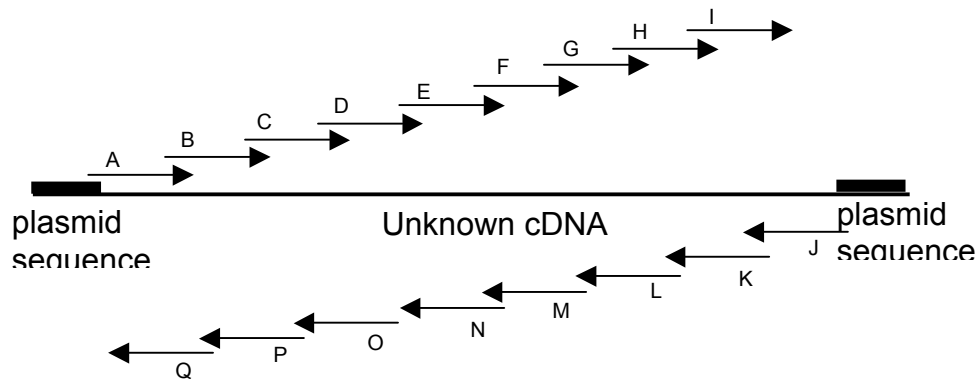


Computer-based analysis of sequence data – the unknown cDNA

Russell Johnson

week #10 (22 April)

For this week's lab, you will work with an "unknown cDNA" isolated from human cells. Your mission will be to produce a complete and correct sequence for the cDNA, to predict the amino acid sequence of its encoded protein, and to carry out some further analyses. The "unknown cDNA", which is about 5000 bp long, has already been placed into a plasmid vector and has been sequenced via primer walking, based on the scheme shown below.



As you can see, sequencing reaction A was carried out using a primer that matched the known plasmid sequence. Another primer was then designed to match some bases near the end of the A sequence, and reaction B was carried out. This was continued until the entire cDNA was covered. You will note that the opposite strand of the "unknown cDNA" was also sequenced, using a series of reactions going in the opposite direction. This is common practice for the sequencing of cDNAs and genes, as it provides a good way to ensure accuracy. The forward and reverse sequences should turn out to be perfectly complementary to each other. If there prove to be any discrepancies, the error in one of the sequences can be identified and corrected.

The results of the sequencing reactions (A through Q) can be found in the BC378 folder on the desktop of the computers in Olin 213. Inside the folder entitled unknown cDNA are individual text files from each of the sequencing reactions. Take a look at the data from the individual sequencing reactions. *From this information can you make a prediction about which strand of the cDNA is the coding strand? How?*

1. Compiling the cDNA sequence. Starting with the individual sequence files, you will need to compile these together to obtain the complete sequence of the entire cDNA. This could be done manually, but it would be extremely tedious. Fortunately, there are computer programs that can do this for you. To align your sequences, you will use a program called CodonCode Aligner.

In the Applications folder of the computer, find the CodonCode Aligner folder. Inside this folder you will find the CodonCode Aligner application. Open the application by double-clicking on it. Once the CodonCode Aligner is open, choose **File → New Project**. Click on the Add Folder button. In the window that pops up, choose the unknown cDNA folder to be added.

When you have finished, all of the sequence files (A through Q) should now appear in the list of Unassembled Samples.

Select the Unassembled Samples folder and click on the Assemble button. The computer will now (very rapidly) compare all of your individual sequences and determine which ones overlap with each other. It will join together each set of contiguous fragments into a “contig”. In your situation, if all goes well, you should expect all of your sequences to be compiled into a single contig. If that is the case, a folder entitled Contig1 will appear in the CodonCode Aligner window. Select the Contig1 folder and choose **View → Contig**. You should now be able to see how the individual sequence fragments have been compiled together. At the top is a graphical representation of each individual sequence fragment and how it fits into the whole. In the middle section, each individual sequence is shown, and you can see the regions of overlap. At the bottom of the screen, the complete compiled sequence is shown.

If all has gone well, you should now have data for every position in the sequence from both directions. *Do you in fact have complete coverage of your cDNA in both directions?* You should now compare to make sure that the two sets of data agree with each other. A very easy way to do this is to click on the Mark Matches button. This will change all of the correctly matching bases into dots. Once you have done this, it is easy to scroll through your entire sequence and see if any bases are still displayed as letters. If so, this would indicate a discrepancy that you would need to resolve. *Do the sequencing data from the two strands of your cDNA match perfectly?*

If you feel confident that you now have a correct and complete sequence for the “unknown cDNA”, you can now take this sequence and work with it. Choose **Edit → Select All** and then **Edit → Copy(selected sequence)**. You will now be able to paste your sequence anywhere you like (into Word, Gene Inspector etc.).

2. Nucleotide sequence analysis. Once you have obtained your complete cDNA sequence, it will be helpful to compare this nucleotide sequence to other previously identified sequences in the database by using the BLAST program. You should also determine the open reading frames that are present in your cDNA and use this information to predict the amino acid sequence of the protein encoded by the cDNA. *Include the information you have obtained here in your notebook.*

3. Peptide sequence analysis. Once you have determined the amino acid sequence of your protein, you can use this to predict the function of your protein. It may be helpful to compare your peptide sequence to previously characterized proteins in the database by using the BLAST program. To compare protein sequences in BLAST you will need to use the blastp program instead of blastn. *Include the information you have obtained here in your notebook.*

4. Design of primers for a qRT-PCR experiment. As part of your future research, you might like to measure the expression patterns for the gene corresponding to your “unknown cDNA”. Of course, a fine way to measure gene expression is by using qRT-PCR. To carry out qRT-PCR on your mRNA of interest, you will need a pair of primers that will specifically amplify a small fragment of your cDNA. Designing successful primers can sometimes be an art as much as a science, but there are a few guidelines that should generally be followed:

- For most purposes, primers about 20 bases long are best
- the G+C content should be close to 50%

- the 3' end of the primer should be rich in A+T
- there should be no self-complementary regions (that could allow hairpin formation)

For a PAIR of primers to work well together in PCR they should:

- have similar melting temperature
- NOT have sequences that are complementary to each other
(this could result in the formation of primer dimers)

When designing a pair of primers for PCR, one also has to take into account the length of the PCR product that will result. For standard PCR, any length from 100 bp up to 2000 bp can usually be amplified successfully. In general, longer products tend to be more difficult to produce. For qRT-PCR, a PCR product of about 80 – 150 bp will give the best results.

While it is certainly possible to design primers manually, it is much more common to use primer design software for this purpose. One very convenient program, which is available free on the internet, is called Primer3. Anyone can use this software just by going to the website (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi)

Your job here will be to design a pair of primers that you could use in a qRT-PCR experiment to measure the abundance of your “unknown” mRNA. *Include the relevant information about the primers you chose, and the PCR product that would be generated, in your notebook.*

4. Protein information. It may be that some useful information is already known about the protein encoded by your cDNA. There are very useful databases available that can help you to find this information. You can find some information (eg. publications about your protein) by using the NCBI web site. Since you are studying a human protein today, you can take advantage of the Human Protein Reference Database (www.hprd.org). Similar databases also exist for proteins from other species.

To fully understand the biological role of a protein, it is often helpful to know what other proteins it interacts with. *Find out what other proteins have been found experimentally to interact with your protein. How do each of these interacting proteins relate to the biological activity of your protein?*

Find three recent papers that have been published about your protein. Read (at least) the abstracts of these three papers to find out what research has been done on your protein. *Include in your notebook the citation information for each paper and a short (2-3 sentences) summary of the findings from each paper. What are two important questions about this protein that need to be answered in the future? What experiments would need to be done to answer these questions?*