

BIOL 362 PRINCIPLES OF GENETICS

SEQUENCE DATA ANALYSIS AND CONTIG CONSTRUCTION

Today we will analyze our sequence files from the cycle sequencing reactions we amplified two weeks ago. Excess BigDye was removed from the sequencing PCR products by your instructor, and they were electrophoresed on the ABI 3100 automated DNA sequencer at the IAB Core Facility for Nucleic Acid Analysis.

Today, we will analyze the sequence files that contain the raw electropherogram data, as well as some sequence files provided by your instructor.

We will use the software program Sequencher. Sequencher takes the raw electropherograms from the sequence files and aligns the data from two or more overlapping sequences to give a consensus sequence. It does this by determining the amount of overlap, the percentage of matching bases, and whether the sequence needs to be reversed and complemented. The end result is a double-stranded consensus sequence based on two or more strands of sequence data.

Obtain the sequence tracefiles provided by your TA:

How to use Sequencher:

1. Import sequences or drag and drop them into a new file.
---Click on FILE
---On pull down menu click on IMPORT—SEQUENCES
2. Open both sequences by double clicking on them.
---Examine the sequences to make sure that they worked. It may be helpful to trim off the excess Ns on the 3'-end of each sequence (these occur where the sequencer kept collecting data, but your sequences ended because they were shorter than the maximum read-length of the sequencer, ~750 bp).
3. If they worked, highlight both sequences by holding down the shift key while clicking on the file.
4. Click on ASSEMBLE AUTOMATICALLY (if it doesn't assemble the sequences then click on ASSEMBLY PARAMETERS and lower the overlap and percentage match numbers).
5. Double click on the assembled file or contig.
6. A window will pop up. Click on BASES, this will show you the alignment.
7. Highlight a portion of your sequences in the lower window and click SHOW CHROMATOGRAM.
8. Now you can go through and examine any miscalled or unidentified (N) bases. The key to calling bases is to compare the overlapping sequence chromatograms that are visible in two new windows.
9. For example: If a + or dot is in the lower window of the sequences, click on it and curser will advance to the same position on the two chromatograms. Examine the chromatograms and determine what base should have been called. Make corrections to the sequence.

- 10.** You should not detect any heterozygous polymorphic positions in the mtDNA control region sequences, but these may exist in nuclear sequences. In these cases, you will immediately detect heterozygous polymorphic positions by the presence to two overlapping base calls (e.g., C & T). Your TA can help you distinguish between polymorphic positions and bad sequence data, which also can look like polymorphic positions. You also might observe insertions or deletions (indels) on one chromosome but not the other. This will appear as frame-shifted “bad” sequence on both strands to the 3’-end side of the insertion.
- 11.** Once you have assembled a contig, try cutting and pasting it to GenBank and searching the sequence databases using the BLAST search function at Genbank: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>