

Supplement 1: Worksheet completed by students on week-5.

Using *iFinch* Bioinformatic Software to Analyze DNA Sequencing Output

Examine the sequence output:

1. Log onto the iFinch website (Look for the shortcut on your computer desktop). The website is: <http://classroom1.bio-rad.ifinch.com/Finch/>

User name: BR_guest Password: guest

2. Click on Folder from the Chromats menu.

The screenshot shows the iFinch website interface. At the top, it says 'iFinch' and 'User: student1 Feb 12, 2008 3:58'. There are links for 'Get FinchTV', 'Quick Search', and 'Logout'. The main navigation menu includes 'Chromats', 'Variants', and 'System'. Under 'Chromats', 'Folders' is highlighted with a red box. A 'Data' table is displayed with the following information:

Data	
Projects	4
Folders	4
Chromatograms	192
Other Files	0
Undistributed Data	0
Active Users	4
Variants	0

Below the table, there are links for 'Sequence Analysis Tools', 'Sequence assembly', 'Sequence Utilities', and 'NCBI BLAST'. A 'Finder' button is also visible in the top right area of the page.

3. A page will appear with a list of folders. Click the folder called "salvia1"
4. A page will appear that presents some of the data from the folder. Each row contains data from a single chromatogram. The four chromatograms are from a single GAPDH clone from the *Salvia* plant that has been sequenced using different primers.

The screenshot shows the 'Contents of Folder A145982_TE' page. At the top, it says 'iFinch' and 'User: sandy1 Feb 8, 2008 1:06'. There are links for 'Get FinchTV', 'Quick Search', and 'Logout'. The main navigation menu includes 'Chromats', 'Variants', and 'System'. Under 'Chromats', 'Folders' is highlighted with a red box. A 'Finder' button is highlighted with a red box. Below the 'Finder' button, there is a search bar with 'Label' selected and 'CONTAINS' in the dropdown. A 'Go' button and a 'Reset' button are also visible. Below the search bar, there is a table with the following columns: Label, Rev, Len, Trim, Q20, Q20/len, Q40, Q40/len, A_sig, C_sig, G_sig, T_sig. The table contains 96 items, and the first 25 items are displayed. A 'Download for MS-Excel' button is visible in the bottom right corner.

Label	Rev	Len	Trim	Q20	Q20/len	Q40	Q40/len	A_sig	C_sig	G_sig	T_sig
QCQP844033.b2_A01.ab1	1	868	722	648	0.75	486	0.56	147	156	215	254
QCQP844034.b2_C01.ab1	1	880	756	749	0.85	610	0.69	215	156	215	254
QCQP844035.b2_D01.ab1	1	868	722	648	0.75	486	0.56	147	156	215	254
QCQP844036.b2_E01.ab1	1	868	722	648	0.75	486	0.56	147	156	215	254
QCQP844037.b2_I01.ab1	1	862	666	631	0.73	465	0.54	107	80	304	176
QCQP844038.b2_K01.ab1	1	867	732	667	0.77	514	0.59	129	59	195	101
QCQP844039.b2_M01.ab1	1	867	812	758	0.87	613	0.71	518	300	487	409

5. Open the first file by clicking the FinchTV icon (located to the left of the chromatogram labels). Answer the questions below.



Contents of Folder DEFAULT

[[Relabel chromatats](#) | [Move chromatats](#) | [Download folder data](#) | [Folder report](#)]

Chromatograms (96) Other (0)

Find: Q20 IS NOT NULL [Go] [Reset] [?]

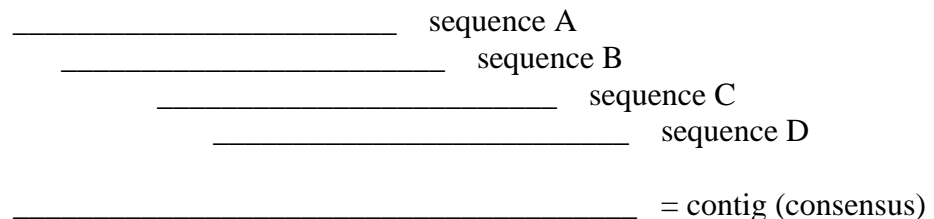
Items 1-25 of 82 ([more](#) | [all](#)) Page 1 of 4 Go To: [1](#) [2](#) [3](#) [4](#)

	Label	Rev	Len	Trim	Q20	Q20/len	Q40	Q40/len	A_sig	C_sig	G_sig	T_sig
<input type="checkbox"/>	 QCDP869442.b2_C17.ab1	1	913	869	771	0.84	603	0.66	801	570	651	1107
<input type="checkbox"/>	 QCDP869395.b2_E05.ab1	1	947	722	765	0.81	567	0.60	504	491	391	643

- What is the length of this sequence? _____ bp
- The low quality sequences (highlighted in gray) are usually located at the beginning and the end of each run. At what base-pair position do you start seeing high-quality sequence? _____
- At what base-pair position do you see the end of the high-quality region? _____
- Examine the low-quality bases. Compare a low quality base with a high quality base. How are they different? What are the characteristics of a low quality base?

Assemble a ‘contig’ from four sequences representing this gene:

The four *Salvia* sequences you see represent overlapping regions of the same gene. iFinch will automatically ignore the low-quality reads of each sequence and prepare the data so it can be pieced together, like a puzzle, into a single ‘contig’ (contiguous DNA sequence). The four runs might fit together in schematic like this:



Computer programs have been written to do this type of thing. Below is an example. Say you had two sentences “Key Slime Pie” and “Big Mickeys”. The computer looks for parts of the phrases that match by building a “Dot Plot Matrix” like the one below. The two axes are made up of the two phrases that are being aligned. To find the alignment put an ‘X’ in every part of the matrix where letters in both phrases match up. Ignore spaces between the words. The first two matches are done for you. You do the rest....

KEY SLIME PIE & BIG MICKEYS

	K	E	Y	S	L	I	M	E	P	I	E
B											
I						X				X	
G											
M											
I											
C											
K											
E											
Y											
S											

Do you see a pattern anywhere? Look for a line of “X”s (like in tic-tac-toe). The overlapping regions are indicated by these patterns. Write below what the full sentence might be if these two phrases were pieced together where they overlap:

(BIG MICKEYS LIME PIE)

This is what the computer does with DNA sequences that contain portions that overlap. Instead of simple “X”s, though, it actually quite mathematical.

- In the ‘salvia1’ folder, click on “Download folder data”. In the ‘Download Sequences’ section, at top, click on the “Export Sequences” button. Select the ‘Open’ option. If you have problems, Save it to the Desktop first and then Open that file using WORD. You might have to Browse through different programs to find WORD.
- This data is in a format called FASTA. In this format each new paragraph is signaled by the arrow (>) symbol. We are going to copy-and-paste everything on this page, so highlight it all and copy.
- Go to the iFinch Home page and click the Sequence Assembly link to access the University of Lyon CAP3 web service (<http://pbil.univ-lyon1.fr/cap3.php>). CAP3 is an assembly program that will piece these four sequences together.
- Paste the copied text into the search box.
- Click the Submit button to start the assembly.

11. When the assembly is complete, a page will appear with links to your results. These links are:

- a. Contigs
- b. Single sequences
- c. Assembly details
- d. Your sequence file

We're going to look at each of these in turn. Use the back button on the browser to return to the assembly results page after viewing each page.

12. Click Your Sequence File first to make certain that you pasted the correct information into the form.

13. Next, click Single Sequences. Sequences appear here if they could not be used in the assembly. Your page should be blank.

14. Click Assembly Details. This shows how the four runs fit together to form a longer contig. It only shows about 60 bases per line. Below that is the 'consensus' sequence (called the contig). Draw a schematic like the one shown on page 1 that accurately portrays how the other three sequences align with the 'A01' sequence. Show where the C01, G01, and I01 sequences fit, below:

_____ A01 sequence

15. Click Contigs. You should see one contig sequence. Notice it is in the FASTA format. Highlight and copy this entire sequence for the next step of the exercise, below.

Use BLASTn to Search the GenBank for Similar DNA Sequences

16. Click the Geospiza Finch in the top left corner to return to the iFinch home page.

17. Click the link to NCBI BLAST. In the BLAST page, look under the BLAST heading and choose nucleotide BLAST.

18. Paste your FASTA sequence into the “Enter Query Sequence” box.
19. Next, in the “Choose Search Set” section, open the database pull-down list and select “Reference genomic sequences (refseq_genomic)”.
20. Click the button to choose BLASTn as the program. BLASTn is more sensitive than the other programs and allows you to find “somewhat similar sequences”.

The screenshot shows the NCBI BLASTn search interface. The 'Enter Query Sequence' section includes a text input field for the sequence, a 'Clear' button, and a 'Query subrange' section with 'From' and 'To' input fields. Below this is an 'Or, upload file' section with a file input field and a 'Browse...' button. The 'Choose Search Set' section has a 'Database' dropdown menu set to 'Reference genomic sequences (refseq_genomic)', with radio buttons for 'Human genomic + transcript', 'Mouse genomic + transcript', and 'Others (nr etc.)'. Below this is an 'Organism' section with an input field and a 'Choose Search Set' button. The 'Program Selection' section has radio buttons for 'Highly similar sequences (megablast)', 'More dissimilar sequences (discontiguous megablast)', and 'Somewhat similar sequences (blastn)', with the last one selected. At the bottom, there is a blue 'BLAST' button and a 'Show results in a new window' checkbox. A note at the bottom right says 'Note: Parameter values that differ from the'.

21. Click the blue BLAST button. After a few moments, the results will appear.

Answer these questions about the BLAST results:

22. How many BLAST results did the search find (look near top of graphic)? _____
23. Look at the top of the list of matches (shown below the graphic). What species and gene is the best match? To find out click on the hyperlink under “Max Score”.
24. Which non-plant species is most similar to our sequence from *Salvia*? _____

