Lab11

Sanger sequence analysis

The goal of the activity is to analyze and interpret the quality of Sanger sequences and generate a consensus DNA sequence for posterior bioinformatics analyses.

- Understand the Sanger method of sequencing
- Interpret chromatograms
- generate a consensus DNA sequence based on forward and reverse Sanger reactions

A few web links:

- https://www.sigmaaldrich.com/US/en/technical-documents/protocol/genomics/sequencing/sanger-sequencing?srsltid=AfmBOorAOPlfZxqZzR7xVUoZQ-waHQk0QbXD75nkGM5MJlOQaGJ0yAo9
- https://www.thermofisher.com/us/en/home/life-science/sequencing/sanger-sequencing/sanger-sequencing-technology-accessories.html
- https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.youtube.com/watch%3Fv%3DxLgOl0b78OE&ved=2ahUKEwj16trT8dqJAxXkMlkFHZpoOd0QtwJ6BAgMEAI&usg=AOvVaw3LCME9e21FoxqVUfDDl7dh

Document all work during the dry lab for each exercise, defining all your tools and input parameters, data output, and interpretation.

From Lab11, there are two Assignments to be submitted to myCourses:
Discussion 11
Activity 11

Resources:

- DNA analysis software: MEGA or FinchTV (https://digitalworldbiology.com/FinchTV).

*MEGA on my Mac sometimes acts up with .ab1 chromatograms. I wouldn't recommend it for analyzing hundreds of sequences, but for today's activity, it is more than enough. You can alternatively download other DNA analysis software like FinchTV or install a commercial demo (Geneious, CLC Workbench, …).

There is multiple software options are available for DNA sequence analysis.

Analyze a Sanger sequence

ABI sequencer data files (*.ab1), known as trace files, include raw data output from Applied Biosystems' Sequencing Analysis Software. .ab1 files also include quality information about the base calls (quality), the chromatogram (also called the electropherogram), and the DNA sequence.
Quality scores indicate the probability that an individual base is called incorrectly during DNA sequencing. For this lab, we recommend a Q score ≥ 40.

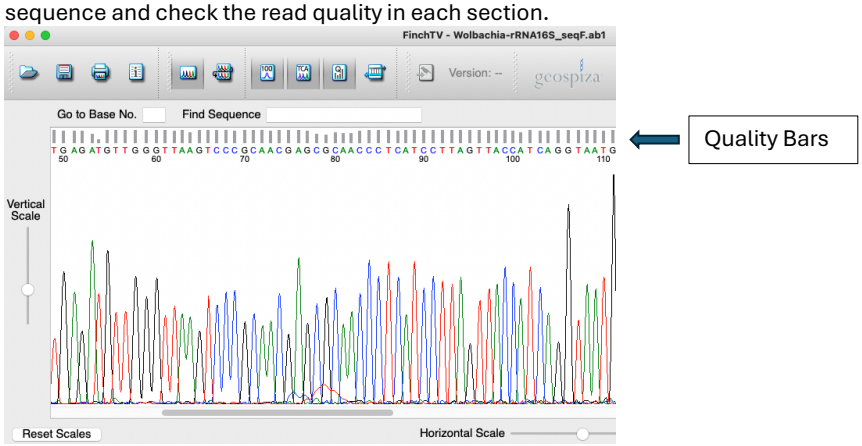
Quality score = -10 x log (probability of error)
For Q20, probability of error = 1 in 100. Base call accuracy 99%.
ForQ40, probability of error = 1 in 10,000. Base call accuracy 99.99%.

In Sanger sequencing of PCR products, DNA is normally sequenced with two reactions: forward (F) and reverse (R). In Sanger sequencing the forward strand uses only the forward primer (the same forward primer used for PCR) while sequencing the reverse strand uses only the reverse primer (the same reverse primer used for PCR).

We will go through the analysis of each sequence separately and then illustrate how to generate a consensus sequence. To generate the consensus, you will align both the forward and reverse sequences to confirm that bases are complementary. If the alignment is not 100% homologous, you **should investigate the discrepancy** in the original chromatogram files.

.ab1 files for Discussion11: (in myCourses > Content > Labs > Lab11-Discussion
Wolbachia-rRNA16S_seqF.ab1
Wolbachia-rRNA16S_seqR.ab1

Open the chromatograms with your preferred DNA analysis program. If allowed, select "show quality values." Quality bars will appear in each residue. Scroll up and down the sequence and check the read quality in each section.



How is the start of the sequence? How is the end? If their quality is no good, usually stated as uncharacterized "N," is it necessary to trim/delete those regions? Let's discuss this section during class.

You can delete/remove the ends, but never trim interior portions. You must use the "N" to call interior regions of poor quality instead. You can export your final sequence to a fasta file. Do the same for the other strand read. Use the Forward (F) or Reverse (R) nomenclature in the name.

Generate the consensus sequence:

Once you have both ends manually curated (trimmed), we can align them and generate a consensus sequence. Some DNA analysis software allows you to do it with one *click* (commercial versions, like Geneious)...we will do it the "hard way".

One option: go to BLAST in NCBI https://blast.ncbi.nlm.nih.gov/Blast.cgi and under **blast nucleotide sequences** select "Align two or more sequences". Paste your forward and reverse fasta files. Use blastn or megablast as the BLAST algorithm.
Hint: the reverse sequence can be use "as is" or convert it in the reverse complementary DNA before doing the alignment.

After checking the alignment tab, the identity should be 100%. If not, you must go back to the chromatograms (trace files) and investigate. Once your alignment is good, save the consensus sequence as fasta: Download > fasta (aligned sequences).

Commented [FRV1]: Blast does not save the consensus seq, it outputs que subjet seq (second seq) alignment. So we are missing the initial (5') section of the consensus.

Analyze not "so nice" Sanger sequences

.ab1 files for Discussion11: (in myCourses > Content > Labs > Lab11-Discussion
Polistes-CO1F.ab1
Polistes-CO1R.ab1

For the low-quality sequencing run, follow the same steps as above. Inspect both forward and reverse and decide which path to take.

What can cause such a lousy quality Sanger sequence?

Discussion 11
Write down the steps and parameters (methods) you used, describing the components and what they represent with your comments. Include significant screen shoots (BLAST alignment, final BLAST result). Once you finish, submit the report with explanations to myCourses (in Assignments).

Sanger sequence of environmental samples

For the activity section (Activity11) you will follow the same steps as before, but we don't know what we are amplifying in the PCR. Sometimes, microbes in the environment cannot be cultured, and no proper identification and genome extraction can be performed. To identify "what's there" we rely on universal primers that can amplify known conserved regions, like mitochondrial genes and ribosomal RNAs. But these universal oligos would not only amplify the targets in our microbes of interest but in others, too.

The Sanger .ab1 trace files are provided from an environmental study: check over the chromatograms, edit and remove the low-quality calls, get consensus sequences, and investigate what is being amplified (gene) and from whom (taxonomy).

.ab1 files for Activity11: (in myCourses > Content > Labs > Lab11-Activity
soil_pt2-07-02B-F.ab1
soil_pt2-07-02B-R.ab1
pond_F1_pt3-03-02-F.ab1
pond_F1_pt3-03-02-R.ab1
pond_polluted_F1_pt1-02-F.ab1
pond_polluted_F1_pt1-02-R.ab1

Activity 11
Once you finish, submit the report to myCourses (in Assignments). **Summarize your report with a table** with columns: ab1 files > consensus sequence length > top blast hit (accession number included) > taxonomic identification.