- JSMC practical course script -

Inferring phylogeny based on sequence information

Collection of coding sequences via BLAST searches

Search for the coding sequences of the following genes from *Arabidopsis thaliana* based on their NCBI accession number and save the nucleotide sequence in FASTA file format. Ensure that you only download the coding sequences (CDS) and not the complete mRNA sequence.

NCBI nucleotide database: http://www.ncbi.nlm.nih.gov/nuccore

NCBI accession numbers:

Gene name	**Gene abbreviation**	**Accession number**
APETALA1	*AP1*	NM_105581
APETALA3	*AP3*	D21125
PISTILLATA	*PI*	NM_122031
AGAMOUS	*AG*	NM_118013
SEPALLATA1	*SEP1*	NM_001125758
SEPALLATA2	*SEP2*	NM_111098
SEPALLATA3	*SEP3*	NM_001198152
SEPALLATA4	*SEP4*	NM_201682

Use the downloaded coding sequences of *AP3*, *PI*, *AG* and *SEP3* to search for orthologous genes in the early diverging angiosperm species *Amborella trichopoda* and *Nuphar advena*; the magnoliid *Liriodendron tulipifera*, and the gymnosperm species *Gnetum gnemon* and *Picea abies*. Perform four individual BLAST searches (one for each gene of interest). Use the coding sequence of the respective gene as query and limit the search set to the five mentioned species. Set the program selection to ‘discontiguous megablast’ to enable the detection of more dissimilar sequences.

NCBI nucleotide BLAST search:
https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch

The following orthologous genes should be detected:

	***Arabidopsis thaliana* query gene**			
	AP3	***PI***	***AG***	***SEP3***
Amborella trichopoda	*AmAP3*	*AmPI*	*AG*	*AGL9*
Nuphar advena	*AP3.2*	*PI1*	*AG*	-
Liriodendron tulipifera	*AP3*	*PI*	-	*AGL9*
Gnetum gnemon	*GGM2*		*GGM3*	-
Picea abies	*DAL11 + DAL12*		*DAL2*	-

Save the coding sequences of the 15 depicted genes in FASTA file format (again ensure that you download the coding sequences only).

For the phylogeny reconstruction and correct rooting of the resulting phylogenetic tree, it is necessary to also include a suitable outgroup sequence i.e. a distantly related gene that is basal to all other examined genes. Thus additionally download the coding sequence of the MADS-box gene *CgMADS1* (NCBI accession AB035567) from *Chara globularis*, which is one of the closest relatives of all extant land plants.

The final sequence collection should contain 24 coding sequences.

Verify that all sequences within your collection are complete coding sequences (i.e. check whether they start with a start codon and end with a stop codon).

Uniform the names of all sequences according to the following style:

Genus name (underscore) species name (underscore) gene name (blank space) NCBI identifier

(e.g. Arabidopsis_thaliana_SEP3 NM_001198152)

It is important to use ‘underscore’ instead of ‘blank space’ to separate genus, species and gene names, respectively.

Multiple sequence alignment and phylogeny reconstruction

The phylogeny reconstruction will be performed based on a codon alignment of the collected coding sequences. To create a codon alignment, first of all, the collected coding sequences are translated into the amino acid sequences of the encoded proteins. There are numerous online tools available to translate nucleotide sequences into amino acid sequences, two commonly used tool are 'Transeq' (http://www.ebi.ac.uk/Tools/st/emboss_transeq/) and 'ExPASy translate' (http://web.expasy.org/translate/). ExPASy translate is suitable to detect the correct reading frame of the input sequence as it simultaneously translates the entered sequence into all six potential reading frames. Transeq is especially comfortable for the translation of numerous sequences as it simultaneously translates multiple sequences. Familiarize with both online tools and translate all coding sequences of your sequence collection into the corresponding amino acid sequences.

Ensure that all coding sequences have been translated correctly by loading the amino acid sequences into Jalview (http://www.jalview.org/). Check whether the sequences have the expected length and search for multiple stop codons per sequence.

If all sequences have been translated correctly create a multiple sequence alignment using MAFFT (implemented in Jalview, web service > alignment > run MAFFT with preset). MAFFT allows for different presets that can considerably influence the resulting alignment.

Inform yourself about the differences among the presets with help of the MAFFT web page (http://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html) and try different presets to compare the alignment quality. Choose which preset creates the best alignment (i.e. compact alignments with few gaps) and save the alignment in FASTA file format.

Use the online tool RevTrans (http://www.cbs.dtu.dk/services/RevTrans/) to 'reverse translate' the amino acid alignment into a codon alignment. Upload or 'copy and paste' the alignment file as well as the coding sequence collection into the corresponding input windows. Set the output format to FASTA file format (advanced options > output format > FASTA) and start the

translation (submit query button). Check for error reports and ensure that all sequences have been translated.

Load the codon alignment into Jalview and make a note of the alignment parts that should not be considered for the tree calculation due to their low conservation (use the nucleotide color mode for better visualization). Save the codon alignment in FASTA file format.

To use the codon alignment for the tree calculation with MrBayes its file format needs to be converted to NEXUS file format. Use the online tool ALTER (https://sing.ei.uvigo.es/ALTER/) to change the file format. Set the input format to MAFFT - FASTA and paste the codon alignment into the input window. Set the output format to MrBayes - NEXUS and convert the file.

The phylogeny reconstruction with MrBayes will be performed via the online tool CIPRES (https://www.phylo.org/portal2/login!input.action). Upload the codon alignment in NEXUS file format, select the input file and the calculation tool (MrBayes) and adjust the following parameter:

Maximum Hours to Run (click here for help setting this correctly)	**2**
My Data Type Is (only one data type can be used through the web form, see help below)	**Nucleic acid**
Specify (only) one outgroup	**(Sequence number of the charophyte sequence)**
Set the number of substitution types (Nst=)	**6**
Set the nucleotide substitution model (Nucmodel=)	**4 x 4**
Exclude these characters from the analysis	**(Previously selected alignment parts)**
Number of Generations (Ngen=)	**5.000.000**
Save branch length information?	**Yes**
Type of consensus tree	**All compatible groups**
Show Tree Probabilities	**Yes**

Save the parameters and start the tree calculation.

After the tree calculation is completed display the output data via the ‘output’ button.

First of all open the error report file and check for any error messages.

Open and skim the main output file and review the following things:

Have all sequences been imported correctly?

Did both tree calculations result in a similar phylogeny? Check whether the split frequency dropped below 0.01 during the calculation and take a look at the ‘generation-probability’ graph. If both calculations resulted in a similar tree the graph displays a scattered cloud rather than two separated curves.

Take a first look at the reconstructed phylogenetic tree at the very end of the main output file.

The consensus tree is located in the ‘infile.nex.con.tree’ file. Download the file and visualize the tree with the FigTree software (http://tree.bio.ed.ac.uk/software/figtree/).