# Phylogenetic analysis practical

Objective of the practical:

* collecting data from sequence databases
* aligning the data
* conducting phylogenetic analysis using Phylip

## 1 Data collection

The initial dataset on webpage is a textfile in fasta-format from the gene brain-derived neurotrophic factor (BDNF) from 12 vertebrate animals. BDNF acts on certain neurons of the central nervous system and the peripheral nervous system, helping to support the survival of existing neurons, and encourage the growth and differentiation of new neurons and synapses. In the brain, it is active in the hippocampus, cortex, and basal forebrain—areas vital to learning, memory, and higher thinking. BDNF itself is important for long-term memory.

The initial dataset includes sequences from one bird (Gallus, chicken) and 11 mammals (two primates: human and chimpanzee, three Artodactyla: pig, cow, horse, two rodents: mouse and rat, the rest being Carnivora). Birds (Aves) and mammals are two “sister-groups” in animal kingdom.

**Expand the dataset by collecting at least 10 additional animals.**

Some suggestions which contribute for making the data a bit more presentable throughout vertebrates and also highlight differences between animal “groups”.

* Take more birds.
* Take also frogs (Amphibia)
* Take more primates (i.e. relatives of human and chimp)

### Finding similar sequences:







Create a new file containing 10 new sequences and the sequences from initial dataset. Be careful with naming the sequences - later used Phylip format uses only first 10 characters and they should be unique for each sequence.

## 2 Create multiple alignment

For that purpose we will be using ClustalX .

Start the program and open the file with sequences.



Select the Phylip format as the output format (Alignment menu>Output Format Options)

Create the multiple alignment (Alignment menu>Do Complete Alignment). Additional parameters for both pairwise and multiple alignments are available under Alignment>Alignment Parameters menu.

An alignment given by a program is always just a suggestion and must be inspected manually = by own eyes and brains. Depending on the case, corrections are needed / not needed.

## 3 Phylip

## General usage of the program

The programs are controlled through a menu, which asks the users which options they want

to set, and allows them to start the computation.

Most of the programs look for the data in a file called infile. If they do not find this file

they then ask the user to type in the file name of the data file. Output is written onto special

files with names like outfile and outtree. Trees written onto outtree are in the

Newick format - you will be using TreeView for viewing the trees.

## Running PHYLIP programs

The programs are used in a sequential way. The output from the first program is used as an input in the next program. The trick is to know how to use the programs in suitable combinations. In Windows, the PHYLIP programs can be invoked by double-clicking on the icon or by typing the name of the program on the command line. It is advisable to use programs from the command line, because then you will be better able to see, e.g., the error messages that might appear.

Most PHYLIP programs run in the same way. The input for a program is taken from a file called infile - if the program does not find this file it then asks the user to type in the file name of the data file. The results are written in a file called outfile. Some programs may write both outfile and a file called outtree or plotfile. **Because most of the programs use the default names for the input and output files, you need to be sure to rename the files you want to save before proceeding to further analysis.**

Otherwise you risk losing your results. For example, you get a distance matrix (outfile) from the program Dnadist, but you want to try different settings for the matrix calculations. Then, before doing the matrix calculation again, rename outfile to Dnadist\_out\_F84 or something similar depending on the type of the analysis and methods used), so that you can tell different analysis results apart after you have ceased to work.

## Essential programs

Here is a list of the programs that can be used for the molecular sequence data analysis. The

programs are divided into the method categories. The choice of the correct analysis method is

left for the user.

### Distance methods

These programs are intended to be used sequentially.

First a distance matrix is calculated by Dnadist or Protdist program from the multiple sequence alignment. The matrix is then transformed into a tree by Fitch, Kitsch or Neighbor program. Programs Dnadist and Protdist create a file outfile. Before running Fitch, Kitsch ot Neighbor, outfile should be renamed, either as infile or with another file name. Fitch, Kitsch and Neighbor programs create both outfile and outtree.

*Dnadist* - DNA distance matrix calculation

*Protdist* - Protein distance matrix calculation

*Fitch* - Fitch-Margoliash tree drawing method without molecular clock

*Kitsch* - Fitch-Margoliash tree drawing method with molecular clock

*Neighbor* - Neighbor-Joining and UPGMA tree drawing method

### Character based methods

These programs read in the sequence alignment, and produce either one or multiple trees in the output files outfile and outtree.

*Dnapars* - DNA parsimony

*Dnapenny* - DNA parsimony using branch-and-bound

*Dnaml* - DNA maximum likelihood without molecular clock

*Dnamlk* - DNA maximum likelihood with molecular clock

*Protpars* - Protein parsimony

*Proml* - Protein maximum likelihood

### Resampling tool

This program reads in a sequence alignment, and generates a specified number of random samples into a file outfile. These random samples are usually used in subsequent analysis as a sequence alignment file with the option M (“use multiple datasets”) turned on.

*Seqboot* - Generates random samples by bootstrapping or jack-knifing

### Consensus trees

This program constructs a consensus tree from multiple trees. For example, Dnapars can produce multiple trees, which can be summarized by the program Consense. Also the results of bootstrapping are summarized by the program Consense as a majority rule tree.

*Consense* - Draws consensus trees from multiple trees

For this practical we will be using DNA sequences for phlyogeny reconstruction and you will create trees with and without bootstraping. For viewing trees use TreeView.

Workflow without bootstraping:

neighbor

tree

dnadist

alignment

Workflow with bootstraping:

alignment

seqboot

dnadist

neighbor

consense

tree

## Quick start

Align your DNA sequences and save the alignment in PHYLIP-format as *alignment.phy*. Start the program Dnadist by typing Dnadist to the command prompt or double clicking on the program’s icon.

First Dnadist (and all the other programs also) checks whether there is a file infile in the folder you started the program in. If it does not find infile it asks you to type in the name of the sequence alignment file.

Dnadist: can't find input file "infile"

Please enter a new file name> alignment.phy

Note that the programs are easiest to use if both the programs and the datafiles are in the same folder as in the example above. If datafiles are in a different folder, you can type in the whole path to the file, e.g., if the files were in the folder D:\data you would type

Dnadist: can't find input file "infile"

Please enter a new file name> D:\data\alignment.phy

All PHYLIP programs are menu-driven. Below is the menu written by Dnadist. Every line in the menu starts with a capital letter or number. You can change the settings of the program by typing in the letter or the number in front of the option you would like to change. For example, typing “d” and pressing Enter, would cycle through different evolutionary models implemented in Dnadist. After you are satisfied with the settings (for this quick start, do not change any options), you should type in “y” and press Enter. This starts the run.



Dnadist prints indications of the run (below). After it has finished calculating all the pairwise distances between the sequences, it tells you so (Done.). These pairwise distances are saved in a file outfile. The file contains just plain text, and you may want rename the file as outfile.txt so that it opens automatically in Notepad when you double-click it.



Next rename outfile as infile, and run the program Neighbor (type in Neighbor). The next menu should appear. Now Neighbor has read the pairwise distances from the file infile, and does not ask you for a new filename. You can again modify the settings to your liking, but for this quick start just type in y and press Enter.



Like Dnadist, Neighbor also prints out indications of the run. After completing the analysis, the program tells you so (Done.).



The tree is now contained in the files outfile and outtree. You can view the graphical tree in outfile by opening it in some text editor or by using Treeview program for graphical representation.

**When doing bootstrap trees make sure you use option to analyse multiple data sets!!!**