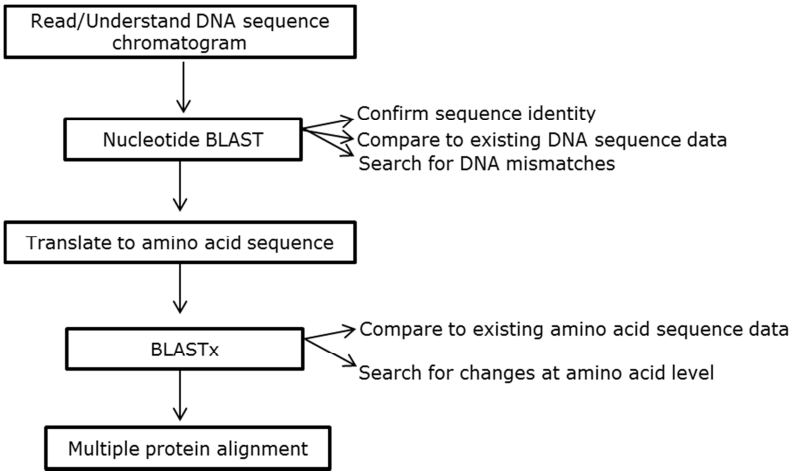
COMPUTER EXERCISE 2: SEQUENCE ANALYSIS

The overall aim of this exercise is to analyse sequence data obtained from sequencing of your cloned PCR product. Briefly, you will:



To begin this exercise, you will use Chromas Lite, which allows you to view chromatograms from automated sequencing runs. First of all, click on 'START' (bottom left of your computer), locate the P-drive and then click on the 'Chromas' file. Open the Chromas application and finally click 'RUN' to launch the program.

Now, navigate to the Molecular Methods Moodle page and under 'Computer Exercises', click on 'Computer Exercise 2'. Right click on the file called '**Cloned nrg product from icebox flies'** and 'save link as' in a folder of your choice.

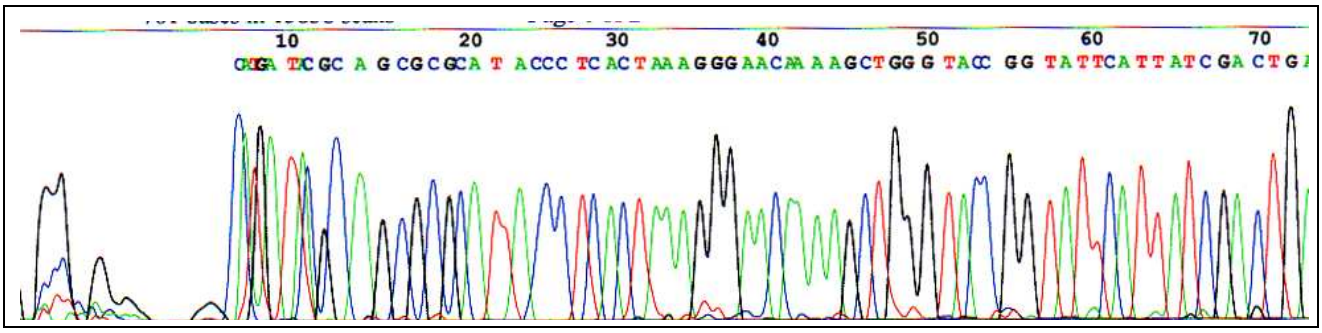
Reading chromatograms from automated sequencing

Return to Chromas Lite and then use **Open** to locate and open the chromatogram file you downloaded from our Moodle site. The chromatogram was obtained from a sequencing reaction run with DNA prepared in exactly the same way as you have done on the Molecular Methods course - a PCR product amplified with primers shown on the previous page was cloned into pBluescript using KpnI and BamHI. The sequencing reaction was run with the pBluescript **reverse sequencing primer** (see multiple cloning site DNA sequence for detail).

Q1. Which restriction site will mark the boundary between vector sequence and the start of the insert when sequencing from the reverse sequencing primer? Which of your cloning primers (used in your PCR) contain this restriction site?

Take a closer look at the chromatogram. Initially, the data has some "noise" - peaks that are broad and look unconvincing - but if you move downstream the signal should settle down to produce a series of sharp, evenly spaced peaks with low background. Locate the restriction site that marks the boundary between vector and insert. You can use the **FIND** tool to locate the sequence of the expected restriction site.

Q2. Note down the restriction site marking the vector/insert boundary in the diagram below:

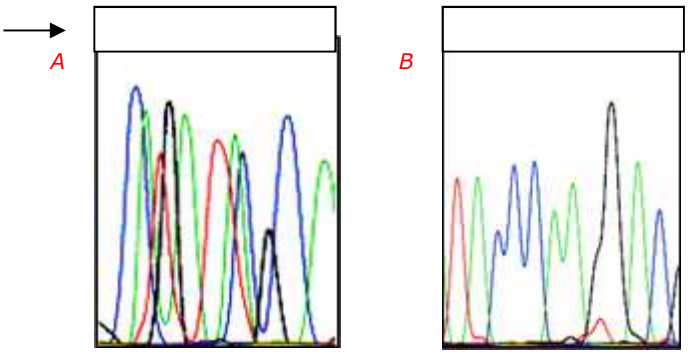


The first 20 bases or so of the insert should match the sequence of the relevant PCR primer used (previous section). The peaks should be clear and evenly spaced, matching the computer-generated sequence that lies above the trace.

Automated sequencing is powerful, but don't accept it uncritically. As you move through the sequence, does the sequence "called" by the computer tally with your interpretation of the data? Does it match what you would expect from the sequence of *Nrg* (you'll find this in Figure 2 of the PCR/Primer Design exercise)? You may find the occasional position where the spacing of peaks is too close for the computer to resolve satisfactorily as the example below. Equally, in regions that are GC rich, the peaks may merge together and the number of nucleotides may have to be inferred from the spacing, rather than the appearance of discrete peaks (see below). If you come across areas like this, your judgement might be better than the computer-called sequence!

For example:

Q3. Write down what you think should be the sequences below (green, A; red, T; blue, C; black, G)



Check your answers by looking at the region on the chromatogram around bases 1-15 for A and 200-210 for B.

Q4. Look carefully at the peaks from around 630 nucleotides. You will notice that the peaks on the chromatogram start to overlap and your faith in the accuracy of the sequence called by the computer may start to wane. From this initial inspection, you should be able to assess how much useful sequence data you have to work with. Subtract the position of the boundary between vector and insert. What are you left with?

Q5. We know that the insert should be 750 nucleotides long. Is it therefore likely that the sequencing reaction has covered the full length of the cloned *Nrg* product?

Q6. If the reaction has sequenced all of the insert and beyond, what features would you expect to see in the data at its furthest point?

Using BLAST searches on the Genbank database

With a bit of patience and an eye to detail, you should be able to confirm that the sequence data has probably come from the *Nrg* gene. The first 20 bases or so should match one of the primers you used for the PCR and beyond that, the data should run into the *Nrg* sequence (shown in Figure 2 of the PCR and Primer Design Exercise).

As you can appreciate, checking by eye is tedious and error-prone - quite often when you have sequenced a piece of DNA it may not be immediately apparent that it has come from the gene of interest or if the sequence is of value to your investigation. One way to assess this is to submit its sequence for comparison with many other known sequences at the GENBANK database. GENBANK is updated nightly, and anyone in the world can send a real or imagined sequence to be analysed for free, using the WWW.

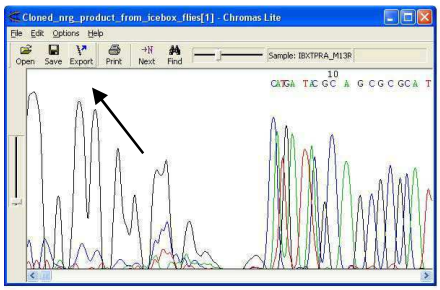
Go back to the web browser and head for the National Center for Biotechnology Information (NCBI) website:

- **http://www.ncbi.nlm.nih.gov/**
- When the page has loaded click on the BLAST button on the right hand side. BLAST stands for Basic Local Alignment Search Tool and it finds regions of similarity between biological sequences.

- Select the **"Nucleotide blast**". This sets up the software to receive a DNA (or RNA) sequence and compare it with other nucleic acid entries on the database. Under section "**Choose Search Set**" select **"others (nr etc)"** and at "**Program Selection**" click on "**highly similar sequences (megablast)".**

At this stage, you need to be able to enter the DNA sequence easily. You could (with patience) type in the data from the chromatogram, base by base but the accuracy will be better if we copy directly from the computer. To do this, select the **Export** button in Chromas Lite (arrowed in the picture below).

Select the folder on the computer where you are accumulating data from this exercise and enter an appropriate file name ("Raw sequence data" would describe it). The data - just the computer-called sequence, not the chromatogram - can then be saved in FASTA format.



Now open a simple text editing program - Notepad would be fine.

This can be found in the Start menu,

- "All programs" and then "Accessories"
- "Open", change "Files of type" from "Text documents (*.txt)" to "All files" or the software won't be able to see the FASTA file that you've just saved!

You should see a header field followed by the sequence taken from your chromatogram - just as bases, not coloured peaks. From your earlier analysis, you should be able to locate where vector sequence ends and the PCR product cloned into Bluescript begins.

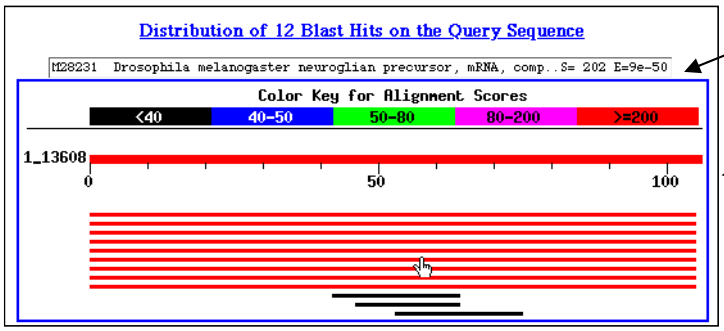
EDIT YOUR SEQUENCE

(1) Delete the vector sequence.

(2) Delete the last 3 or 4 lines of sequence to ensure that the data is of good quality for the next stage of the exercise.

- Now **save** the edited file back to your designated folder using "Save as". Choose a sensible name and save as a text file.
- Click and drag across the edited sequence file and copy. If you return to the NCBI BLAST page, you can now paste the data into the "Search" box. To compare your sequence against the standard database of all nucleotide sequences click the "**Blast!**" button. If you are lucky your results will come back very quickly. If the server is busy you will then get a message which says the page will be updated in X seconds. Use the time you are waiting to read ahead.

When you get your results, scroll down the page and you will see something like this:



This box displays the identity of any sequence match. Simply place the mouse pointer over any of the sequence matches below.

Thick red bar – linear representation of your submitted sequence.

Sequence matches: coloured lines show extent and significance of matches to your sequence found in the database

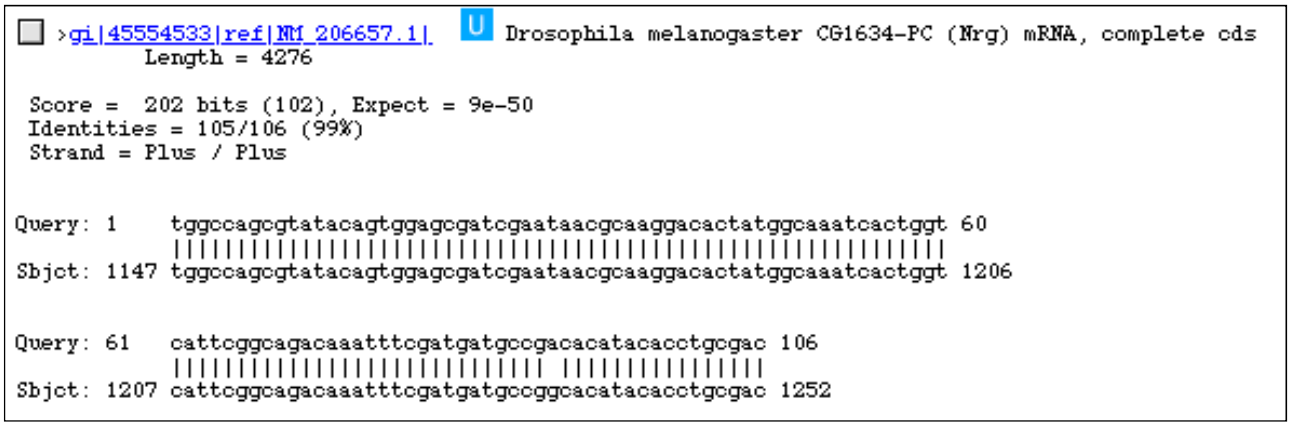
The red bar with numbers in the middle of the frame shown above represents your submitted sequence (we submitted 106 bases of sequence in this example). Below this are a series of lines of varying length. These represent sequences in the database which match your submitted sequence. The extent of the line tells you which part of your sequence is matched by the database sequence, and the colour of the line tells you how good the match actually is. The scale (as shown under "Color Key for Alignment Scores") runs from black, which means "pretty awful", to red which means "extremely good match". The scoring has two main components: the length of the matching sequence and the number of mismatches in the aligned part of the sequence. Thus a sequence of 300 bases which matches your submitted sequence with 10 mismatches between the sequences would score more highly than a 300 base match with 30 mismatches. Both of these would score higher than a perfectly matched alignment of only 20 bases in length.

If you hold the mouse pointer over a match line (as in the example above), the name of the matching sequence will appear in the box above the Color Key; in our example the match is to the "*D. melanogaster* neuroglian precursor

mRNA". Moving the mouse pointer across some of the other matches you will see displayed the identity of the other matching sequences. Further down the page you will see the names of all the matching sequences, together with the probability of the match being a mere coincidence.

Q7. Are you satisfied that the sequence you have gathered has actually come from the Drosophila *Nrg* gene?

Scroll down even further and you'll see a section that shows the actual alignment for each match, as in the following example:



Here you will see a "query" sequence (in this example, a neuroglian mutant called *Nrg3*, which is a temperature sensitive recessive lethal) aligned with the matching "subject" sequence found in the database (in this case, the *Nrg* mRNA, as it says at the top). In the example shown here, the alignment is between nucleotides 1 to 106 of our query sequence and 1147 to 1252 of the neuroglian mRNA. The two sequences are virtually identical but you will see that there is a mismatch between the two sequences, a G > A change at position 1236 in the neuroglian mRNA. Perhaps this nucleotide change could be responsible for the *nrg3* mutant phenotype. We shall pursue a similar analysis to try and see if we can identify mutations that might explain the behaviour of *icebox* flies.

If you scroll down you will come to neuroglian sequences named 'transcript variants'. Choose one of the transcript variants A, B or C to see if there are any mismatches which might be the *icebox* mutation. The other sequences represent, for example, genomic DNA whereas the transcript variants show the mRNA sequence.

Note the position of any nucleotide change(s) together with a six to eight bases on either side so that you can locate this region later. Annotate your notes with the sequence of the "subject" sequence at these points of mismatch.

Q8. Just because there is a DNA sequence change does not mean that the amino acid sequence will change – why?

Translation of DNA sequence

ExPASy is one of many useful sites where analysis tools for molecular biology are freely available. With this **Translate** software, you can paste DNA sequence into the open box and the software will generate all possible protein coding sequences for you.

Set up a new browser window, and go to

- **http://www.expasy.ch/tools/dna.html** (you can go to this web site from the Molecular Methods Moodle site – Computer Exercise section)
- Go back to Notepad (or whatever text editor you have been using) and open up the *edited* sequence file (*not the original, unprocessed data!*). Click, drag and copy the data.
- Paste in the data from the file with our sequenced clone DNA sequence. From the drop-down menu beneath your pasted sequence, select **Includes nucleotide sequence**. Now click **TRANSLATE SEQUENCE**.

The software takes the DNA sequence and translates it in all forward and reverse reading frames, producing aligned protein and nucleic acid sequences. **'Stop' or '–'** marks potential stop codons for each of the reading frames. Remember that our PCR product covers part of an intron as well as exon 3. The natural *start* codon will not therefore be present and stop codon(s) might be present in sequences from the intron. Nevertheless, the most likely reading frame should be free of stop codons.

Q9. What is meant by the term 'reading frame'?

Q10. Which reading frame looks most likely to encode the Nrg protein from the icebox flies?

Translated Blast Searches

The BLAST software at NCBI has been adapted for a range of jobs. We have made DNA/DNA nucleotide comparisons using "**blastn**". Another useful search tool, "**blastx**", assumes that your sequence encodes a peptide, and it compares all 6 possible reading frames of your sequence with the deduced peptides of all known sequences in the database. We now know that we have a potential reading from in the sequence data - we can now use blastx to see if there are any amino acid changes as a result of nucleotide substitutions.

The different 'flavours' of blast. Each has a use – ask you demonstrator if you would like to know more.	
blastn	Search a nucleotide database using a nucleotide query
blastp	Search protein database using a protein query
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

To carry out a "**blastx**" search:

- go back to the BLAST front page and this time select "**blastx**".
- Submit your *icebox* mutant DNA sequence as before; your results page will be colour coded, similar to the results from blastn, but note that even a black–coloured match from blastx does indicate a fairly good relationship.

Here is our result with the *Nrg3* mutant sequence, the example that we used before:

```
>gi|24640619|ref|NP_727274.1|  CG1634-PB [Drosophila melanogaster]
gi|14286138|sp|P20241|NRG_DROME   Neuroglian precursor
gi|22831957|gb|AAN09236.1|   CG1634-PB [Drosophila melanogaster]
          Length = 1302

 Score = 76.6 bits (187), Expect = 2e-13
 Identities = 34/35 (97%), Positives = 34/35 (97%)
 Frame = +2

Query: 2   GQRIQWSDRITQGHYGKSLVIRQTNFDDADTYTCD 106
           GQRIQWSDRITQGHYGKSLVIRQTNFDDA TYTCD
Sbjct: 284 GQRIQWSDRITQGHYGKSLVIRQTNFDDAGTYTCD 318
```

The *Nrg3* mutant sequence has a change of G to D at amino acid 313. Looking at the DNA sequence data from blastn a couple of pages back in the manual and using the genetic code table on the last page of the manual, we can tell that this change arose from GGC (encoding glycine) changing to GAC (aspartic acid). This alters the smallest possible amino acid side chain (–H) for a larger, acidic side chain ($-CH_2COOH$), and therefore is a candidate for causing the *nrg3* phenotype. We can write it using standard notation as: G313D (GGC>GAC).

Look at your own blastx results. You may see your sequence aligned with several different versions of the *Drosophila melanogaster* neuroglian sequence – these are sequences submitted to the database by different labs, and you may find that there are differences between the sequences from the database, even though they are all from wild-type flies.

Q11. Suggest two reasons why there might be differences between the neuroglian sequences from the database.

Q12. What is the common domain that is present in all of the proteins discovered in your BLAST search (Nrg/hemolin/L1CAM/neurofascin)?

Focus on any differences between your sequence and the database subjects that are consistent for all of the *D. melanogaster* hits.

Q13. Which amino acid changes might be responsible for the icebox phenotype? Write down the change(s), their position(s) in the protein sequence and a few residues at either side to aid location of the altered amino acid(s).

Multiple Protein Alignment

Multiple protein alignments are important tools in studying proteins. The basic information they provide is identification of conserved sequence regions. This is very useful in designing experiments to test and modify the function of specific proteins, in predicting the function and structure of proteins and in identifying new members of protein families. In order to help us do this, we can export the protein files and do a protein alignment using software called ClustalX.

Once you have the sequence of the same protein from several species, there are several ways to get information from this. One of the most straightforward is a multiple sequence alignment. To understand how this works, think about American versus English spellings:

E: Note the catalogue number, your favourite colour, and send a cheque with your order.
A: Note the catalog number, your favorite color, and send a check with your order.

These sentences can be aligned, so that the equivalent bits coincide, by artificially introducing gaps (-)

E: Note the catalogue number, your favourite colour, and send a cheque with your order.
A: Note the catalog-- number, your favo--rite colo-r, and send a check-- with your order.

Thus we can align functionally equivalent bits of protein sequence, even though evolution has introduced many changes in sequence composition and length.

We have downloaded and aligned the following sequences:

Cloned Nrg	Nrg sequence that you cloned in the lab
Drosophila WT Nrg	Nrg protein from wild type *Drosophila melanogaster*
Anopheles Nrg	Nrg from *Anopheles gambiae* (mosquito)
Moth Hemolin	Hemolin from *Pseudoplusia includes* (moth)
C. elegans L1CAM	L1CAM (a gene of the same family) from *Caenorhabditis elegans* (worm)

- Go back to your Moodle page and right click on '**Nrg alignment file'** and save to a folder of your choice.
- You should have **ClustalX2** installed on your computer. Locate the P-drive as before, then open the ClustalX2 file. Click on the Clustal application and then 'run' to launch the programme.
- In ClustalX go to **FILE** then open the saved aligned file (**Nrg sequences of Nrg.aln**)

Q14. Can you identify the 2 mutations that are consistently different between your cloned icebox Nrg sequence and wild type Nrg from Drosophila, Mosquito and Moth? Does this provide you with any information which might be useful in deciding if the mutation you identified might be responsible for the *icebox* mutant phenotype?

Q15. Why is one amino acid mutation highlighted in the same blue colour? What does this represent?

Q16. L1CAM is a human gene of the same gene family as *Drosophila* neuroglian. Mutations in this gene lead to what disease phenotype? (To find out Google L1CAM)

It is relatively quick and easy to make *Drosophila* transgenic. How could you establish beyond doubt that a mutation you have identified, and not one of the other changes, is actually responsible for the icebox phenotype?