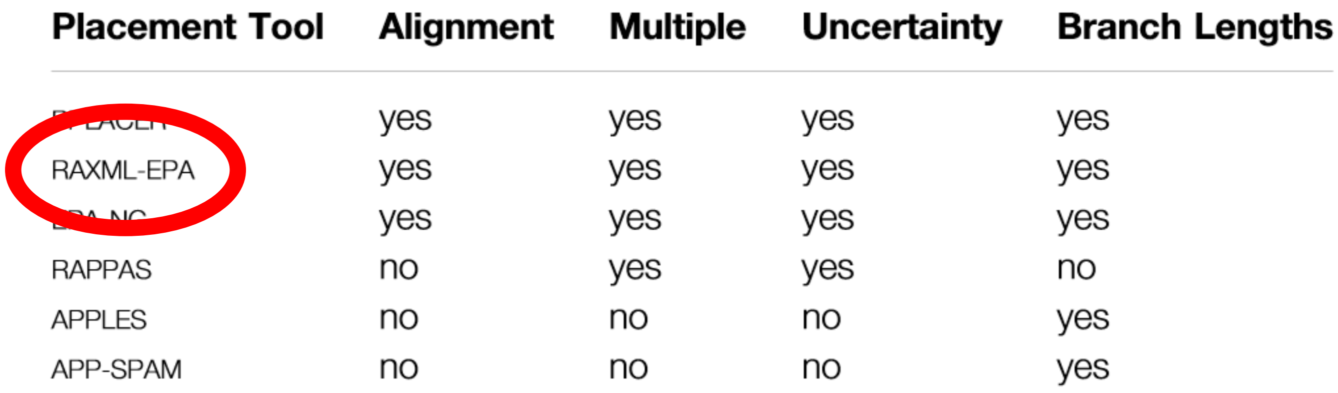
- Several algorithms exist

frontiers | Frontiers in Bioinformatics

REVIEW
published: 26 May 2022
doi: 10.3389/fbinf.2022.871393

Metagenomic Analysis Using Phylogenetic Placement—A Review of the First Decade

Lucas Czech[1], Alexandros Stamatakis[2,3], Micah Dunthorn[4] and Pierre Barbera[5]**



Placement Tool	Alignment	Multiple	Uncertainty	Branch Lengths
PPLACER	yes	yes	yes	yes
RAXML-EPA	yes	yes	yes	yes
EPA-NG	yes	yes	yes	yes
RAPPAS	no	yes	yes	no
APPLES	no	no	no	yes
APP-SPAM	no	no	no	yes

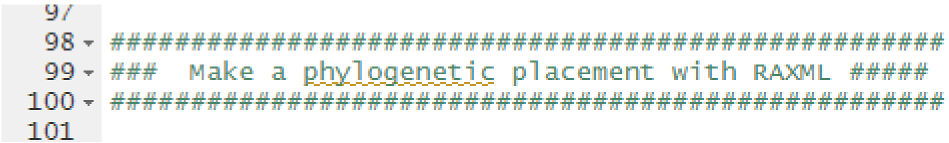
3.3 Example with RaxML in R

- Go back to the script and go to

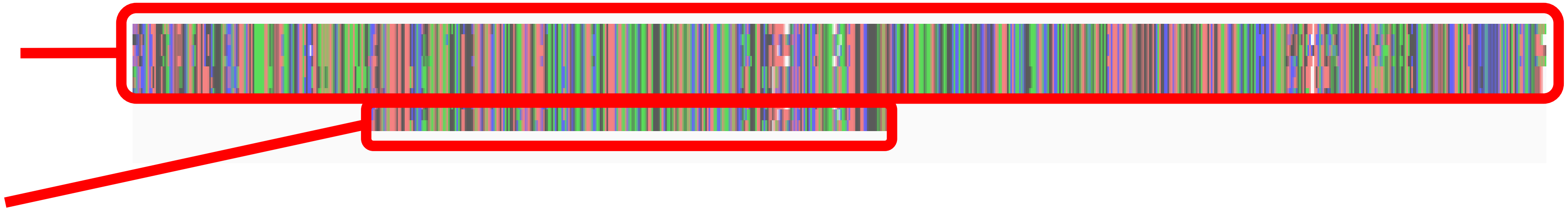
"Make a phylogenetic placement with RAXML"

We will use this alignment:

« rbcl-diatbarcode-ASV.fasta »



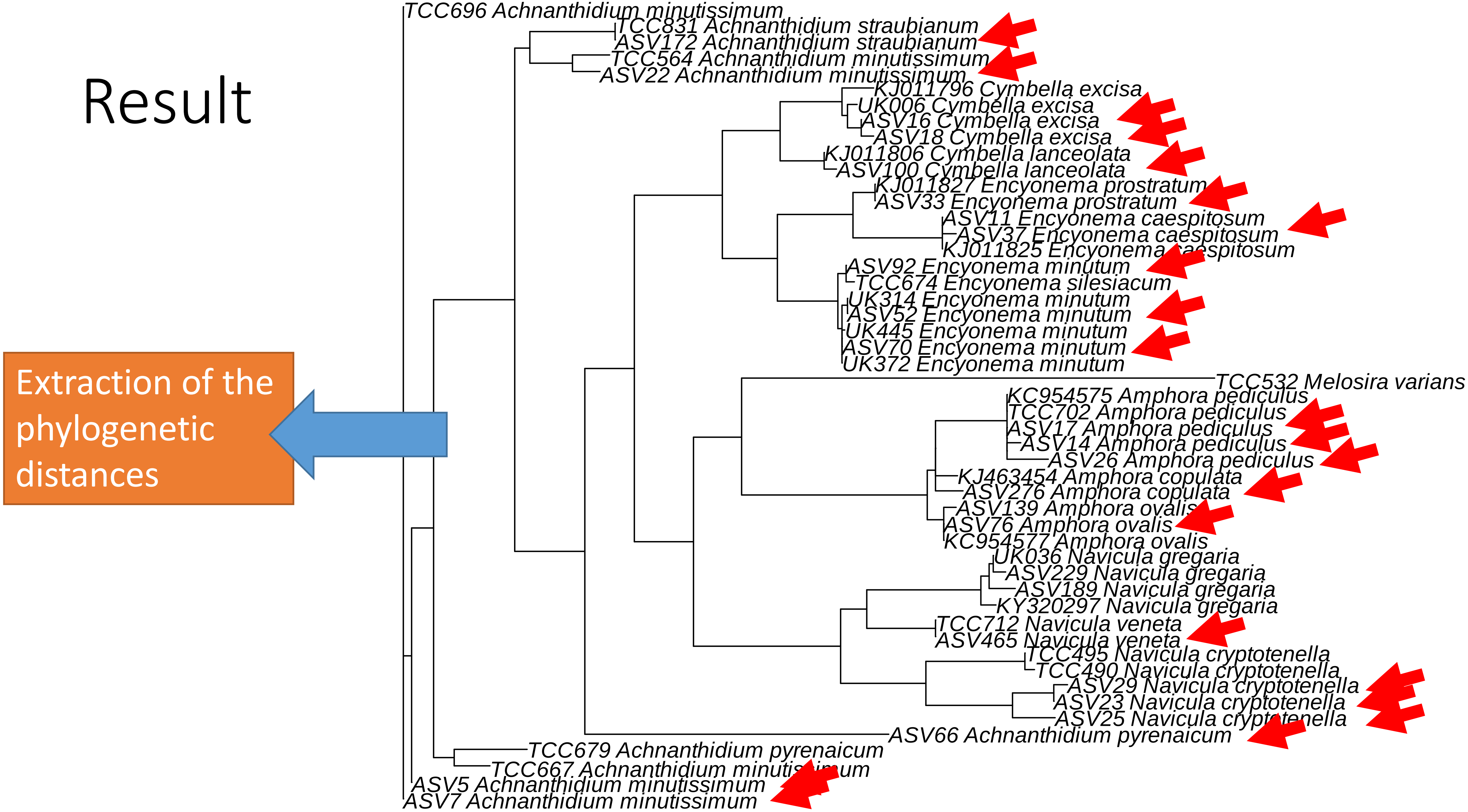
```
####################################################
###  Make a phylogenetic placement with RAXML #####
####################################################
```



Sequences used in the reference tree > 1000 bp

Query sequences 263 bp

Result



Extraction of the phylogenetic distances

TCC696 Achnanthidium minutissimum
TCC831 Achnanthidium straubianum
ASV172 Achnanthidium straubianum
TCC564 Achnanthidium minutissimum
ASV22 Achnanthidium minutissimum
KJ011796 Cymbella excisa
UK006 Cymbella excisa
ASV16 Cymbella excisa
ASV18 Cymbella excisa
KJ011806 Cymbella lanceolata
ASV100 Cymbella lanceolata
KJ011827 Encyonema prostratum
ASV33 Encyonema prostratum
ASV11 Encyonema caespitosum
ASV37 Encyonema caespitosum
KJ011825 Encyonema caespitosum
ASV92 Encyonema minutum
TCC674 Encyonema silesiacum
UK314 Encyonema minutum
ASV52 Encyonema minutum
UK445 Encyonema minutum
ASV70 Encyonema minutum
UK372 Encyonema minutum
TCC532 Melosira varians
KC954575 Amphora pediculus
TCC702 Amphora pediculus
ASV17 Amphora pediculus
ASV14 Amphora pediculus
ASV26 Amphora pediculus
KJ463454 Amphora copulata
ASV276 Amphora copulata
ASV139 Amphora ovalis
ASV76 Amphora ovalis
KC954577 Amphora ovalis
UK036 Navicula gregaria
ASV229 Navicula gregaria
ASV189 Navicula gregaria
KY320297 Navicula gregaria
TCC712 Navicula veneta
ASV465 Navicula veneta
TCC495 Navicula cryptotenella
TCC490 Navicula cryptotenella
ASV29 Navicula cryptotenella
ASV23 Navicula cryptotenella
ASV25 Navicula cryptotenella
ASV66 Achnanthidium pyrenaicum
TCC679 Achnanthidium pyrenaicum
TCC667 Achnanthidium minutissimum
ASV5 Achnanthidium minutissimum
ASV7 Achnanthidium minutissimum

Schedule

0. Including phylogenies into ecological studies

0.1 Phylogenetic diversity

0.2 Measuring phylogenetic signal

0.3 Measuring and testing community phylogenetic structure

1. Theory of sequence alignment

1.1 What is an alignment?

1.2 Objective of aligning sequences

1.3 Edit distance

1.4 Local alignments

1.5 Global alignments

2. Phylogenies

2.1 Choose appropriate model of sequence evolution

2.2 Phylogeny inference

2.3 Graphical representation of phylogenies

3. Phylogenetic placement

3.1 Definition

3.2 Why using this

3.3 Example with RaxML in R

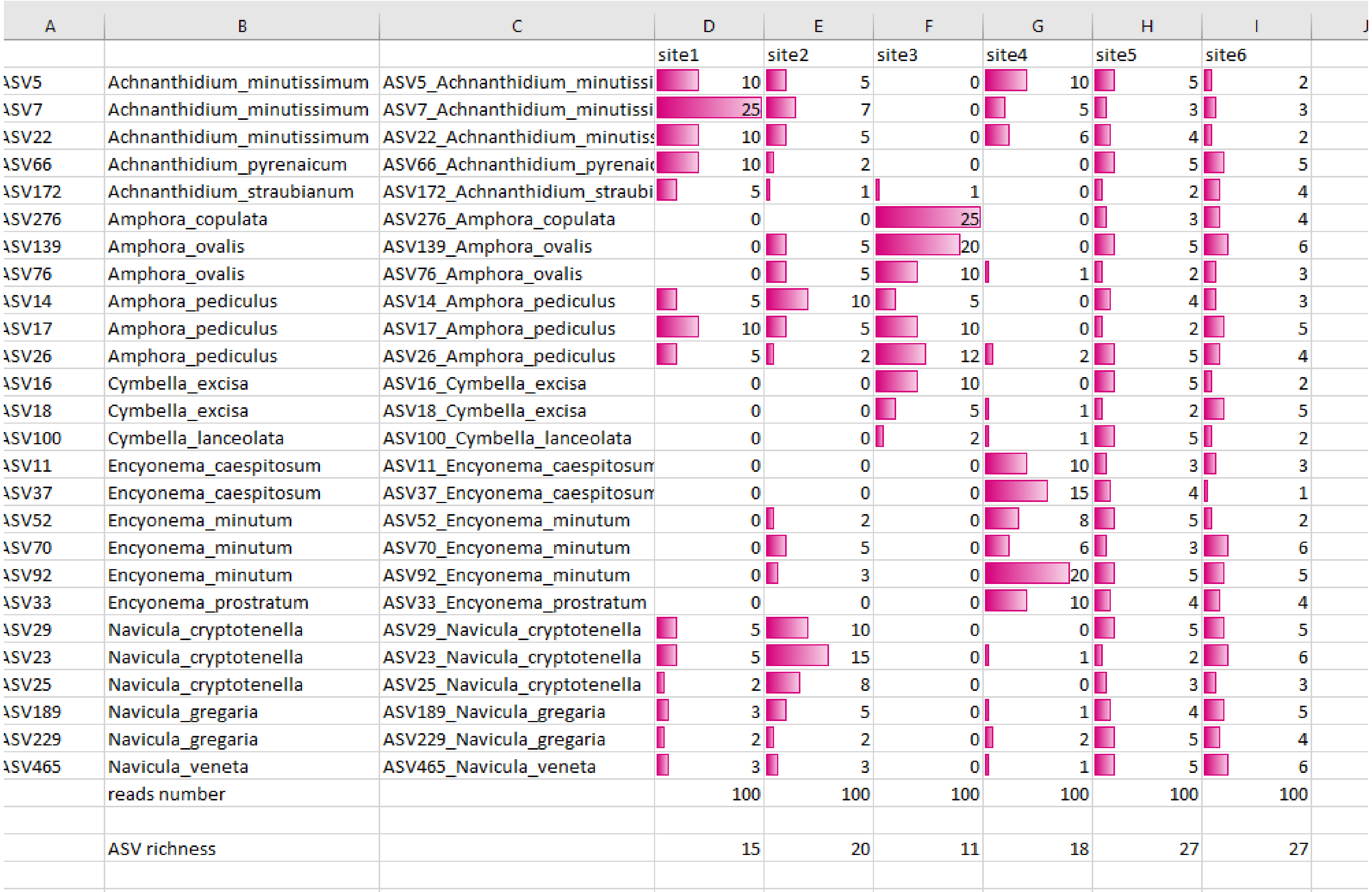
4. Ecophylogenetic training in R (R studio)

NRI NTI

PD

Funded by European Union

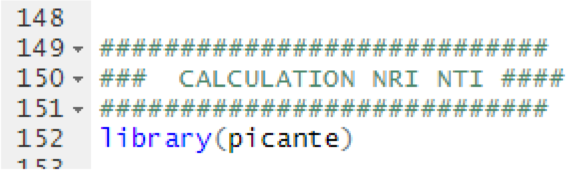
Example on mock communities



A	B	C	D	E	F	G	H	I	J
			site1	site2	site3	site4	site5	site6	
ASV5	Achnanthidium_minutissimum	ASV5_Achnanthidium_minutissi	10	5	0	10	5	2	
ASV7	Achnanthidium_minutissimum	ASV7_Achnanthidium_minutissi	25	7	0	5	3	3	
ASV22	Achnanthidium_minutissimum	ASV22_Achnanthidium_minutiss	10	5	0	6	4	2	
ASV66	Achnanthidium_pyrenaicum	ASV66_Achnanthidium_pyrenaic	10	2	0	0	5	5	
ASV172	Achnanthidium_straubianum	ASV172_Achnanthidium_straubi	5	1	1	0	2	4	
ASV276	Amphora_copulata	ASV276_Amphora_copulata	0	0	25	0	3	4	
ASV139	Amphora_ovalis	ASV139_Amphora_ovalis	0	5	20	0	5	6	
ASV76	Amphora_ovalis	ASV76_Amphora_ovalis	0	5	10	1	2	3	
ASV14	Amphora_pediculus	ASV14_Amphora_pediculus	5	10	5	0	4	3	
ASV17	Amphora_pediculus	ASV17_Amphora_pediculus	10	5	10	0	2	5	
ASV26	Amphora_pediculus	ASV26_Amphora_pediculus	5	2	12	2	5	4	
ASV16	Cymbella_excisa	ASV16_Cymbella_excisa	0	0	10	0	5	2	
ASV18	Cymbella_excisa	ASV18_Cymbella_excisa	0	0	5	1	2	5	
ASV100	Cymbella_lanceolata	ASV100_Cymbella_lanceolata	0	0	2	1	5	2	
ASV11	Encyonema_caespitosum	ASV11_Encyonema_caespitosum	0	0	0	10	3	3	
ASV37	Encyonema_caespitosum	ASV37_Encyonema_caespitosum	0	0	0	15	4	1	
ASV52	Encyonema_minutum	ASV52_Encyonema_minutum	0	2	0	8	5	2	
ASV70	Encyonema_minutum	ASV70_Encyonema_minutum	0	5	0	6	3	6	
ASV92	Encyonema_minutum	ASV92_Encyonema_minutum	0	3	0	20	5	5	
ASV33	Encyonema_prostratum	ASV33_Encyonema_prostratum	0	0	0	10	4	4	
ASV29	Navicula_cryptotenella	ASV29_Navicula_cryptotenella	5	10	0	0	5	5	
ASV23	Navicula_cryptotenella	ASV23_Navicula_cryptotenella	5	15	0	1	2	6	
ASV25	Navicula_cryptotenella	ASV25_Navicula_cryptotenella	2	8	0	0	3	3	
ASV189	Navicula_gregaria	ASV189_Navicula_gregaria	3	5	0	1	4	5	
ASV229	Navicula_gregaria	ASV229_Navicula_gregaria	2	2	0	2	5	4	
ASV465	Navicula_veneta	ASV465_Navicula_veneta	3	3	0	1	5	6	
	reads number		100	100	100	100	100	100	
	ASV richness		15	20	11	18	27	27	

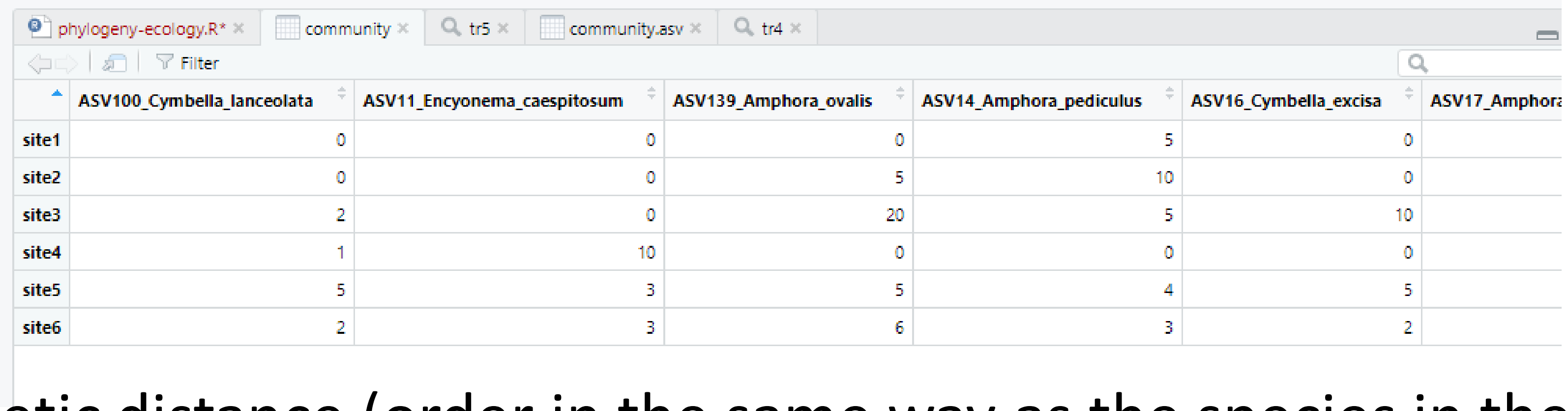
Calculation of NRI NTI

- Back to R



```
############################
###  CALCULATION NRI NTI ####
############################
library(picante)
```

- Load the file « community-asv.csv »

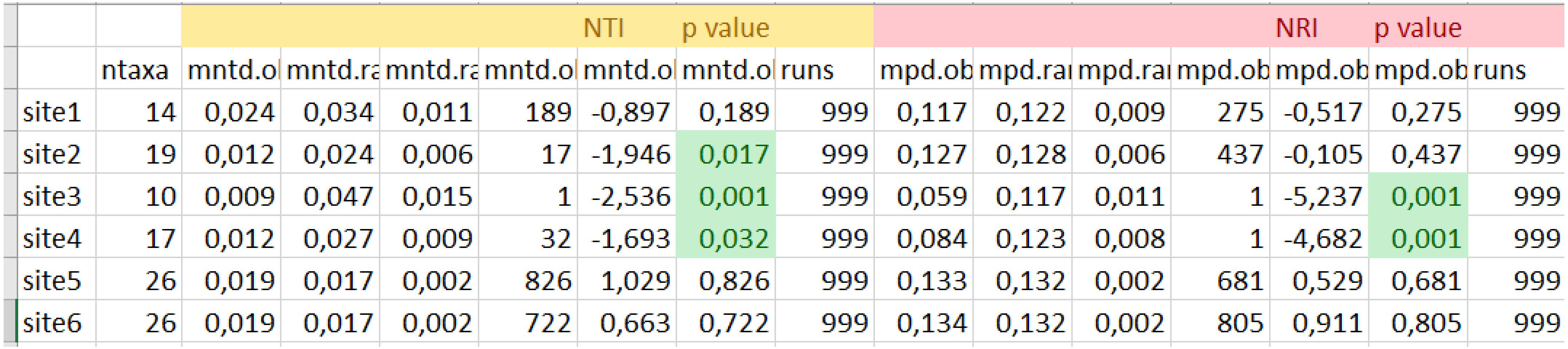


	ASV100_Cymbella_lanceolata	ASV11_Encyonema_caespitosum	ASV139_Amphora_ovalis	ASV14_Amphora_pediculus	ASV16_Cymbella_excisa	ASV17_Amphora
site1	0	0	0	5	0	
site2	0	0	5	10	0	
site3	2	0	20	5	10	
site4	1	10	0	0	0	
site5	5	3	5	4	5	
site6	2	3	6	3	2	

- Load the phylogenetic distance (order in the same way as the species in the community data) « distfromtree_asv-ord.csv »

Calculation of NRI NTI

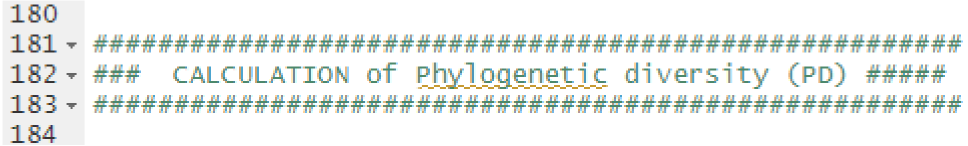
- Look at the NTI.csv and NRI.csv



						NTI	p value						NRI	p value	
	ntaxa	mntd.ol	mntd.ra	mntd.ra	mntd.ol	mntd.ol	mntd.ol	runs	mpd.ob	mpd.ra	mpd.ra	mpd.ob	mpd.ob	mpd.ob	runs
site1	14	0,024	0,034	0,011	189	-0,897	0,189	999	0,117	0,122	0,009	275	-0,517	0,275	999
site2	19	0,012	0,024	0,006	17	-1,946	0,017	999	0,127	0,128	0,006	437	-0,105	0,437	999
site3	10	0,009	0,047	0,015	1	-2,536	0,001	999	0,059	0,117	0,011	1	-5,237	0,001	999
site4	17	0,012	0,027	0,009	32	-1,693	0,032	999	0,084	0,123	0,008	1	-4,682	0,001	999
site5	26	0,019	0,017	0,002	826	1,029	0,826	999	0,133	0,132	0,002	681	0,529	0,681	999
site6	26	0,019	0,017	0,002	722	0,663	0,722	999	0,134	0,132	0,002	805	0,911	0,805	999

- Significant overclustering for sites 2, 3, 4

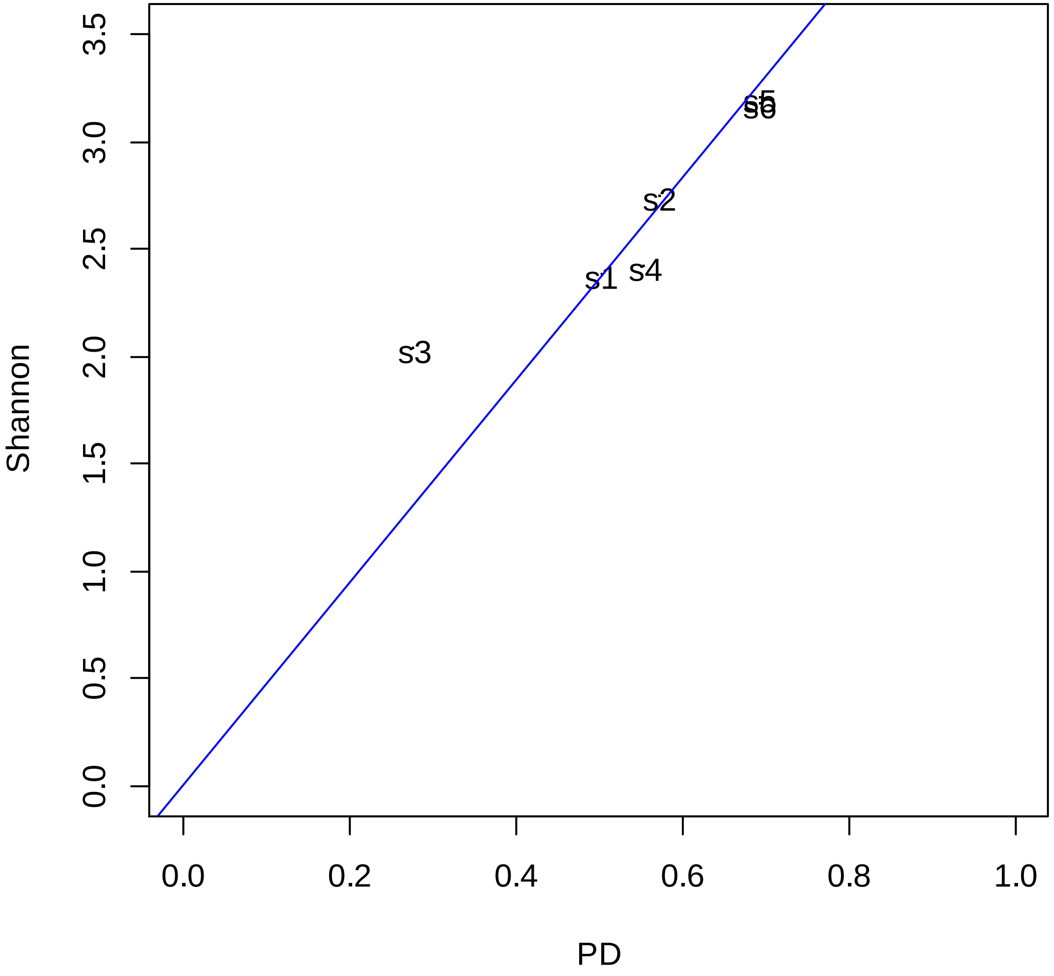
Phylogenetic diversity



```
####################################################
###  CALCULATION of Phylogenetic diversity (PD) #####
####################################################

```

- Comparison of PD and Shannon diversity



Acknowledgement

This project has received funding from European Union's Horizon 2020 research and innovation programme under grant agreement No. 101079234

Thank you for your attention!

www.biolaweb.com