Lab04

Introduction

Phylogenetic analysis of mitochondrial cytochrome b gene and NADH Dehydrogenase subunit 1(ND1) (partial) sequences in chamois (mountain ungulate in Eurasia).

Document all work during the dry lab for each exercise, defining all your tools and input parameters, data output, and interpretation.

From Lab04, there are three Assignments to be submitted to myCourses:
Discussion 4.1
Discussion 4.2
Activity 4.1

Distance-based phylogenetic trees

Following the paper in https://pubmed.ncbi.nlm.nih.gov/18796461, you are going to build a phylogenetic tree of *cytb* sequences (nucleotides).

- In Perusal: https://app.perusall.com/courses/bioinformatics-algorithms-spring25

- Read Materials and Methods & Results
  - ➢ Check for my comments
  - ➢ Feel free to post questions & comments in Perusal

- Download sequences from GenBank (multi fasta file)
  - ➢ There are 20 sequences (haplotypes)
  - ➢ 2 sequences as outgroups

- Have everything ready:
  - ➢ Files (fasta)
  - ➢ Fasta multiple sequence alignment in phylip format
  - ➢ software: PHYLIP, MEGA, Tree viewer (Lab00)
  - ➢ PHYLIP – **MacOS users**:

    Download http://phylipweb.github.io/phylip/download/phylip-3.695-osx.dmg
    Copy folder in your directory
    Go to /../phylip-3.695/exe
    Copy linkmac from source dir: cp ../src/linkmac .
    chmod +x linkmac
    Execute linkmac: ./linkmac

Steps:

1. Sequence files in fasta format.
   a. Get the sequences from the article.
      Tip: Search by Nucleotide in NCBI "cytochrome b AND rupicapra[orgn]"
      Too many sequences?
      Pro-Tip: Select one of the haplotypes (349 bp) and click on PopSet.

   b. Rename fasta headers with **only nine characters**: Population+haplotype (eg. AsiCytb22, that would be *Rupicapra r. asiatica* haplotype cytb-22).

2. Alignment.

3. Input file in phylip format.

4. Calculate distance matrices.

5. Build tree joining branches and nodes.

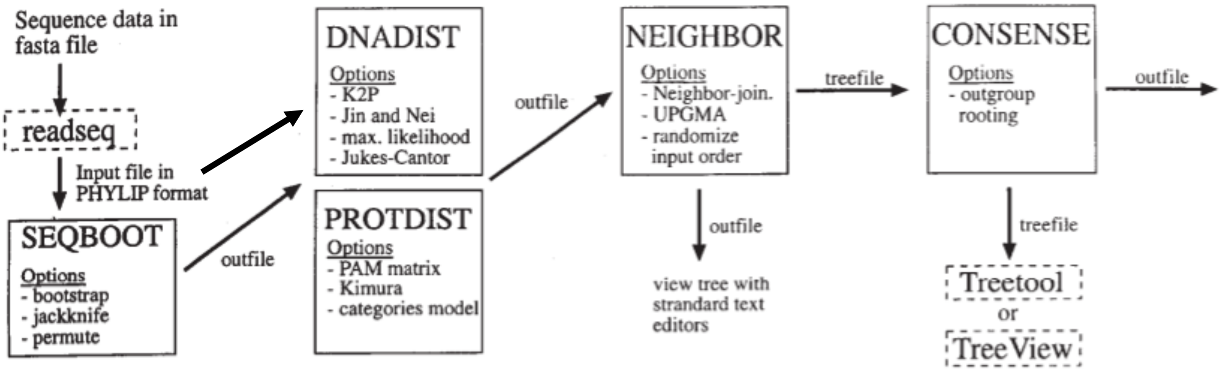
Workflow for a phylogenetic analysis using the PHYLIP program package
https://phylipweb.github.io/phylip/

An excellent guide to using PHYLIP with molecular data is available:
https://phylipweb.github.io/phylip/tuimala3.pdf

Another good PHYLIP practical guide:
https://vcru.wisc.edu/simonlab/bioinformatics/programs/phylip/doc/main.html

For today's practice, we are not implementing the bootstrap process or building consensus trees (next week).

- ~~SEQBOOT~~ accepts the PHYLIP file as input and sequences are bootstrapped a user-defined number of times.
- **PROTDIST/DNADIST** calculates distance matrices
- **NEIGHBOR** program joins nodes and branches according to the calculated values. *NEIGHBOR uses a fontfile file to produce an output tree as text. You can use any fonts (font1, font2, ..., font6) available in PHYLIP. But you can rename it to "fontfile".
- **TREEVIEW** or **TreeTool** allow the user to manipulate the tree (e.g. re-rooting, making branch rearrangements, and changing fonts).
- ~~CONSENSE~~ program reduces the 1000 trees to the one that includes only those nodes that are present in the majority of the trees in the set of all possible trees and indicates the bootstrap values by the nodes.



Source: *Bioinformatics*, Baxevanis.

You can read the documentation for each program in /phylip-3.695/doc/ folder.

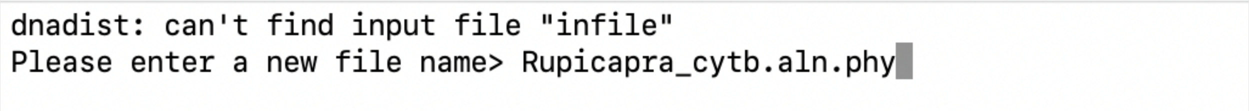
In **Oedipus**

Go to your folder with the alignment files.
PHYLIP is installed in
`/mnt/sde_dir/software/phylip/phylip-3.697/`

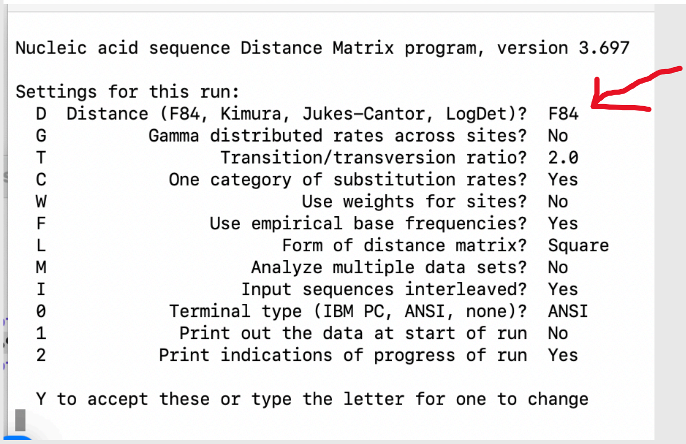
With exec files in
`/mnt/sde_dir/software/phylip/phylip-3.697/exe/`

DNADIST:
`/mnt/sde_dir/software/phylip/phylip-3.697/exe/dnadist`

*BETTER COPY THE NAME OF YOUR FILE BEFORE.



```
dnadist: can't find input file "infile"
Please enter a new file name> Rupicapra_cytb.aln.phy
```



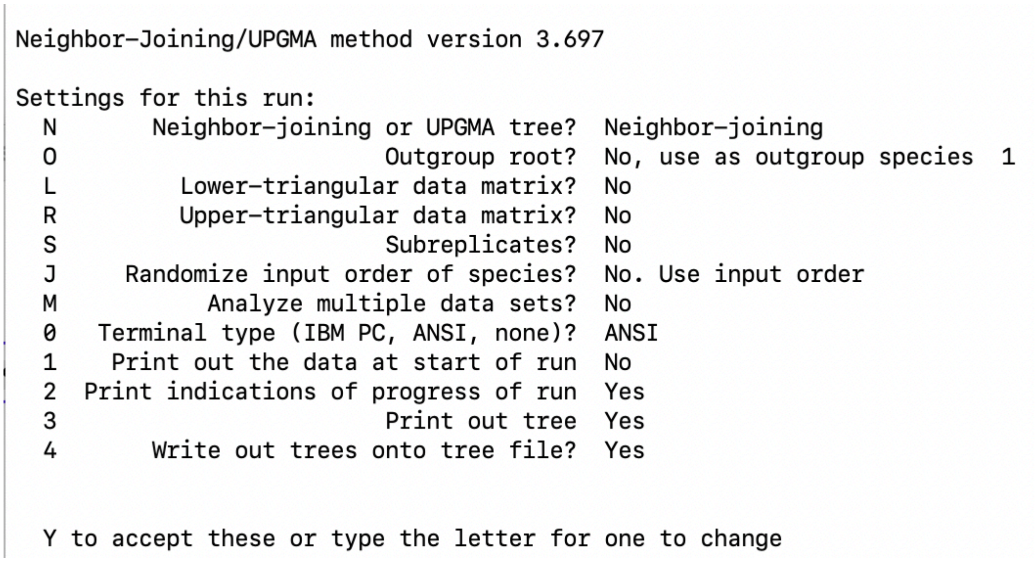
```
Nucleic acid sequence Distance Matrix program, version 3.697

Settings for this run:
  D  Distance (F84, Kimura, Jukes-Cantor, LogDet)?  F84
  G          Gamma distributed rates across sites?  No
  T                 Transition/transversion ratio?  2.0
  C            One category of substitution rates?  Yes
  W                         Use weights for sites?  No
  F                Use empirical base frequencies?  Yes
  L                       Form of distance matrix?  Square
  M                    Analyze multiple data sets?  No
  I                   Input sequences interleaved?  Yes
  0            Terminal type (IBM PC, ANSI, none)?  ANSI
  1             Print out the data at start of run  No
  2           Print indications of progress of run  Yes

  Y to accept these or type the letter for one to change
```

Feel free to change the Distance method with “D” to one of the models discussed in class (Jukes-Cantor or Kimura-2P). Once you modify the parameter, type “Y” and “Return”. You’ll have a new file “output”, which I recommend renaming (mv output Rup_cytb_JC-distance.txt). That will be the input for NEIGHBOR.

After DNADIST, look at the output matrices generated. Use Jukes-Cantor and Kimura 2-parameter. For K2P, what is the Transition/transversion ratio? What ratio would you use?

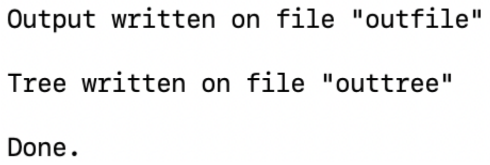
Build trees using different combinations: distance-based methods (Jukes-Cantor, Kimura), and tree construction methods (UPGMA, Neighbor Joining).
Tip: rename every *outfile* with the options used. That way, your final tree - *treefile* (newick format) and *outfile* (text format) would be something like Cytb_K2P_UPGMA* or Cytb_JC_NJ*.

```
/mnt/sde_dir/software/phylip/phylip-3.697/exe/neighbor
```



```
Neighbor-Joining/UPGMA method version 3.697

Settings for this run:
  N       Neighbor-joining or UPGMA tree?  Neighbor-joining
  O                        Outgroup root?  No, use as outgroup species  1
  L         Lower-triangular data matrix?  No
  R         Upper-triangular data matrix?  No
  S                        Subreplicates?  No
  J     Randomize input order of species?  No. Use input order
  M           Analyze multiple data sets?  No
  0   Terminal type (IBM PC, ANSI, none)?  ANSI
  1    Print out the data at start of run  No
  2  Print indications of progress of run  Yes
  3                        Print out tree  Yes
  4       Write out trees onto tree file?  Yes


  Y to accept these or type the letter for one to change
```

You’ll have two outputs:



```
Output written on file "outfile"

Tree written on file "outtree"

Done.
```

Explore their formats. Do you recognize one from Lab01? Which one would you need for the tree viewer program?

After NEIGHBOR, once the tree file is renamed, open it in a tree viewer program (MEGA, Figtree). Check the topology and how the different haplotypes and populations are

clustered together. Is your topology like the phylogenetic tree figures in the article (Figure 2)?
Tip: work around with the tree viewer program. I like the FigTree program more (I am more used to it) for tree configuration:

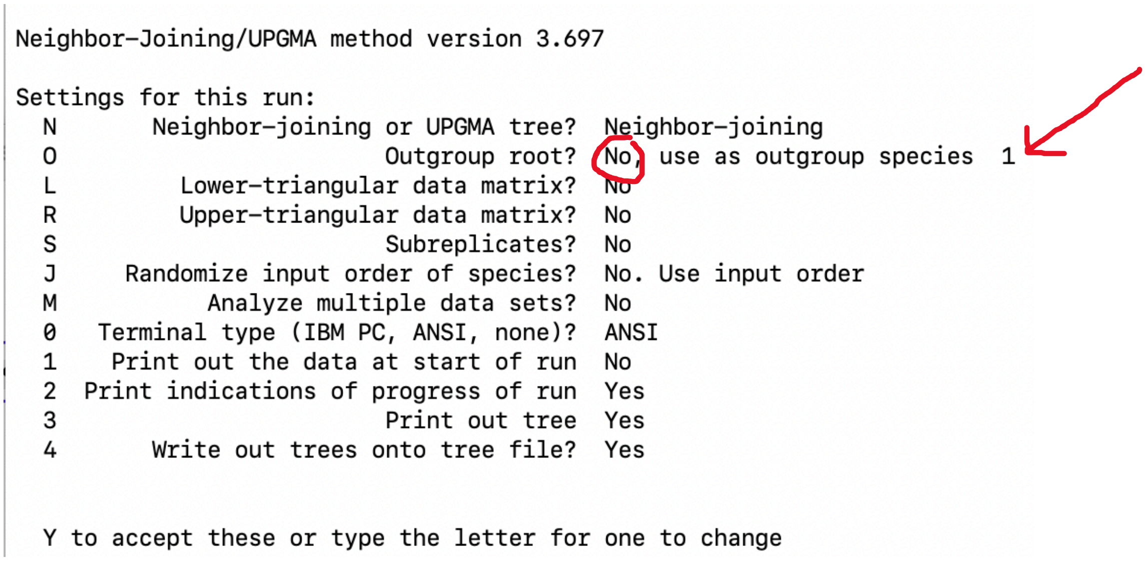
- Check the different layouts (rectangular, polar, radial).
- Node labels. See what node labels you have…. what are those numbers?
- Scale bar. Choose a different scale range.
- Make **good-looking trees** (check the zoom) that can be saved in a pdf file.

Discussion 4.1
Try reconstructing two trees, one rectangular (like Figure 3 in Rodriguez *et al.* 2009) and one radial (like Figure 2) that look as much as the trees in the article. Write down the steps and parameters (methods) you used. (Once you finish, submit the report with explanations to myCourses (in Assignments).

Distance-based phylogenetic tree with outgroup (rooted tree)

Now, let's build a rooted tree. We need an outgroup—a sequence or organism that falls out of the clade of study (not too far) but shares a common ancestor. The domestic goat (*Capra hircus*) is a suitable candidate for chamois.

Use the fasta file in myCourses content/Labs: *Lab04-Rupicapra_ND1_Capra.fasta,* which contains NADH Dehydrogenase subunit 1(ND1) sequences, and follow the previous steps (create the multifasta file, rename fasta headers, multiple sequence alignment, save alignment as phylip format). When generating a tree with NEIGHBOR, the Neighbor-Joining (not UPGMA) method must be used to associate a sequence as the outgroup under "Outgroup root?". Open the tree in the tree viewer and modify it accordingly.
Pro-tip: you have to specify which sequence is the outgroup, determined by its position in the alignment file (1, 2, 3….., 21). Select "outgroup root?" to Yes, and the number.



```
Neighbor-Joining/UPGMA method version 3.697

Settings for this run:
  N       Neighbor-joining or UPGMA tree?  Neighbor-joining
  O                        Outgroup root?  No, use as outgroup species  1
  L         Lower-triangular data matrix?  No
  R         Upper-triangular data matrix?  No
  S                        Subreplicates?  No
  J     Randomize input order of species?  No. Use input order
  M           Analyze multiple data sets?  No
  0   Terminal type (IBM PC, ANSI, none)?  ANSI
  1    Print out the data at start of run  No
  2  Print indications of progress of run  Yes
  3                        Print out tree  Yes
  4       Write out trees onto tree file?  Yes


  Y to accept these or type the letter for one to change
```

Discussion 4.2
Try to reconstruct one ND1 tree with Capra as outgroup. Submit your final tree in pdf.

For the Activity assignment, retake the initial fasta file with *cytb* haplotypes, and add an outgroup. A good outgroup sequence for *cytb* would be GenBank EF158834. Why? Pick the same model as in Figure 3 (Kimura and NJ) and repeat the same steps to build a rooted tree.
Export (pdf) the final tree and make sure it is as nice-looking as the tree in Figure 3 ;)

Activity 4.1
Once you finish, submit the report to myCourses (in Assignments).