

DNA Sequence Analysis Tutorial

Genetics Laboratory

By Jason Evans, Luke Sheneman and Celeste Brown

1 What is VectorNTI?

From the Informax website:

VectorNTI equips laboratories with an extensive range of software tools for sequence analysis and molecule manipulation. Integration of data and analysis tools is achieved using an intuitive, object-oriented database for the storage and organization of DNA, protein and oligonucleotide sequences, and other molecular biology data. From the Database Explorer, users can launch comprehensive, publication-quality views of any sequence in an integrated, multi-pane Molecule Display window.

Within this window, make a selection of some sequence graphically--an Open Reading Frame (ORF), for example--and apply all the tools necessary to design multiple sets of PCR primers, initiate a BLAST search against any division of GenBank, translate using any one of numerous genetic codes, or automatically design a cloning strategy to create a new recombinant molecule. The VectorNTI Local Database stores any new protein sequences generated, or PCR primers designed, and keeps a record of the parent-descendant relationships that might exist between molecules. These workflows can be performed without any reformatting of data, allowing the researcher to concentrate on the science, rather than the technology, underlying their bioinformatics analyses. Many additional bioinformatics workflows are possible within Vector NTI.

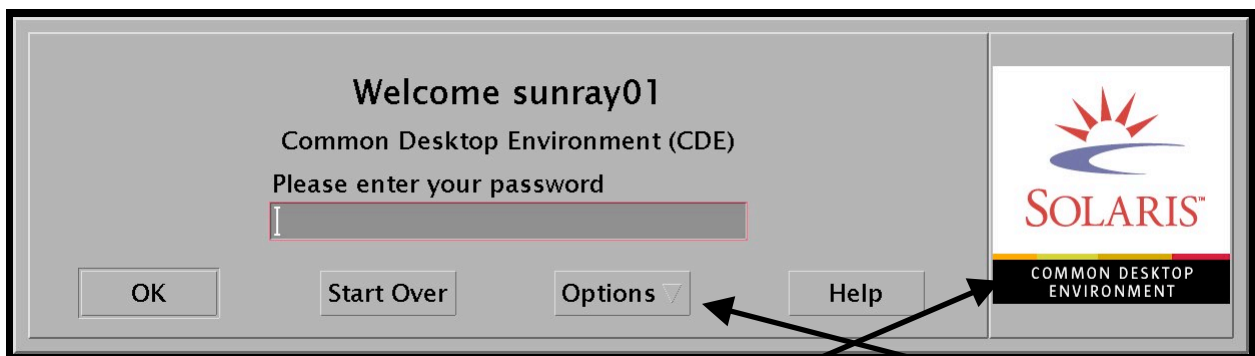
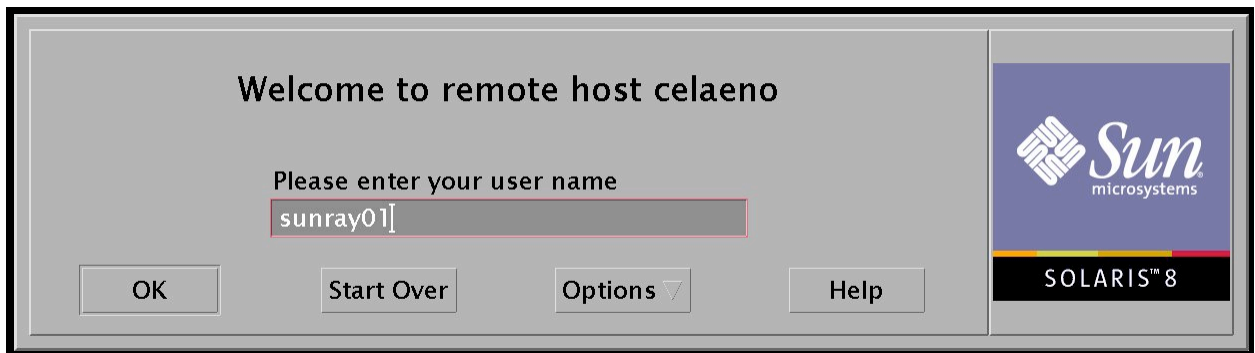
- ❖ Molecular biology data management: storage of DNA/RNA Molecules, Protein Molecules, Enzymes, Oligos, Gel Markers, Citations, BLAST Results and Analysis Results
- ❖ Creating, mapping, analyzing, annotating and illustrating DNA and protein sequences
- ❖ PCR, sequencing and hybridization primer/probe design
- ❖ BLAST searching, viewing of results, and recovery of hits
- ❖ Recombinant cloning strategy design
- ❖ PubMed/Entrez searching, viewing and recovery of results
- ❖ *In silico* gel electrophoresis
- ❖ Connectivity to numerous Internet analysis tools
- ❖ Primer Design
- ❖ DNA Sequence Analysis
- ❖ Protein Sequence Analysis
- ❖ BLAST Searching
- ❖ Database Explorer

2 Goal

You were given the DNA for an organism, but all you know about the organism is that it is a mammal, and you were asked to determine what species the sample came from. You have amplified CytB and now have the sequence for the amplified product on both strands. You will edit the sequence data to correct any ambiguities between the two sequences. You will use the resulting sequence to search for organisms most similar to your mystery organism.

3 Getting Started

Log onto the computer using the name of the terminal (bottom left hand corner of the monitor) as the login name, and the name of the terminal preceded by an exclamation point as the password:
USER name: sunray01 PASSWORD: !sunray01 (the password will not appear in the window).



Be sure that the system is Common Desktop Environment by selecting Options: Session: Common Desktop Environment. **REMEMBER which computer you are working on so you can find your results for the next computer lab.**

Click on the Windows like icon (rdesktop) in the bottom left hand corner of the screen to access the Windows 2000 computer. Use the same login and password.



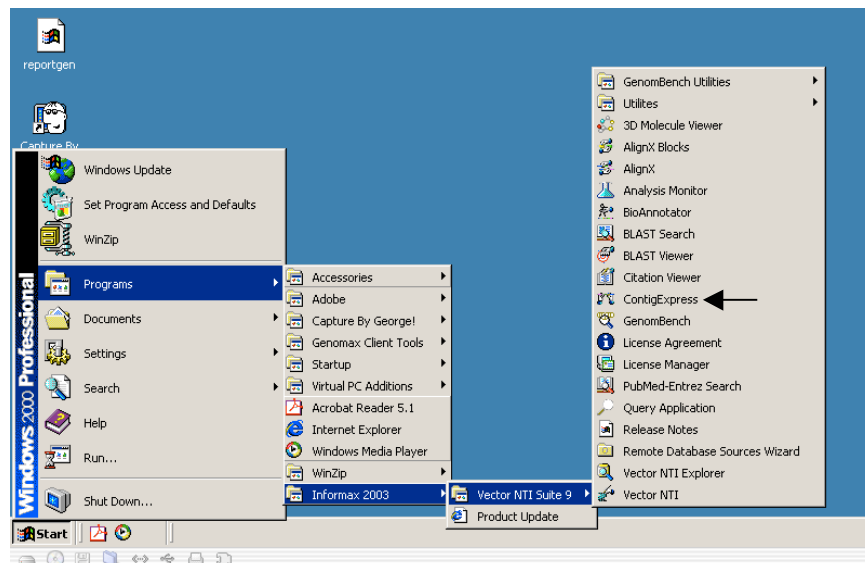
4 Using ContigExpress

We will use ContigExpress to create a contiguous DNA sequence from your raw chromatograph data produced by the automated DNA sequencer. ContigExpress is a tool that allows you to assemble DNA fragments, both text sequences and raw chromatograms directly from an automated DNA sequencer, into overlapping sequences known as “contigs”. Users can create ContigExpress projects and store sequences, contigs, and the options used when assembling the contigs.

In order to create a contig from distinct DNA fragments, you will learn how to

- Create and manipulate a ContigExpress project
- Assemble contigs
- Edit a contig to remove ambiguities
- Export a contig as a single DNA sequence for further analysis

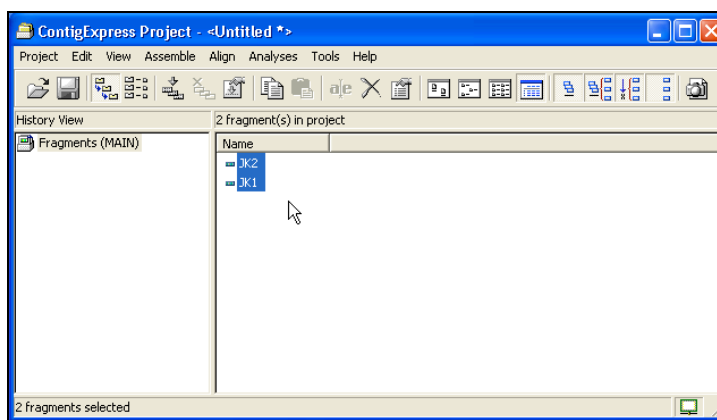
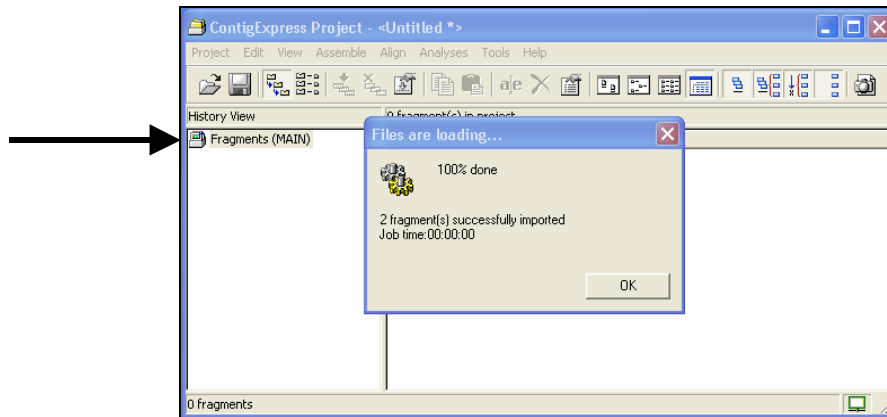
To start a new ContigExpress project, click on **Start: Programs: Informax 2003: VectorNTI Suite9: Contig Express**:



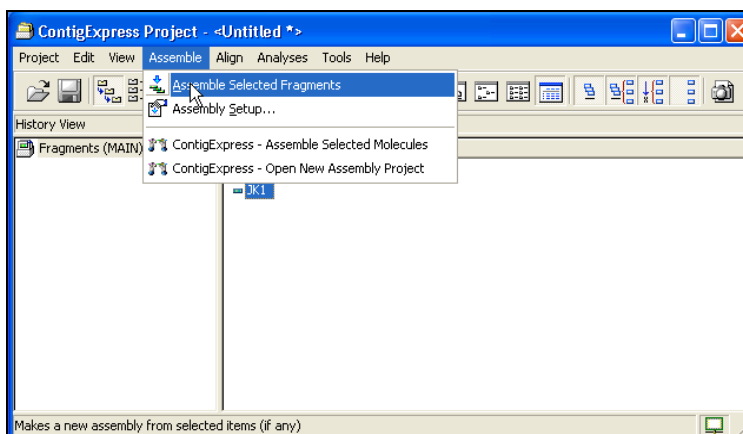
The ContigExpress Project window will open. At this point, you need to get your sequencing results. Click on **My Computer** and then click on the CD drive. Open the folder for your lab

section, and then open the folder with your group's two abi files. Select the two files and drag them to the **Fragments(MAIN)** folder in the ContigExpress Project window.

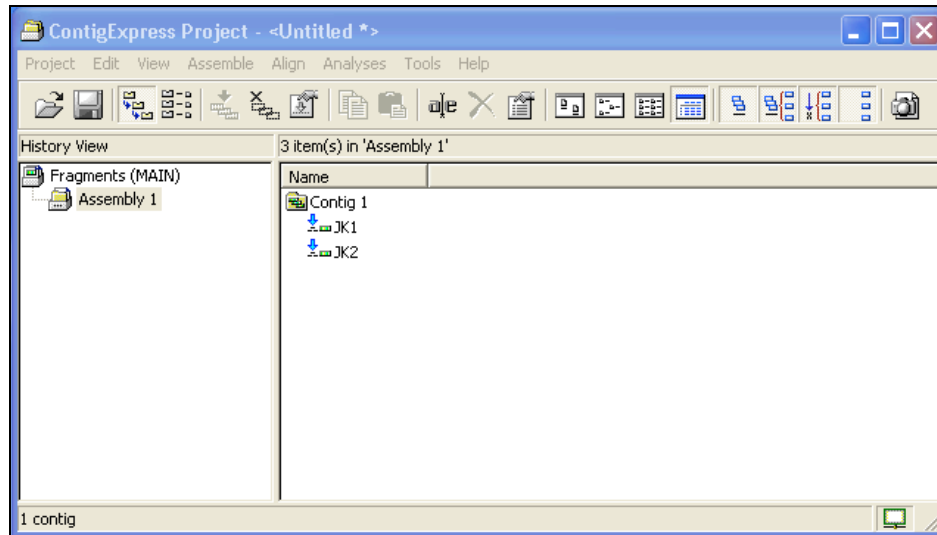
Once you drop the fragments into ContigExpress, you should get a message indicating that the fragment importing was successful.



Once both of the sequence fragments have been imported into ContigExpress, select both of the fragments inside of the ContigExpress project window. Select **Assemble** → **Assemble Selected Fragments** from the ContigExpress pull-down menu. The program tells you how many contigs were assembled and how long it took. Click on **OK**.

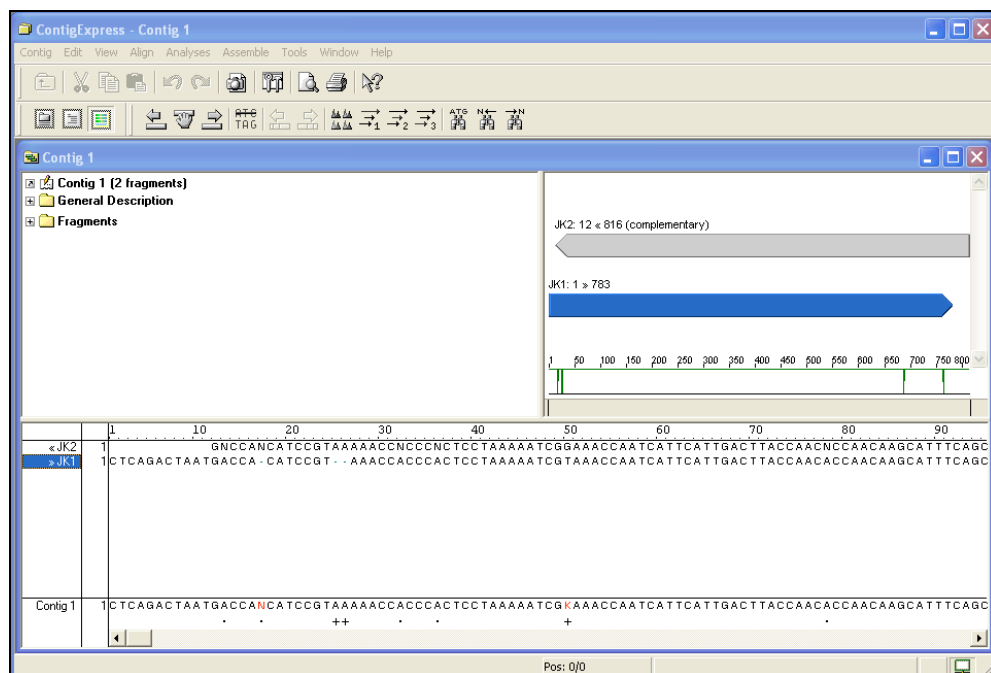


Assembling DNA fragments essentially performs an alignment of the overlapping region of the two fragments. Your window should now look like this:



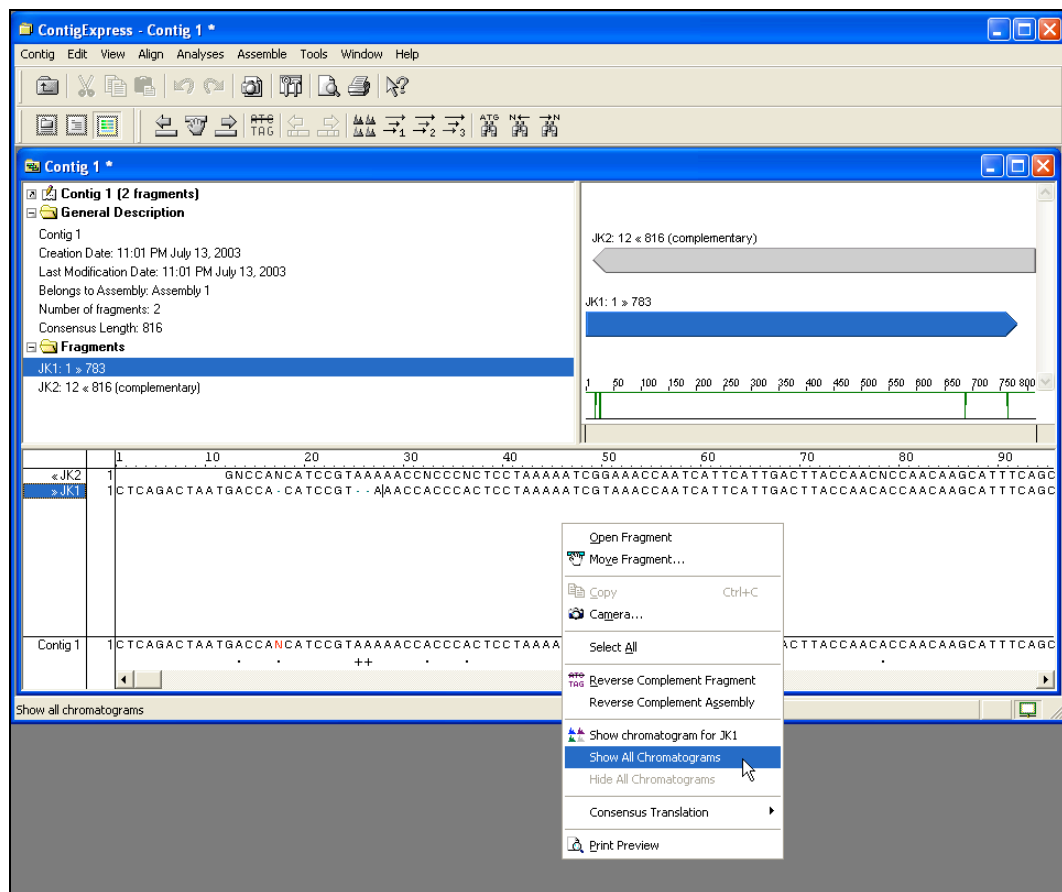
Contig 1 is now the only contig in Assembly 1. Rename the Assembly by clicking on “**Assembly 1**” and then waiting a second and clicking again. Change the name to reflect your Class Period and Name (eg. TuAM_CBrown). Contig 1 is shown with two sub-fragments. You can rename your Contigs and your individual fragments if you like.

Now double-click on your contig (i.e. “**Contig 1**”) in order to edit it. A contig editor window will open:

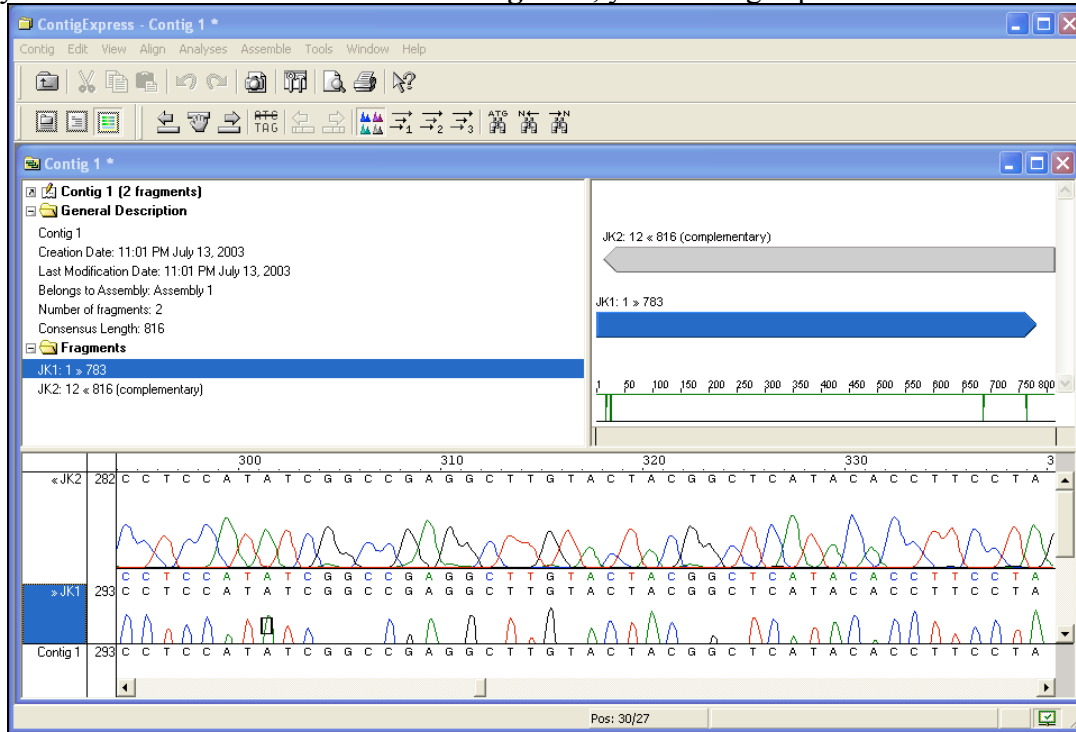


This window is broken up into three main panes. The upper left pane contains simple information about the contig, a textual annotated description of the contig, and a description of the fragments within the contig. The upper right pane presents a graphical representation of the contig and the way in which the fragments overlap in the contig. Also shown in the lower portion of this pane is the location of known discrepancies in the contig. These discrepancies represent nucleotides that differ between the two sequences, and will need to be resolved. The lower pane depicts the actual sequences themselves and depicts the fundamental details of the contig, showing the exact way in which the contig was constructed from the sequence fragments. Ambiguities and gaps are presented here in detail, and can be resolved manually in this lower pane.

The next step is to manually resolve all discrepancies in the contig. Prior to doing this, it is best to show the detailed chromatogram output from the automated sequencer. This can be done by right-clicking anywhere in the lower pane and selecting **Show All Chromatograms** from the pop-up window.



Once you have selected **Show All Chromatograms**, your ContigExpress window looks like this:



In order to manually remove all ambiguities from this fragment assembly, you will:

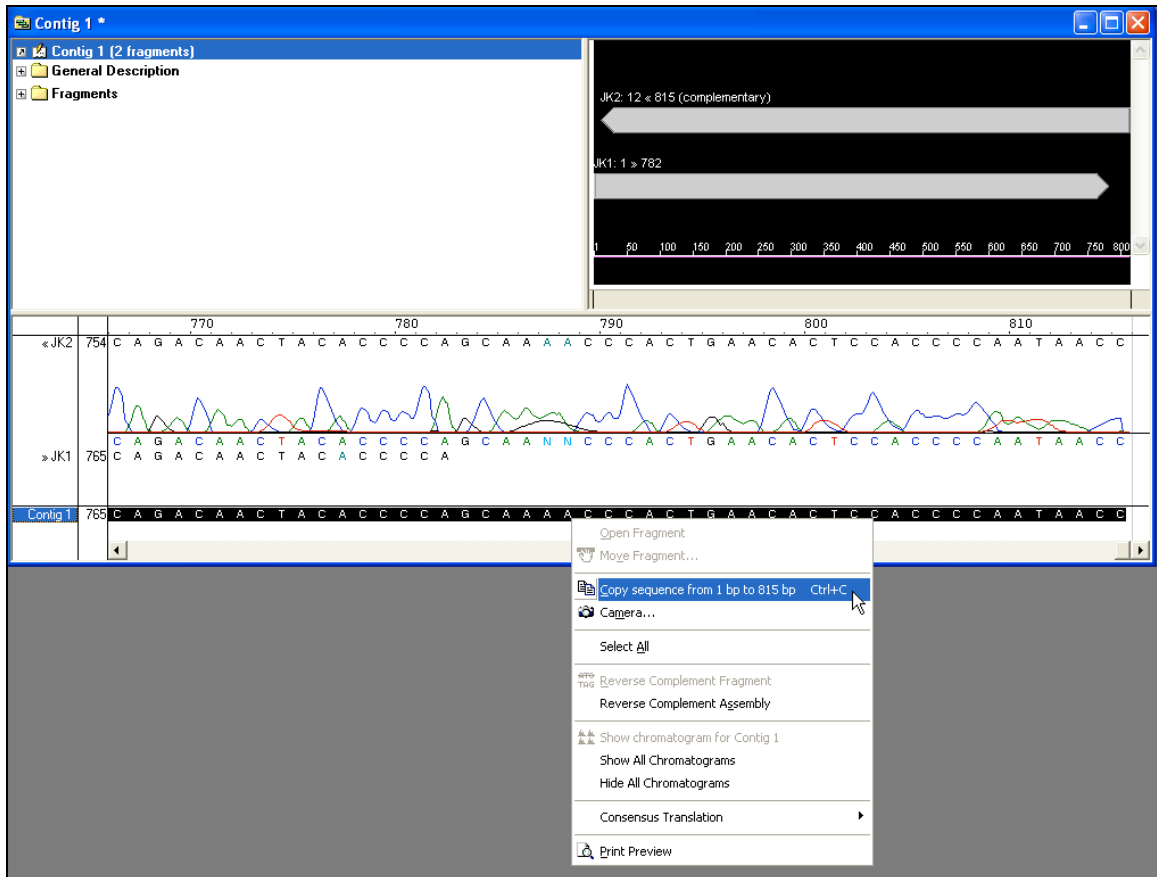
- ❖ Locate all ambiguity codes (depicted in red in the consensus sequence at the bottom of the lower pane, as a symbol underneath the consensus sequence and as green vertical lines in the graphics pane) and resolve the ambiguity character with the most likely alternative. In most cases, the most appropriate solution is simply to choose the unambiguous code from the other sequenced fragment in the same aligned column (where it overlaps).
- ❖ Locate and remove all gaps. In some cases, the solution here is to remove the entire column, while sometimes the solution is to change the gap code in one of the fragments to the most obvious alternative character. Use the chromatogram as your guide. Also, choose the better chromatogram. In the picture above, the top chromatogram is good and the lower chromatogram is very poor.

Action	How to Perform	Sequence Pane Result	Chromatogram Pane Result
Delete	Select residues; press Delete	(□) replaces NTs; Nts moved below strand	(- - -) appear in upper sequence
Insert	Place caret; type new Nts	(↑) appears below new NTs; new NTs are colored	A break appears in the chromatogram
Replace	Select NTs; type new Nts	New NTs appear in strand; replaced NTs moved below strand	New NTs appear in upper sequence; no break in chromatogram

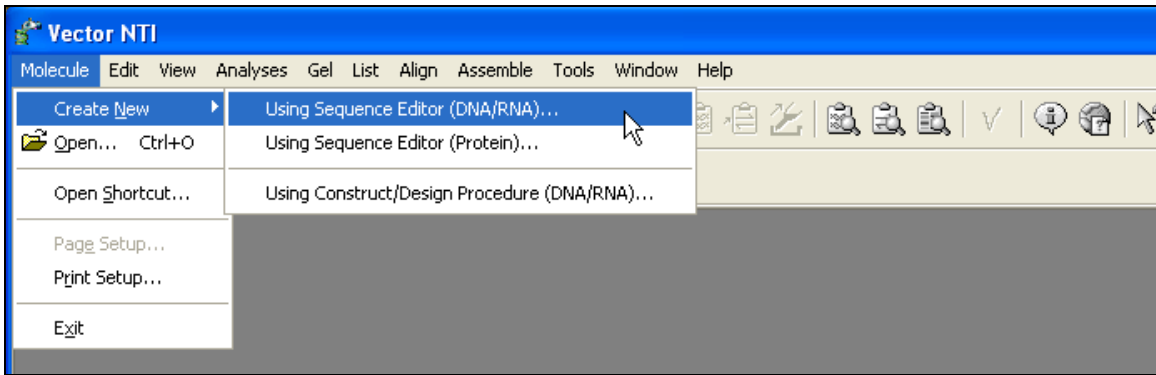
Finally, you should delete the consensus sequence where there is only sequence in one direction.

5 Saving the Contig as a full, standalone DNA sequence

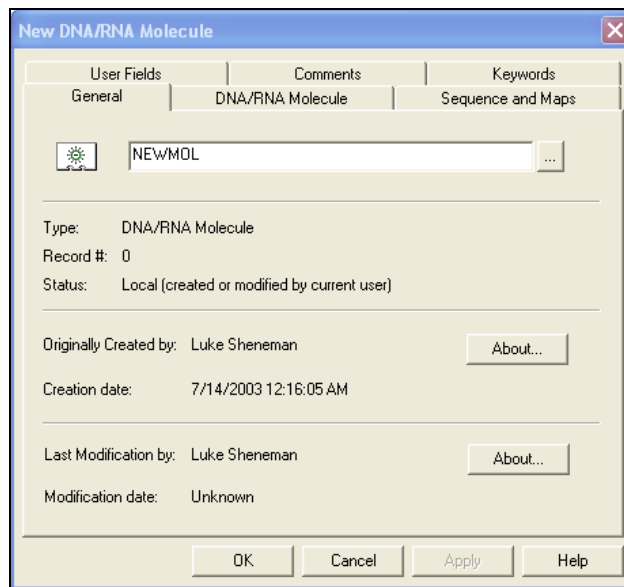
Once all ambiguities are resolved, select (highlight) the entire consensus sequence. The easiest way to do this is to right click on **Contig1** in the bottom pane and select **Select All**. Right click on the consensus sequence and select **Copy sequence from 1 bp to n bp**



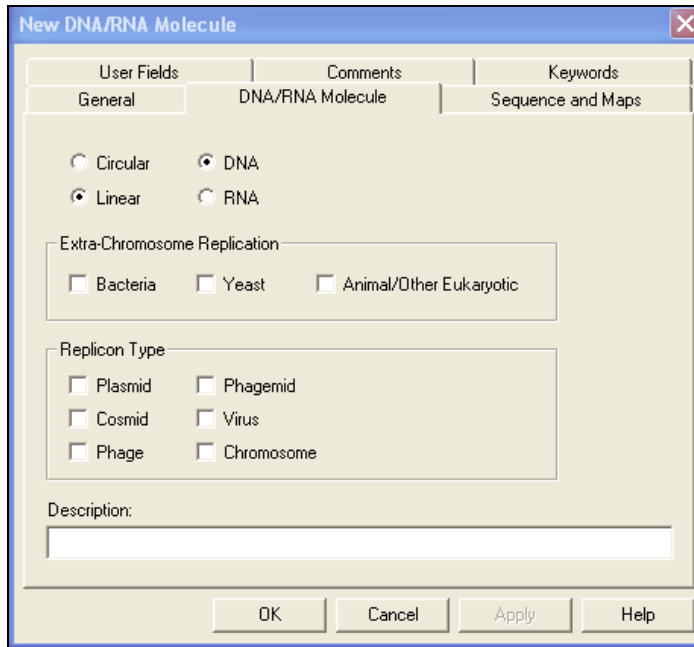
At this point, the consensus sequence representing the assembled contig of the two DNA fragments is in your cut/paste buffer. Open the main Vector NTI window **Start: Programs: Informax 2003: VectorNTI Suite 9: VectorNTI**. Select **File → Create New Sequence → Using Sequence Editor (DNA/RNA)**



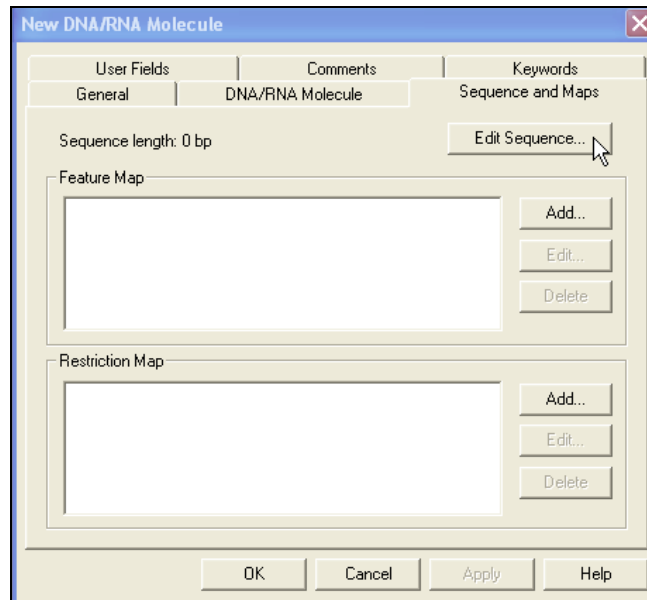
A dialog pops up which allows you to annotate, name, and edit the characteristics of the new molecule:



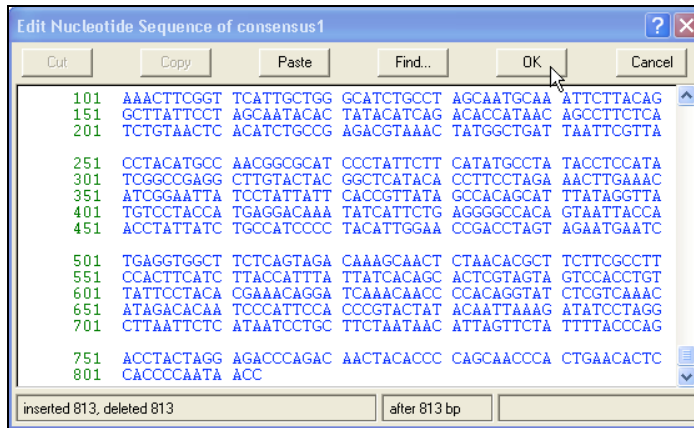
In the General tab, give the new molecule a name that reflects your lab Day and Time and your Name (Tu_AM_CBrown). Next, click on the **DNA/RNA Molecule** tab and make sure that the DNA is marked as linear and that the molecule is treated as DNA rather than RNA.



Next, click on the **Sequence and Maps** tab in this dialog, and then click on the **Edit Sequence** button in this dialog.

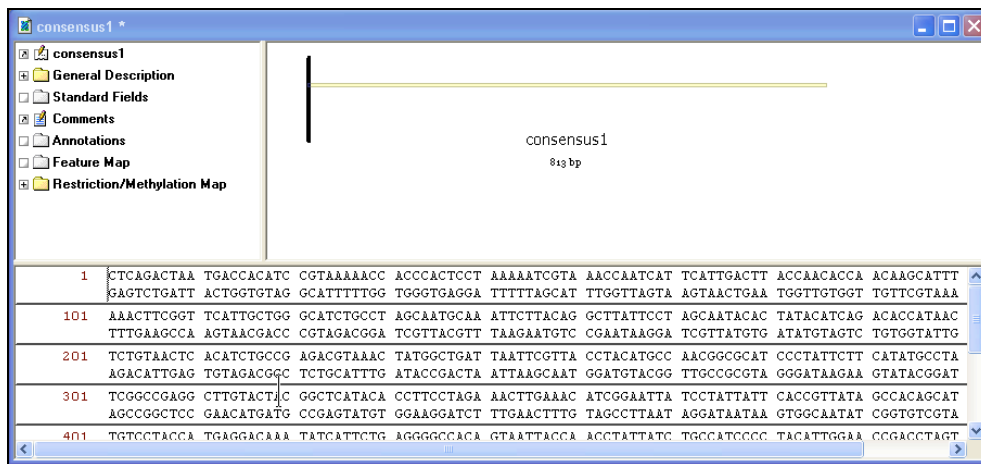


A dialog box will open allowing you to enter the complete DNA sequence of the consensus contig. Paste the consensus contig sequence into this window by clicking on the **Paste** button, and then pressing **OK** and then **OK** again.



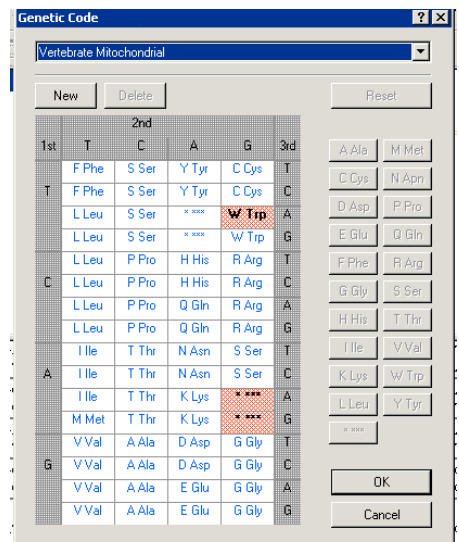
You should now have a linear DNA sequence that represents the fully assembled and cleaned-up DNA sequence which was output from the automated DNA sequencer. Your next job is to identify the species from which this DNA sample originated.

Open your new DNA sequence by double-clicking on its name inside of your main Vector NTI window. The Molecule viewer will pop up with your sequence inside:

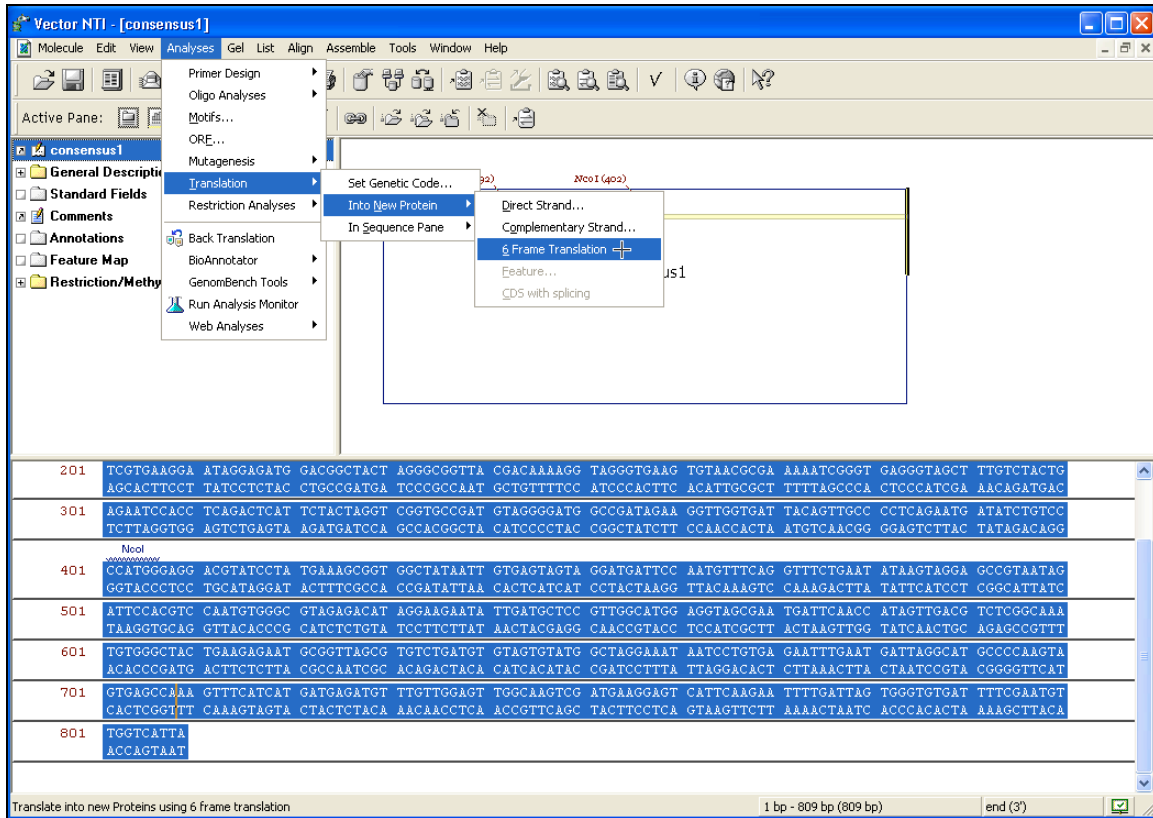


6 Translation into all 6 reading frames

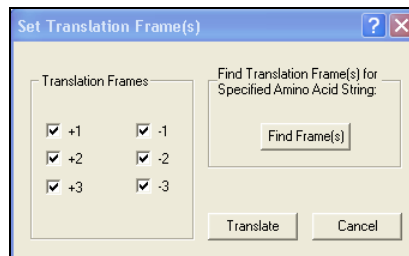
At this point, we want to translate the DNA into protein. Since we don't know the proper reading frame for this translation, we need to perform a translation for all 6 reading frames. The sequences that you are working with are encoded on the mitochondrial DNA, and mtDNA uses a genetic code that is different from the Standard genetic code. In order for your translations to be accurate you will change to the vertebrate mtDNA genetic code. Select **Analyses** → **Translation** → **Set Genetic Code**. And then change from Standard to Vertebrate Mitochondrial in the pull down menu.



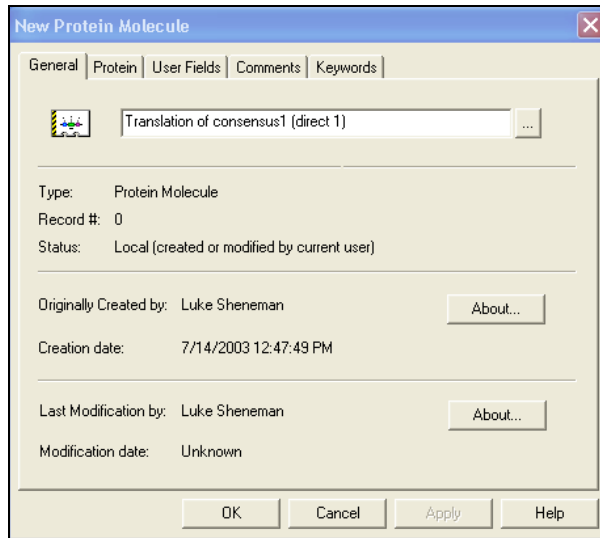
Click OK and when the Code window closes, select the entire nucleotide sequence by clicking in the sequence pane and then going to **Edit → Select all**. Now you can translate the sequence: **Analyses → Translation → Into New Protein → 6 Frame Translation**



A pop-up dialog will appear, asking which frames need to be translated. Just press **Translate** to get all 6 possible reading frames.



After pressing **Translate**, the following dialog will appear:



Change the name of the molecule to reflect your class period and name but keep the strand designation (direct 1), then press **OK**. One of these dialogs will appear for each of the 6 translated proteins; change the names accordingly for each of them.

At this point, you have added 6 new protein sequences to your protein database. In order to see all of the files, go to **Window → Cascade** and all 7 sequences will be visible.

7 Identify Correct Reading Frame

The goal here is to identify the most likely reading frame, under the assumption that there are no major sequencing errors. Only one of these proteins was correctly translated, since only one of them could be in the correct reading frame during translation.

Look through all six translated proteins and identify the correct translation/reading frame. Identifying the most likely protein is relatively simple. Look for a protein sequence with the longest contiguous section of amino acid residues which lack stop codons. Stop codons are represented with asterisk character. For example, the following sequence was not likely translated in the correct reading frame, because it is sprinkled with asterisks and has no long sections without stop codons:

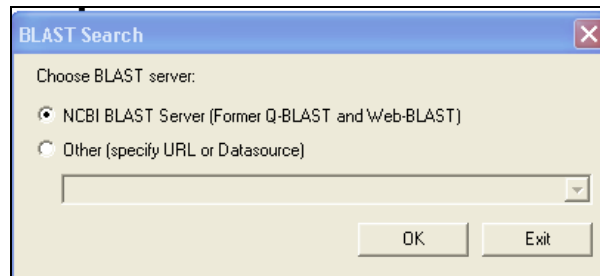
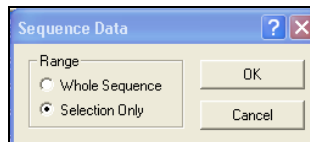
1	MTNIRKSHPL IKILE*LLHR LANSNKHLIM MKLWLTWGM PNHSNSHRII
51	S*PYTTHQTR *PHSLQ*PTF AETSTMVESF ATSMPTEHQY SSYVSTPTLD
101	VESITAPTYI QKPETLESSY YSQL*PPLS* DTSSHGDRYH SEGQL*SPTF
151	YRPSPTSAPT **NESEVDSQ *TKLPSPDFS RYTSPLYLLS* PP**PSISYS
201	FTRQDPTTPQ A*SPIPTKFKH FIHTTQSKTS *AYSSSS*SY SY*PYSPQTC
251	*ENPDNYTPA NPLSTPPHI

8 Blasting proteins against SwissProt

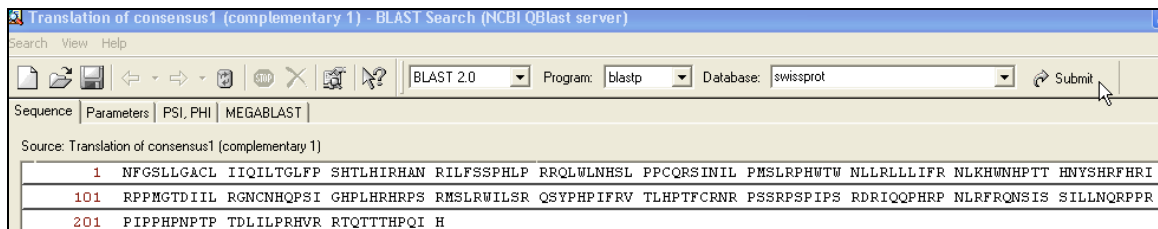
Once you have identified the best possible protein sequence, highlight the largest section of the sequence between stop codons (as shown below), and perform a BLAST search against the SwissProt database using this selected subsequence. Here, you will identify the sequence in the SwissProt database with the most similarity to your sequence.

1	**PTFENHTH *SKFLNDSFI DLPTPTNISS **NFGSLLGA CLIIQILTGL
51	FPSHTLHIRH ANRILFSSPH LPRQLWLNH SLPPCQRSIN ILPMSLRPHW
101	TWNLRLLLI FRNLKHWNHP TTHNYSHRFH RIRPPMGTDI ILRGNCHQP
151	SIGHPLRHR PSRMSLRWIL SRQSYPHPIF RVTLHPTFCR NRPSSRPSPI
201	PSRDRIQQPH RPNLRFQNS ISSILLNQRP PRPIPPHPNP TPTDLILPRH
251	VRTQTTHP QIH*APHPT

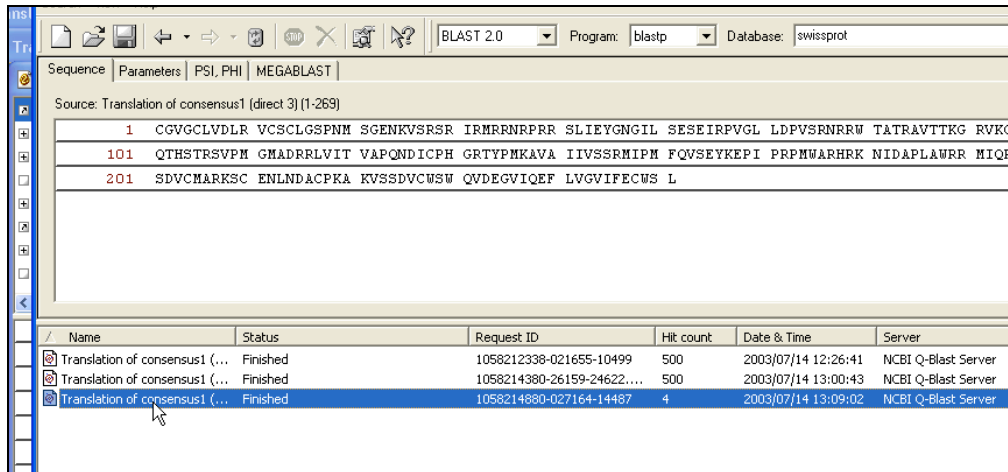
Perform the BLAST search by selecting **Tools** → **BLAST Search**, and then performing the search against the selected portion of the sequence:



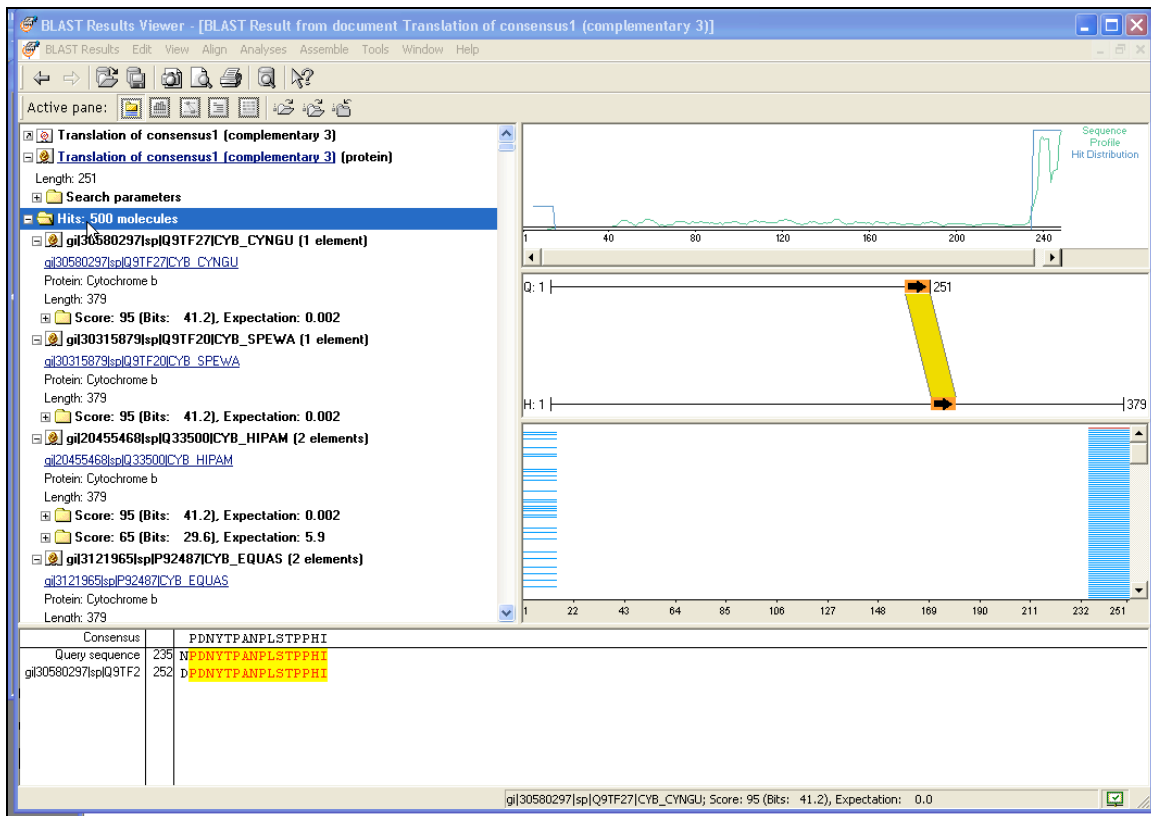
Make sure that you are searching using the `blastp` program, and that you are searching against the swissprot database and then press the **Submit** button:



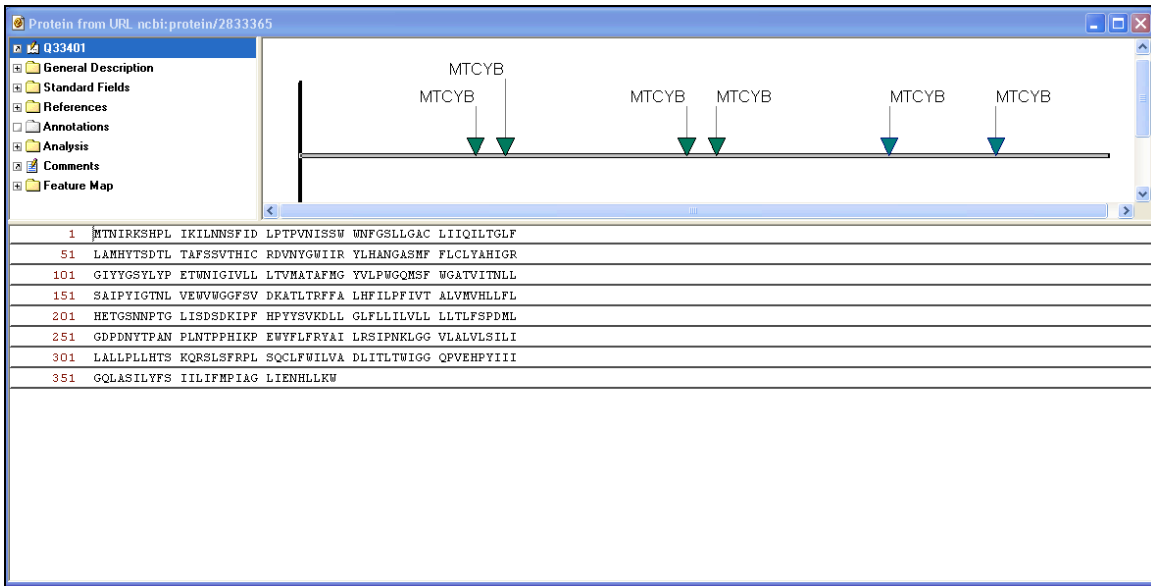
The status column tells you what the program is doing. After awhile, the BLAST search will conclude. Double-click on the name of the sequence to review the results of the BLAST search:



To see the hits from the BLAST search, expand the hits folder in the leftmost pane in the BLAST Results Viewer window:



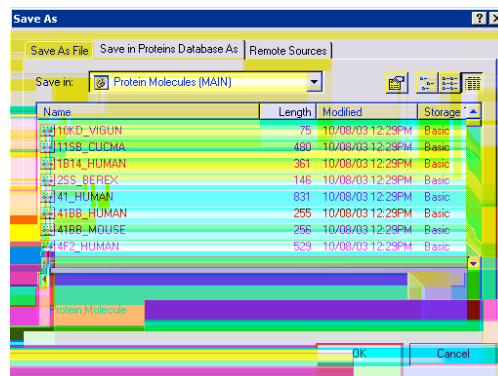
A set of scored results are shown in the upper-left window pane in the BLAST Results Viewer window. Click once on the blue, underlined link of the “Hit”. Doing this downloads the annotated SwissProt file for the molecule and displays it in the molecule viewer:



Open the Standard Fields folder and discover what species is the closest to your sample.

9 Saving your results

If your first hit was not a cytochrome B sequence, you have chosen the wrong reading frame. Try again until you find the correct reading frame. Once you are satisfied that you have the closest sequence to yours, save this sequence by selecting **File → Save As** then select the **Save in Proteins Database As** tab. Change the name of this protein to reflect your class period and name.



Similarly, save your Blast Results in the **Database BLAST Results** giving this file a name that reflects your class period and name. When you close Contig Express, it will ask if you wish to save your contigs, say yes, and save with a name that reflects your class period and name.

Log off of Windows using **Start: Shut Down**. When Windows has closed, click on the **Exit** button at the center bottom of the screen and click on **Log Out**. **Remember what computer you were working on!!!!!!**