# Workshop Practical on concatenation and model testing

Jacob L. Steenwyk & Antonis Rokas

*Programs that you will use: Bash, Python, Perl, Phyutility, PartitionFinder, awk*

To infer a putative species phylogeny using the concatenation method, a concatenated supermatrix is used as input. In this workshop practical, we will perform the common steps of creating a concatenated supermatrix, a partition file, and a determining the appropriate models of evolution for partitions.

Note, steps within objectives that have fill-in-the-blank prompts are indicated as such using blue color font. Please fill in these prompts.
Additionally, if the software you are trying to use isn't in your path, it is likely in *~/software*

**Protocol**

**1) Download and examine the dataset**

• The dataset for this practical will amino acid FASTA files

Objectives:

i) Go to your Guacamole terminal interface and download the data set by typing *wget,* pasting the link address available from the website, and pressing *enter*.
On your keyboard, press *enter* or *return* to catalyze the download

ii) Confirm that you have downloaded the tar zipped directory *FILES_concat_and_model_testing_practical.tar.gz*

iii) Unzip the directory using the following command:

*tar -zxvf FILES_concat_and_model_testing_practical.tar.gz*

iv) Now change directory into the newly unzipped directory using the following command:

*cd FILES_concat_and_model_testing_practical*

v) Examine the contents of the directory using the *ls* command

vi) How many directories are in *FILES_concat_and_model_testing_practical*?

**2) Concatenate the amino acid FASTA files with *phyUtility***

• Inferring putative species tree using the concatenation method requires constructing a supermatrix of all your genes. In this step, we will familiarize ourselves with the dataset and create a supermatrix using *phyUtility*

• *phyUtility* is a command-line phyloinformatics tools to facilitate phylogenomic and related analyses (publication link: https://www.ncbi.nlm.nih.gov/pubmed/18227120 ; github: https://github.com/blackrim/phyutility)

Objectives:

i) Change directory to *FILES_aligned_and_trimmed_fastas* using the *cd* command

ii) Examine the contents of the directory using the *ls* command

iii) How many FASTA files are there?
(hint: they are aligned and trimmed using *mafft* and *trimal* and have *".fa.mafft.trimal"* as suffixes)

iv) All FASTA files have the same set of taxa. How many taxa are in each FASTA file?
(hint: use *"grep"* and *"wc -l"*)

v) Use *phyUtility* to create a concatenated supermatrix using the following command:

*phyutility -concat -in EOG092D* -out concatenated_supermatrix.nexus*

where *-concat* specifies that you want to use the concatenation functionality, *-in* is how to specify the FASTA files you want to concatenate, and *-out* species the name of the output file

vi) Examine the output file *concatenated_supermatrix.nexus* using the *more* or *head* command

vii) Create a partition file using the following command:

*head -n 3 concatenated_supermatrix.nexus | tail -n 1 | sed 's/^.*\[//g' | sed 's/\s\].*$//g' | sed 's/.fa.mafft.trimal_gene.\s/=/g' | tr " " "\n" | sed 's/^/AUTO,\s/g' > partitions.txt*

where

- *head -n 3 concatenated_supermatrix.nexus* shows you the first three lines of *concatenated_supermatrix.nexus*,
- *tail -n 1* will print just the last line of the stdout,
- *sed 's/^.*\[//g'* removes everything from the beginning of the line till the open bracket,
- *sed 's/\s\].*$//g'* removes everything from the close bracket to the end of the line
- *sed 's/.fa.mafft.trimal_gene.\s/=/g'* removes removes the end of the file name and replaces it with an equals sign,
- *tr " " "\n"* replaces spaces with new lines,
- *sed 's/^/AUTO,\s/g'* replaces the beginning of the line with *AUTO,* and
- the stdout is redirected to *partitions.txt*

viii) Examine *partitions.txt*, how long is the supermatrix?

**3) Concatenate the amino acid FASTA files with a custom script**

• Alternatively, a custom script can be written to concatenate FASTA files. Concatenating FASTA files programmatically is great practice as the task is of mild difficulty (like a 2/5 on a difficulty scale).

• Here, I will be sharing with you a custom script I wrote to facilitate my projects.

Objectives:

i) Read the help message of the script that will be used to create the concatenation matrix by executing the following command:

*python ../FILES_scripts/create_concat_matrix.py -h*

ii) What are the four required parameters?

iii) To create the <*list of alignment files*> file, execute the following command:

*ls EOG092D* > alignment.list*

iv) To create the <*list of taxa names*> file, execute the following command:

*cat EOG092D1MLK.fa.mafft.trimal | grep ">" | sed 's/\s.*$//g' | sed 's/>//g' > taxa_names.list*

Explain what each step in the pipe command does. Hint: if you are having difficulty determining what something does, execute the command in pieces and compare the stdouts. For example, if I am unsure of what *sed 's/\s.*$//g'* does, I would first execute the following command:

*cat EOG092D1MLK.fa.mafft.trimal | grep ">"*

and then I would execute the following command:

*cat EOG092D1MLK.fa.mafft.trimal | grep ">" | sed 's/\s.*$//g'*

I would then compare the stdouts of both commands to understand what *sed 's/\s.*$//g'* does.

Explanation of *cat EOG092D1MLK.fa.mafft.trimal:*

Explanation of *grep ">"*:

Explanation of *sed 's/\s.*$//g'*:

Explanation of *sed 's/>//g'*:

Explanation of > *taxa_names.list:*

v) Execute the *create_concat_matrix.py* script using the following command:

*python ../FILES_scripts/create_concat_matrix.py -a alignment.list -c prot -t taxa_names.list -p concat*

vi) Examine the output files especially *concat.partition* and ensure consistency between this one and the previously made partition file, *partitions.txt*

vii) Certain programs require nexus formatted partition files for increased other analyses that have increased complexity. To prepare you for creating nexus files, modify *concat.partition* to be in nexus format using the following command:

*bash ../FILES_scripts/create_nexus_partition.sh concat.partition > concat.partition.nexus*

viii) Examine the contents of the resulting file

**3) Determine models of sequence evolution using *PartitionFinder***

• Different genes may have different models of sequence evolution that best describes their evolution. As an example using models for nucleotide sequences, the simplest model is JC69 (Jukes and Cantor, 1969; https://www.sciencedirect.com/science/article/pii/B9781483232119500097?via%3Dihub) where assumptions of equal base frequencies and mutation rates is assumed for A, T, C, and G while the GTR (generalized time-reversible; http://www.damtp.cam.ac.uk/user/st321/CV_&_Publications_files/STpapers-pdf/T86.pdf) model is the most general and may require up to six substitution rate parameters and four equilibrium

base frequency parameters. As you may imagine, the evolution of some sequences may be best described by GTR as opposed to JC69 (or vice versa).

• There are numerous methods to determine the best fit models for sequences. In fact, many phylogenetic software (e.g., *RAxML* and *IQ-tree*) have this built into existing frameworks, which is why *AUTO* was specified in the model section for 'auto selection'. However, one may want to determine the best fit model separately for various reasons (e.g., to split up the task up into smaller chunks)

• In this step, we will use *PartitionFinder* (https://www.ncbi.nlm.nih.gov/pubmed/28013191) to determine the best fit model in our data

Objectives:

i) Change directory to a subdirectory of the parent directory using *../FILES_PartitionFinder* and the *cd* command.

ii) Examine the contents of the directory using the *ls* command

iii) Move the *concat.fa* file made during step 2 to the current directory using the *mv* command

iv) *PartitionFinder* requires that the input sequence file be in phylip format. Convert the fasta file to phylip format using the following command, which uses the script *Fasta2Phylip.pl* and was originally written by Joseph Hughes (https://github.com/josephhughes):

*perl ../FILES_scripts/Fasta2Phylip.pl concat.fa*

Examine the resulting file named *concat.fa.phy*.

v) *PartitionFinder* requires its own configuration file (or input file). Examine the configuration file *partition_finder.cfg*

vi) In *partition_finder.cfg*, the sections *"Alignment File"* and *"Models of Evolution"* have been left blank but all other sections have been filled for you. Populate these sections with the newly generated phylip formatted alignment file and LG, JTT, and BLOSUM62 such that your models line looks like *"models = LG, JTT, BLOSUM62;"*

To do so, the easiest text editor to use in *nano*

vii) Next, execute the PartitionFinder program by using the following command:

*PartitionFinderProtein.py* .

viii) Examine the *best_scheme.txt* file in directory *analysis*

ix) What model(s) of sequence evolution best fit the data?

x) How many partitions were made?

xi) PartitionFinder has prepared partition files for input into three explicitly stated programs. What are these programs?