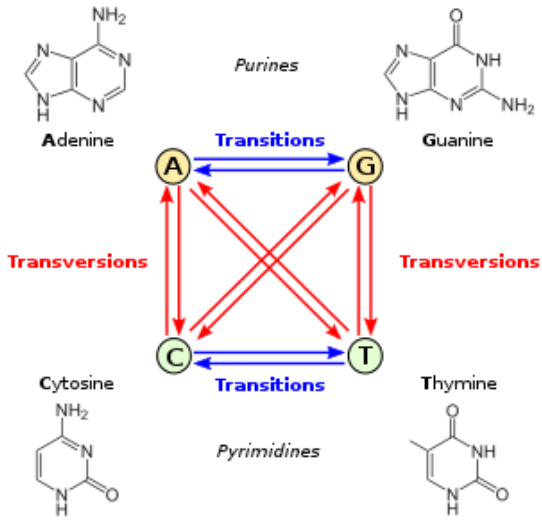
Assignment 9

Your name:
Your email address:

Answer all the questions in red in the provided boxes.

You can skip the parts in green.

Substitution Models



JC69 : All rates equal, all frequencies at 25%

HKY85: All transition rates equal but different from transversions. All frequencies estimated from data

GTR: General Time Reversible: All rates are estimated, and all frequencies are estimated form the data.

Note that the rate for X -> Y = the rate for Y -> X
This is in keeping with considering trees as unrooted, but certainly this is not true in general.

Assignments:

1) Maximum likelihood ratio tests using phyml as implemented in Seaview.
We will test the following models:

	#of rates	#of frequencies	Gamma	Invariant sites	degrees of freedom to previous model
JC	1	1	N	N	
HKY 85	2	4	N	N	4
GTR	6	4	N	N	4
GTR + Gamma	6	4	Y	N	1
GTR + Gamma + Invariant sites	6	4	Y	Y	1

Open the file at
(https://j.p.gogarten.uconn.edu/mcb3421_2021/labs/Lab9/Yeast_vma1_all_aligned.mase)
in Seaview. Select all sequences. Select sites Extein.
Under Trees select phyml.

- Select JC69 from the first pull-down menu. Select none for Invariable sites and none for Among Site Rate Variation. For the tree search use Nearest Neighbor Interchange (NNI). Run.
  Before you click ok when the calculation is done, copy the log likelihood value into the table below. Keep the tree window open.
- Select HKY from the first pull-down menu. Select none for Invariable sites and none for Among Site Rate Variation. For the Transition / Transversion ratio and the nucleotide frequencies select optimized. For the tree search use Nearest Neighbor Interchange (NNI). Run. Write down the log likelihood value into a table.
- Select GTR from the first pull-down menu. Select none for Invariable sites and none for Among Site Rate Variation. For the nucleotide frequencies select optimized.
- Select GTR from the first pull-down menu. Select none for Invariable sites and optimized for Among Site Rate Variation. For the nucleotide frequencies select optimized.
- Select GTR from the first pull-down menu. Select optimized for Invariable sites and optimized for Among Site Rate Variation. For the nucleotide frequencies select optimized.

	Log(likelihood)	2*deltaLogL (comp. to previous model)	Degrees of freedom	P-value:
JC				
HKY85				
GTR				
GTR + Gamma				
GTR + Gamma + Invariant Sites				

One important condition that has to be fulfilled before one can use a Likelihood Ratio Test (LRT) to compare two models, is that the models should be "nested". This means that the simpler model must be a constrained version of the parameter-rich model. The likelihood ratio test is performed by doubling the difference in log-likelihood scores and comparing this test statistic with the critical value from a chi-squared distribution having degrees of freedom equal to the difference in the number of estimated parameters in the two models. The parameter-rich model will always have a better fit, due to the extra parameters and

will therefore have the highest log-likelihood, so the difference should be a positive number. The **degree of freedom** between each of the models is given in the above table - plus/minus gamma shape parameter is one parameter (even though is approximated by 4 rate categories) and the % invariant sites also counts as a parameter.

Use this online chi-square calculator to determine the significance of the test.

Are all the more complex models a significant improvement over the more simple ones? Enter twice the log likelihood ratio and the P-value with which the simpler hypothesis is rejected.

Doing this using a GUI and copying numbers back and forth is tedious. An older program called modeltest automatically tested for a few dozen models. A more recent program iqtree incorporates testing for the appropriate model.

To get the alignment into a format readable by iqtree,
in seaview:

select sites extein,
select all sequences,
then select File > Save **Selection**
Enter a filename (e.g., Yeast_vma1_extein_aligned.fst), select fasta format,
select File -> save selection as

The software is available via a web interface (e.g. here). However, today, we will use the version as available on xanadu. To run iqtree

Establish an ssh connection to **`xanadu-submit-ext.cam.uchc.edu`**

Open an ssh terminal window (either using PuTTY or Terminal), log into Xanadu:
`ssh mcb3421usrXX@xanadu-submit-ext.cam.uchc.edu`
Enter your password

- login
- **`srun --pty -p mcbstudent --qos=mcbstudent --mem=2G bash`**

Create a directory for lab9, and transfer the aligned sequences for exteins only (as a multiple fasta file created via save selection as in seaview - see above) into that directory.

- **`mkdir lab9`**
- **`cd lab9`**
- load the iqtree module
  `module load iqtree/2.2.2`

Get the sequences:
`curl -O https://j.p.gogarten.uconn.edu/mcb3421_2023/labs/Lab9/exteins_yeast_vma_1.fst`

- execute iqtree
  `iqtree2 -s exteins_yeast_vma_1.fst`
  this will be pretty fast.

In case you want to look at the tree, use filezilla to move the files created by iqtree to your desktop computer. You can open the .treefile in seaview. You can also look at the text file exteins_yeast_vma_1.fst.iqtree which includes an ascii version of the tree

Inspect the logfile using more:

`more *.log` or **`more *.iqtree`**

At the end of listing of the for the lnL for the individual models, is the listing of the best models under the different criteria.

Which models were chosen, and what do the abbreviations mean? (see the iqtree documentation chapter 11.6 Rate heterogeneity and 12.1 on DNA substitution models. For info on the different criteria (AIC, BIC, ... see here)

Exercise 2: Parsimony, distance matrix and ml analyses: sensitivity to LBA and missing data.

PHYLIP is a collection of programs for phylogenetic analyses written by Joe Felsenstein. The programs are freely available (including source code), and can be used on a variety of different operating system. The programs are modular. Different modules exist to create bootstrap samples, calculate distance matrices and calculate trees from the distance matrices (Fitch and Neighbor), calculate consensus trees, etc.. All programs either use files called infile or intree, or alternatively the user needs to provide the file name. We will use the sequences in https://j.p.gogarten.uconn.edu/mcb3421_2023/labs/testseq1b.txt

We will use the program programs as implemented in **seaview**.

(IGNORE THIS PARAGRAPH, if you are in class) This paragraph has comments on phylip that you can safely ignore for now: If you use protpars directly (without the seqview interface), note that phylip by default treats gaps as a 21st character. If you want to treat the gap as missing data, you need to replace the gap symbol with "?"'s. In case you want to use one of the programs on your own, you need to read the excellent manuals that come with the software.

To calculate a phylogenetic tree from the aligned sequences using protein parsimony, open seaview, and drag the file testseq1b.txt into alignment window.

Align the sequences using muscle. (click on align, then align all).

In the trees menu select parsimony. Uncheck "ignore all gap sites" and check "gaps as unknown state". To do a more thorough search for the tree that explains the data with the least number of substitutions, select "randomize seq order" 5 times provides a reasonable number of starting points for the heuristic search.

Four notes:

1. Sulfolobus and Thermococcus are Archaea, Borrelia is a Spirochete (bacterium), Acetabularia is a green algae, Daucus is a flowering plant (carrot), Candida, and Saccharomyces are yeasts, Neurospora is another fungus (not a yeast though), Drosophila is an animal (fruit fly) and Trypanosomes are protists.
2. **the Borrelia sequence actually is an archaeal type ATPase acquired through gene transfer**
3. Most of the sequences in testseq1b.txt (V/A-ATPase catalytic subunits) are quite similar to one another. To test the effect of long branches, I added a homologous (paralogous), but only distantly related sequence to this file (the ATPase involved in flagellar assembly from Salmonella).
4. Remember that some sequences contain inteins -- What are potential problems that might be caused by the intein sequence?

In the seaview tree viewer window select re-root. Then select the node between pro- eand eukaryotic sequences.

How are the fungal sequences resolved? (What does this tell us about parsimony and missing data?)

Where does the paralogous Salmonella sequence go?

How many equally parsimonious trees did you obtain? Compare with your neighbors: did the most parsimonious tree have the same number of steps?

Repeat the parsimony analysis using 100 bootstrap samples (do not ignore positions with gaps, but do not check 5 random starting trees--if you do a bootstrap analysis, repeated heuristic searches for each dataset are not worth the time). In the tree window, re-root the tree and place a checkmark in Br support.

What is the support for the two long branches (flSalmonella and Borrelia) going together? Check with your neighbors, did you get the same result?

Explore the Swap feature in the tree viewer window.

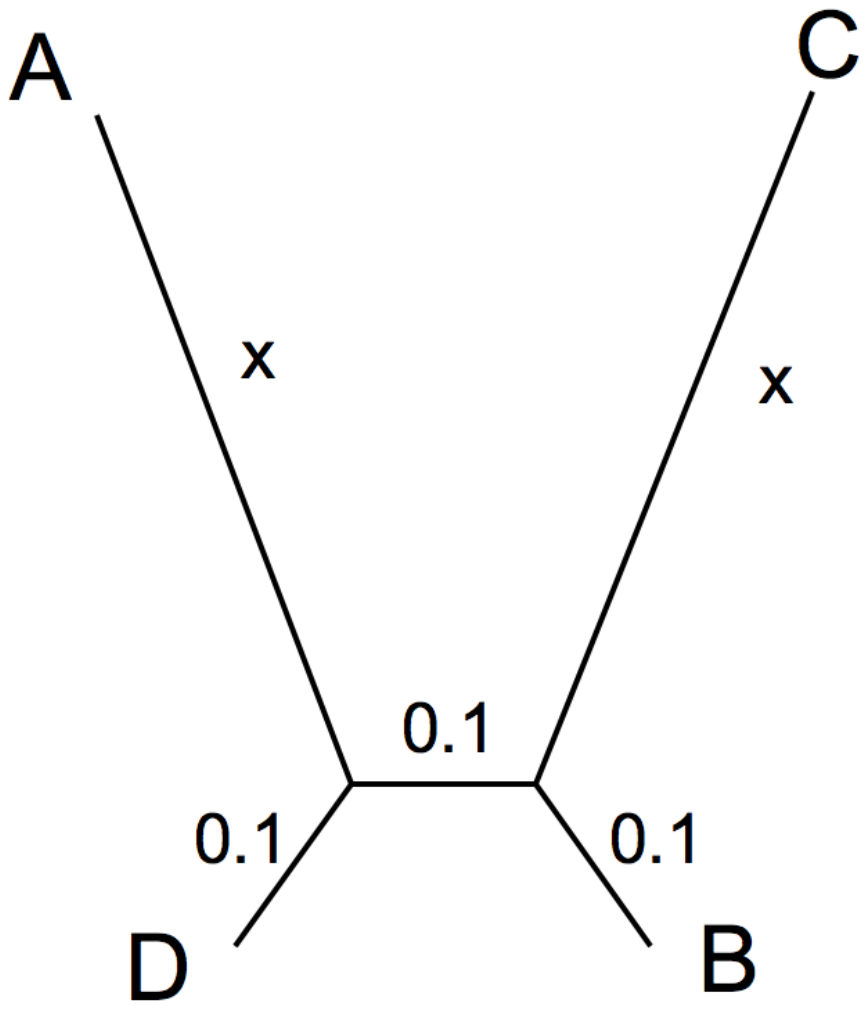
Perform a distance analysis in Seaview. Select bootstrap, Kimura distances, NJ, and do **not** check ignore all gap sites. Inspect the bootstrap support values.
Does the resulting tree indicate that this distance matrix phylogenetic analysis was very sensitive to Long Branch Attraction?

What might have gone wrong with the fungal sequences? (If you need a hint, repeat the analysis with the ignore all gaps option checked).

Exercise 3:

Long Branch Attraction (LBA) is a serious problem in phylogenetic reconstruction. LBA denotes the fact that long branches tend to be grouped together with significant support, even though the organisms representing the long branches did not share more recent common ancestry. The support usually is measured through bootstrap support values for the different trees. We have simulated the evolution of 4 sequences (named A,B,C,D)

according to the following tree:



Files containing these sequences in multiple sequence fasta format were generated and named according to the length chosen for the two long branches (all scaled in substitutions per site). For the simulation we assumed that the Among Site Rate Variation could be described with a gamma distribution that has a shape factor of 1 (equal to an exponential distribution).

These files in a single zipped file are at
https://j.p.gogarten.uconn.edu/mcb3421_2023/labs/simsequences.zip

Your task is to explore the sensitivity of different phylogenetic reconstruction algorithms towards LBA. At the minimum you should use **protein parsimony** and one protein **distance** matrix or **ml** analysis approach. In this case we know that the sequences are aligned as given; however, to explore the effect that the alignment algorithm has on LBA, we can align them before phylogenetic reconstruction. To keep track of things, name the files accordingly.

NOTE I: If you want to explore the effect of alignment, it might be a good idea to use seaview and muscle as alignment program - especially for the more divergent sequences. We will use the GUI provided in seaview.

Note II: You can divide the labor with your neighbor, distributing different sequences to different students.

We will use programs as implemented in SEAVIEW

2A: To test parsimony, choose the files with x = 0.1; 0.3; 1; 3.

For the datasets with x = 0.1, 0.3, 1, 3, use the tree menu in seaview, select parsimony, uncheck "ignore all gap sites", check "gaps as unknown states", check "bootstrap with 100 replicates", and move the consensus tree level lever to the left. (Note: If you are interested in the best parsimony tree, then you want to use the original dataset (not bootstrapped) and randomize the input order for several independent heuristic searches, if you do a bootstrap analysis, repeated heuristic searches for each dataset are not worth the time.)

In the following box list the files that you chose, aligned or as provided, and the bootstrap support for the correct tree ((A,D),(B,C)), or the support for the LBA tree ((A,C),(B,D,)) (note: seaview will show them arbitrarily rooted)

File used	Aligned yes/no	Support for True tree ((AD),(B,C)_	Support for LBA tree ((A,C),(B,D,))

2B) **(or do 2C)** Explore a distance matrix based approach with respect to LBA (Neighbor joining using Poisson corrected or observed distances work well). Depending on the settings, these might be less sensitive to LBA. x = 0.3, 1, 3, 10 are good choices to explore.

In the following box list the parameters you selected in seaview, the files that you chose (aligned or as provided), and for each file indicate the bootstrap support for the correct tree, or the support for the LBA tree:

File used	Aligned yes/no	Parameters used	Support for True tree ((AD),(B,C))	Support for LBA tree ((A,C),(B,D,))

2C) **(or do 2B)** Explore the sensitivity of phyml towards LBA. This may work better on a fast computer

- use the default setting for phyml in seaview (go with nearest neighbor interchange (nni) and the LG substitution matrix).
- the search converges much faster, if you do not align the sequences first
- use the aLRT (approximate likelihood ratio test) support values (not the real non-parametric bootstrap). The aLRT values are between 0 and 1, with one corresponding to maximum probability for the branch to be present in the true-tree.
- x=1, 3, and 10, 30 are good values to explore.

In the following box list give the parameters you chose for phyml, the files that you chose, indicate if you aligned them or used them as provided, and for each file give the support value for the correct tree, or the support for the LBA tree:

File used	Aligned yes/no	Parameters used	Support for True tree ((AD),(B,C))	Support for LBA tree ((A,C),(B,D,))