**Week 7: Sequence Analysis**

(Directions modified from Sarah Deel’s Bio 125 directions)

This handout describes the procedures you’ll use to clean up and analyze the DNA sequences from *Chamaecrista*.

**Computer Software**

This week you’ll be using two different software packages to handle your DNA sequences. First you’ll clean up the sequences in a program called 4Peaks, which will allow you to look at the chromatogram of each DNA sequence, trim the ends of the sequence, edit individual nucleotides, and create the reverse complement as necessary. Next, you’ll use an online resource, the Student Interface to Biology Workbench (SIB), to create a consensus sequence from the two sequences (forward primer sequence & reverse primer sequence) for a DNA region and then to compare the sequences from the different ecotypes. Some of you may prefer to use a software program called LaserGene, which is available on the computers in the basement of Hulings to create your consensus sequences and to compare the sequences from the different ecotypes in your search for SNPS. Those directions are also provided.

**Cleaning Up Sequences: Trimming**

1. On the computers in Hulings, log in on the **Mac** side. 4Peaks is available only for Macs. If you need help working in the Mac OS, ask for help from your lab instructor or TA as you go.

2. In the Applications folder, open up **4Peaks**.

 *Access this folder through the Go menu. Double-clicking one of the sequence files may automatically open 4Peaks.*

3. Download your sequences from the *Chamaecrista* Explorer to the Scratch Disk on your computer. You will eventually want to save them to your electronic journal and possibly your HOME directory, a flash drive, or email them to yourself, but you can most easily work with them from the Scratch Disk.

4. **Save** the file with a new name immediately, using “File | Save As…”.

 *The “undo” function in this program does not seem to work, so you may want to save your files with new names as you go; one way to do this is to add “ver” (for version) and a number (1) to the end of the file name, and just increase the number with each new version of the file. In your lab notebook, keep a running list of file names and what each represents.*

5. Next, you should familiarize yourself with 4Peaks.

a. On the screen, you’ll see the chromatogram of fluorescence, with each base indicated by a different colored line.

b. Written near the top of the chromatogram is the sequence of nucleotides, in colors matching the chromatogram. If you see an “N” among the nucleotides, it indicates a position where the sequencing instrument was unable to determine which nucleotide was present.

c. Below the nucleotide sequence is the amino acid sequence (in grey); we won’t be using this information during this lab project, but you can imagine how it might be useful.

d. Behind the chromatogram is a grey-blue bar graph. This graph represents the quality of the data (see box below).

**Base Calling Quality Scores**

The sequencing instrument uses a computer algorithm to “interpret” the fluorescence data and list a sequence of nucleotides. This base calling software also determines a Quality Score for each nucleotide, which indicates the reliability or accuracy of the base called. Quality Scores typically range from 4-60, with 4 being low quality and 60 being very reliable. The scale is a logarithmic one:

 **Quality Score** **Accuracy**

10 90 %

20 99 %

30 99.9 %

40 99.99 %

50 99.999 %

Generally, scores above 20 are considered reliable; another way to think of a score of 20 is that the probability of the base call being incorrect is 1 in 100.

e. If you move your mouse over the chromatogram (without clicking), you should see information appear about each nucleotide in the upper right hand corner of the screen, including the base, the number of the base in the sequence, and the Quality Score.

 *If this is not working for you, try single-clicking somewhere in the chromatogram and then moving your mouse around.*

**Accessing Help in 4Peaks**

There is a known bug in the current version of 4Peaks, so you cannot access the help information directly. However, if you are in the Quality Overview screen (see #6.h.i.), you can access the Help system by clicking the round button with the question mark in it. Once in the Help pages, you should be able to move around and find the information you are looking for.

f. You’ll notice a slider bar directly under the chromatogram that runs the width of the chromatogram. You can scroll through the chromatogram using this slider.

 Note that a little ways into the chromatogram of Tutorial Sequence 1 the peaks appear clean, well-defined, separate from one another, with a fairly low background (the small, secondary peaks are barely visible at the bottom). Very few nucleotides are identified as having “N” bases. This region contains good, reliable sequence information (you can check this with the Quality Scores); contrast this to the region near the end of the trace.

g. Under the main slider bar, there is a lower slider bar that lets you compress or stretch the chromatogram. If you hold down the “Shift” key while moving this slider, it will let you compress or stretch the data vertically rather than horizontally.

 *You can also resize the window to allow you to see patterns more clearly.*

h. There is a row of buttons on the lower right of the window. Here is information about the buttons you’ll want to use:

*i.* Clicking the button with the image of a DNA strand will show you the sequence in a window on the side. Clicking a nucleotide within that window will take you to that point in the trace.

*ii.* Clicking the button with the three stars lets you further investigate the quality of the sequence. Check out the Quality Overview option: the region shown in white is generally Quality Scores over 20, while the regions in pink have lower Quality Scores. Close the Quality Overview window and click the three stars button again, but this time select the Show Threshold option. This allows you to see the white and pink regions on the actual chromatogram. Go back to the three stars button to Hide Threshold when you’re ready to get back to the normal view.

i. Feel free to explore other options in the program, but remember that the Undo function may not work, so you may need to re-open the saved file if you get stuck somewhere.

6. Using the following strategies, **trim** your sequence so only the high-quality center region remains:

a. Click the three stars button and select **Show Threshold**. Look at the chromatogram and confirm that the pink areas look like lower quality data; you will be trimming the pink areas off the sequence, so you want to make sure you are not losing good sequence data.

*If you feel like there is additional sequence in the white area that is low quality, that’s okay; we’ll deal with that a little later.*

b. If you feel the area marked in pink actually contains valuable sequence data, go to the three stars button and click **Quality Overview**. Then use the **slider bar** on the left to move the threshold to something lower than 20. Re-check the pink regions on the chromatogram to make sure you’re happy with the selection.

c. When the area marked in pink looks like it is correctly identified as sequence to be trimmed off, go to the three stars button and click **Quality Overview**. Then click the **Crop** button (on the left of the Quality Overview screen) to remove the pink regions from the chromatogram and sequence.

d. **Save** your sequence (maybe with a new name).

e. Now go back and examine the ends of your chromatogram again. There may still be a few nucleotides which overlap one another or are unclear. You’ll want to trim these off as well. To do this, **select** the region you want to trim by clicking and dragging your mouse across it to select it. Go to the Edit menu and choose **Delete Selection**. This region of the chromatogram will remain on the screen, but the sequence will be deleted.

f. **Save** your sequence (maybe with a new name).

**Cleaning Up Sequences: Identifying “N” Nucleotides**

7. Identify and **edit “N”** **nucleotides** in the remaining sequence.

a. Use the magnifying glass in the upper right corner of the 4Peaks window to **search on “N.”**

b. Look carefully at the chromatogram where the N appears; do you agree with the computer’s assessment?

c. If the chromatogram shows two peaks right at the same point, or if for some other reason you can’t determine the sequence from the chromatogram, leave the sequence as an “N”.

d. If you would like to change the base, **double-click the “N” on the chromatogram to edit it**. You can change it to another letter or delete it altogether.

*If you happen to notice there is a nucleotide missing in the sequence (the chromatogram shows a peak but no corresponding nucleotide), you can add in a base by clicking on the trace to the right of where you want to add the nucleotide and, in the Edit menu, choosing “Insert Base.”*

e. To search for the next instance of an “N” in the sequence, hold down the Command key (⌘) and type “g” (or go to the Edit menu and choose “**Find Again**”). Repeat the editing process for all the remaining “N” nucleotides in your sequence.

g. **Save** your sequence (maybe with a new name).

8. Once you have cleaned up your sequences, you should look at each of them carefully at this point. How long are they? Does the sequence based off the forward primer go all the way to the reverse primer?

9. After saving each of your sequences, export them as a text file for use in later steps.

a. In the File menu, select **Export…** and accept the default option Sequence. The file name should be fine as-is.

b. Click the **Export** button to create the file.

*4Peaks will save your file in FASTA format, a common, simple format for sequences. In addition to the sequence, it includes a line for the sequence name and additional notes or descriptions (this line is preceded by a “>” symbol). When you use this file later during the lab, leave this descriptive line intact.*

**Consensus Sequences: Creating a Reverse Complement (not for all sequences!)**

10. To create a **reverse complement** of your sequence, all you have to do is go to the Edit menu in 4Peaks and select “**Flip Sequence**.” You will want to do this for your reverse primer sequences.

11. Save the reverse complement with a new name which will be clear to you later. Keep noting the file names you’re using (and what they represent) in your electronic lab notebook.

15. Export the reverse complement sequence to a text file as in #9 above.

**Consensus Sequences: Aligning Sequences using Student Interface to the Biology Workbench (SIB)**

*(Note: You may prefer to use Lasergene instead of SIB. Those directions are found after the SIB directions)*

At this point, all your sequences should be cleaned up. For each *Chamaecrista* transcript amplified with a primer set, you should have two sequences which represent that region. One sequence should have been generated from the left primer, and the other sequences should have been generated from the reverse primer. One of these sequences should have been converted to the reverse complement so that now both sequences read the same strand in the same direction. (This should be in the 5’ to 3’ direction along the coding strand of the gene which contains the DNA region.)

1. Double-check to make sure you can identify this pair of sequences.

2. Now you will use the Student Interface to the Biology Workbench (SIB) to align the pair of sequences and create a consensus sequence which takes into account information from both sequences. In Moodle, find the link to SIB on the Lab Week 5 Resources page and click on it.

3. You will need to register to use SIB, but they do not require very much information, and they will not fill your email box.

a. Follow the instructions to register.

b. On the SIB login page, after you type in your username and password, click on the Submit button rather than hitting the Return key. (For some reason, hitting Return sometimes sends you to the registration page.)

4. Learn how to get around in SIB.

a. You should be on the Preferences page. Start a new session by clicking on the NEW button; create a name for the session (like "asters") and click the "Start New Session" button. You will be returned to the Preferences page, and your new session will be highlighted at the bottom of the page.

*There are four different primary pages in SIB; here is a list, including commonly used features:*

*Preferences****:*** *start or resume sessions*

*Protein Tools****:*** *search for, get information on, and align protein sequences*

*Nucleic Tools: search for, get information on, and align DNA or RNA sequences*

*Alignment Tools:* *compare sequences and make trees*

*Creating new sessions can help you keep your data organized; to choose a particular session, go to the Preferences page, click the button beside the session you want (at the bottom of the page), and click on the RESUME button. The highlighted session is your current session.*

5. Upload your edited aster sequences.

a. Click on the Nucleic Tools button at the top of the page.

b. Scroll down until you see the Add button on the right side of the page. Click the **Add** button.

c. Click the **Browse…** button and navigate to the first sequence you would like to upload.

d. Click the **Upload File** button.

e. Type the file name (or other logical description) into the **Label** box. Leave the “>” symbol and descriptor on the top line of the box containing the sequence.

f. Click the **Save** button at the top of the page.

g. Repeat steps a-f for each of the sequences you are using.

6. Choose a pair of sequences which represent the same sequence. Select these species by clicking in the checkbox to the left of their names.

 *One of these sequences was generated from your forward primer, and the other from your reverse primer.*

7. Align these sequences by clicking the **CLUSTALW** button on the right side of the screen. Observe the number of differences between the sequences; if the sequences seem very different, check to see that you selected the correct pair, and that you generated the reverse complement sequence correctly.

8. Click the **Import Alignment** button to save the alignment.

9. Note that you are now on the Alignment Tools page, not the Nucleic Tools page, in SIB. Click in the checkbox next to your alignment, click in the “Select Box if using a nucleic alignment” checkbox at the top of the page, and click on the **BoxShade** button to the right.

10. This plot lets you easily find any differences between the files.

 *Note that BoxShade sometimes drops the last two nucleotides in the line, so if you seem to be missing nucleotides, this may explain your confusion.*

a. If you have nucleotides which do not match, open these same two sequences in 4Peaks. Note: you want good matches in your consensus, you aren’t expecting perfect matches when you move on to compare consensus sequences from different ecotypes.

b. Using the magnifying glass tool, search for the sequences surrounding any mismatched bases to locate them in the chromatograms. Once you have located them, determine whether one base or the other seems more accurate.

c. Open one of the text files containing one of these sequences, either in a program called TextWrangler (if available), or in Word. Save this file with a new name which indicates it is a consensus sequence. Adjust the descriptive information after the “>” symbol to reflect the new name of the consensus sequence. Edit this file using the information from SIB and 4Peaks.

 *Avoid using TextEdit to edit your files; it inserts formatting which can cause problems later.*

d. In addition to editing individual bases in TextWrangler or Word, you will usually need to add nucleotides to one end of the sequence you are editing. Look at the consensus sequence in SIB to determine where you need to add the nucleotides.

 *Your two sequences will be mostly overlapping, but you will probably find “tails” on either end which come from only one sequence. Your consensus sequence should include both tails.*

e. Save your document as a text (.txt) file. You should add a “C” to the end of the file notation to denote that it is a consensus sequence.

 *Your file name should retain the primer notation (F or R) for the sequence you did not take the reverse complement of. This means that for E-rbcL, N-rbcL, and W-ITS, the last two letters of your file name will be “FC.” For K-matK, the last two letters of the file name will be “RC.” Please be careful with this naming convention, since you will be sharing this file with others in the class.*

f. Repeat this procedure for each transcript for which you have sequences.

11. Repeat this procedure to compare the consensus sequences from the different ecotypes to determine whether or not your predicted SNP is present.

**Check the DNA Sequence Against the *Chamaecrista* transcriptome**

Next, you need to double-check that the region you have sequenced is the region you thought you were sequencing. You will do this by comparing your sequence to the sequence in the *Chamaecrista* transcriptome using the local Blast in the *Chamaecrista* Explorer.

**Using LaserGene software – An alternative to SIB**

**Consensus Sequences: Aligning Sequences**

At this point, all your sequences should be cleaned up from your work with 4Peaks. For each *Chamaecrista* transcript amplified with a primer set, you should have two sequences which represent that region. One sequence should have been generated from the left primer, and the other sequences should have been generated from the reverse primer. One of these sequences should have been converted to the reverse complement so that now both sequences read the same strand in the same direction. (This should be in the 5’ to 3’ direction along the coding strand of the gene which contains the DNA region.)

1. Double-check to make sure you can identify this pair of sequences.

2. Now you will use the LaserGene program SeqMan to align the pair of sequences and create a consensus sequence which takes into account information from both sequences. Find the Applications folder on the Mac and navigate to the DNASTAR folder; inside that is a LaserGene 8 folder, and inside that is the SeqMan application. Double-click **SeqMan** to open the program.

3. This program is quite powerful, and you won’t be exploring it like you did 4Peaks. First, you need to add the sequences to be aligned.

a. In the window labeled “Unassembled Sequences,” click the **Add Sequences…** button.

b. Choose a pair of sequences which represent the same DNA region from the same ecotype. Add these species by selecting them (either one at a time or at the same time) and clicking the **Add File** button. When they are both in the Chosen Files box, click **Done**.

 *One of these sequences was generated from your forward primer, and the other from your reverse primer.*

c. Click the **Options…** button in the Unassembled Sequences window.

d. **Uncheck “Trim sequence ends”** and click the **Scan Later** button.

e. Click the **Assemble** button in the Unassembled Sequences window.

f. A window called “Untitled.sqd” will open. Click “**Contig 1**” in that window to select it.

g. In the Contig menu, choose **Alignment View** to see your sequence alignment.

h. Scroll through the alignment. using the scroll bar at the bottom of the Alignment window. Note that bases or regions which do not agree are highlighted in red.

i. In the Contig menu, choose **Strategy View** to see an overview of the sequence. If you click in the Conflicts box, you can see red lines where there are regions which do not agree.

j. If all your bases are in alignment and perfect agreement, skip down to 3. o. below.

k. If you have nucleotides which do not match, determine where they are (note their position number by selecting them in the Alignment window).

l. Open the sequence files in 4Peaks again to determine if one base or the other appears more accurate at a given position.

m. You can **edit** the nucleotides directly in the Alignment window of SeqMan as needed; just select them in the Consensus (top) sequence and change them to the correct nucleotide.

n. Repeat this procedure with other nucleotides not in agreement.

 *You should not have large areas of disagreement if your sequence quality is reasonable. If you do, double-check to make sure you selected the correct sequences and that one is the reverse complement of the other.*

o. **Save** the alignment project in SeqMan by going to the file menu and choosing “Save…”

p. Save a more accessible copy of your consenus sequence. Select Contig 1 in the project window (it now has the title you just gave your project as the window title). In the Contig menu, choose **Save Consensus | Single File…** When the dialog comes up, save the file in FASTA format. Add a “C” to the end of the file notation to denote that it is a consensus sequence.

q. Repeat this procedure for each pair of sequences (i.e. the ones sequenced with matched left and right primers).

4. Repeat this procedure to compare the consensus sequences from the different ecotypes to determine whether or not your predicted SNP is present.