Additional materials and details for Bacterial Phylogenetics module in Physics 498EBP.
Prof: Seppe Kuehn
TA: Laura Troyer

Videos for understanding background concepts:

(1) The structure of DNA
https://www.youtube.com/watch?v=o_-6JXLYS-k
(2) 16S rRNA background.
https://www.youtube.com/watch?v=wFMcTIEMVJk
https://en.wikipedia.org/wiki/16S_ribosomal_RNA
(3) PCR how does it work and what does it do? Hilarious but effective:
https://www.youtube.com/watch?v=iQsu3Kz9NYo
https://www.youtube.com/watch?v=aUBJtHwHASA
(4) Sanger DNA sequencing: how does it work?
https://www.youtube.com/watch?v=e2G5zx-OJIw
https://www.youtube.com/watch?v=Jnk_4Maf5Fk
(5) DNA sense/anti-sense, the meaning of forward and reverse primers on a gene.
https://www.youtube.com/watch?v=0vdbf4F90_o
(6) DNA electrophoresis
https://www.neb.com/products/n3200-1-kb-plus-dna-ladder#Product%20Information
https://www.youtube.com/watch?v=mN5IvS96wNk
https://en.wikipedia.org/wiki/Chimera_(EST)
(7) Jukes Cantor
https://www.youtube.com/watch?v=a5Dgha7EX7Y
(8) UPGMA method for creating trees.
https://www.youtube.com/watch?v=J-RjA889SKY
https://www.mathworks.com/help/bioinfo/examples/building-a-phylogenetic-tree-for-the-hominidae-species.html
https://en.wikipedia.org/wiki/UPGMA

Original paper by Woese and Fox in PNAS: https://www.pnas.org/content/74/11/5088

Note the very confusing conventions for DNA directionality.
"Forward" primer primes on the 3' end of the anti-sense/template/**non-coding**/(-) strand of the gene. During extension it creates the sense/non-template/**coding**/(+) strand of the gene. The sequence resulting from the reverse primer then makes the **non-coding** sequence. The sequence of a gene is reported as the "coding" sequence which are then read during the process of translation (but of course, not for the 16s rRNA because this is not coding a protein!). So when constructing the sequence you need to take the reverse complement of the sequence from the REVERSE primer!!

Details of the primers you will use for PCR Primers are 20nt long.
27F, 1492R gives a fragment of 1492-27 = 1465 - 40 = 1425nt long. --- but the true length of the sequence you get will vary from strain to strain due to insertions/deletions during evolution.

How do I go from my Sanger data to a sequence of the 16S gene?
Combining forward and reverse reads: (cross platform, free tool for all types of sequence analysis). We have it installed on one machine in the lab, or you can install it yourself locally on your own machine.
http://ugene.net/
Instructions:

(1) Locate your Forward and Reverse read '.ab1' files from the Sanger facility
(2) Open Ugene
(3) Tools>Sanger Analysis>Reads denovo assembly
(4) Click Add and select the file for the **Reverse** read
(5) Click Add again and select the file for the **Forward** read
(6) Click Run
(7) Select "Multiple sequence alignment in…."
(8) Confirm in the left hand panel that the reverse complement of the Reverse read was used.
(9) On the very right hand side of the window select the fourth icon down for export
(10) Select the three dots and choose a path to export the file.
(11) Click export
(12) Choose "Raw Sequence"
(13) Sequence is now exported as a text file.

Notes on alignment with "multi-align" in Matlab.

First, create a structure array which has two fields "Header" and "Sequence" for each of your samples. On this datastructure run the matlab function "multialign"

Second, on the output run the function "showalignment" -- recall our discussion of hypervariable regions above. Can you spot them? Where are the conserved regions?