

## Computer-based analysis of sequence data – the *COT1* cDNA

Russell Johnson

week #9 (15 April)

Over the next two weeks you will learn how to use a variety of computer software tools for the analysis of nucleotide and peptide sequences.

Fairly detailed instructions are provided for this week's activities to help you get accustomed to the sequence analysis programs. For next week, there are not as many instructions and you will be expected to work more independently, using what you have learned this week and investigating on your own any additional information that you may need.

## Computer-based analysis of sequence data

### Introduction

In recent years great strides have been made in the development of faster and faster techniques for sequencing genes and entire genomes. The mountains of sequence data obtained from large-scale sequencing projects would be of little value, however, without the ability to quickly and accurately analyze this information. The development of efficient and user-friendly computer software for the analysis of DNA, RNA and protein sequences has provided molecular biologists with the ability to obtain important information about the genes they are studying. Computer programs are now available that can search DNA sequences for open readings frames, predict amino acid sequences from DNA sequences, search large sequence databases to identify similarities between genes, predict secondary structure of RNA and protein molecules based on their sequences, and many other tasks that are of great help to molecular biologists.

We will be using two different programs in lab today. One program, called BLAST (Basic Local Alignment Search Tool) is installed on the computers of the NCBI (National Institute for Biotechnology Information) in Bethesda, MD. To use this program all we have to do is to send a DNA sequence electronically to the NCBI and the BLAST program will compare this sequence with all of the sequences in a huge database and report back to us if there are any known sequences similar to ours. The second program, Gene Inspector, is a software package designed for the Macintosh that is able to do a wide variety of useful analyses on any sequence that we are interested in.

### The *COT1* cDNA

The famous mock scientist Elena Nito del Bosque has been toiling in the lab day and night for several weeks and has finally succeeded in isolating and sequencing a mock cDNA for a gene that is highly expressed in muscle cells of mock turtles that have been subjected to cold temperature stress. She has named this mock gene the *COT1* (cold tolerance) gene. By doing some RNA blotting experiments she has already confirmed that the mRNA from this gene is present in very low quantities in the muscle cells of mock turtles that are living in a warm (30°) environment. When the mock turtles are transferred to a cold environment (10°) the *COT1* transcript becomes very abundant within a few hours. Now that she has obtained the sequence of a *COT1* cDNA she needs your help to figure out what sort of protein is encoded by the *COT1*

gene and what sort of role it might be playing in helping the animals to tolerate cold temperatures.

The sequence of the *COT1* cDNA, according to the sequencing she has done, is as follows.

```
GAATTCGGCGCCGCTCTCCACTCCACTCCACACCAAAACGGGAAGAATCAGTTCCTTCCGTTTCTTGAGATCTCAGCCAAATGGAACACTGTT
ATCTCAGCAGAAATTTCCGCTCCTAGTTGCAGCATCGCCGGACTGAGAGATATAGTATAACAAAGGCTGTAGTTCAGAGCTTCATCAACATGTCACCTC
CTTTCAAATACCGTTGGTCACTGTGGATGTGACTGGCACTCTTATTGCTTACAAAGCCGGCTTGGTGATTACTACTGTATGGCAGCTAAGTCTGCTG
GAAAGCCATGCCCGACTATAACCGAATGCACGAAGGCTCAAGCTTGACATCACTGAGATGGCTAGAAAAATACCCATGCTTTGGGTTTGGCGCAAAGAT
GCCAACGATCGAATGGTGGCGGATCTGTGTCAAGGATTCGTTTGTAGGGCTGGCTATGATTATGACGATGAGACATTTGACAAAGCTTCAAACGTATT
TATTCTGCCTTTGGCTCCCTGCACCATACTCGGTGTTTCCTGATGCACACTCGAGCTGCATTTCAAGCTGTGTGAGCCGGCGGGTCCGAGTTTCAGAGA
GCTCAAGCACCAGTATAAGAAAAGGTGTCACTATGAGGAGAATCGAGGCCATCATGGGCCAGTAAGAGATTCCGACATTTGTTTAAATGAGTTTGTGAG
GGGCCAAGGAGGGGACACACATCCGCTATCACCTTCTGTCTTGGGCACCTTCGGTGTAGTCAATGTGAGAAATGTCTTGTTCAGAGCAACCTGAGC
AAACACGAGGGAATTCATGTCAGAAAGATCTGGTGTCTGAGTGTGGCAATCCCTGGGCAGCAAAATCAACATTAGCAAGCACGAGCGAATTCACGGGG
AGGGCGGGCATGGGGCCAGTGTGCACACCACAGTGGTCAGCATTTGTGTCAGTAACACGTCACAGCATGCACACGTGGCAGCAGCATCGATGCGACGT
GACGACGAGTGACACAGCAGTCATAAAAAAAAAAAAAAAAAAAAAAAAAA
```

To prevent you from having to type the whole sequence into the computer to do your analyses, she has already entered it for you. It is in the file entitled cot1cDNA and can be found on the desktop of your computer.



A mock turtle shown with Alice and the Gryphon

## Comparison of the *COT1* cDNA sequence to previously identified sequences using BLAST.

**1. Sending a search request.** It is standard practice to compare any newly acquired sequence to the sequences of genes that have already been identified. This will tell you whether or not someone else has already identified a gene that is very similar to yours. If so, you may be able to very easily determine what the biological role of your gene is. In our case, we might find that other researchers have already isolated a gene similar to the *COT1* gene from another animal species. If they have already determined the role of the protein encoded by this other gene, this may provide us with very useful clues about the role of the *COT1* protein.

The analysis can be done quite simply by sending your sequence to NCBI (through the BLAST website) to be compared to their DNA databases using the BLAST program.

The address for the BLAST website is <http://www.ncbi.nlm.nih.gov>.

**2. Interpreting the results of your search.** When the BLAST search is finished, you will receive a list of sequence segments from the database that have a reasonable amount of similarity to some part of your sequence. The name of each sequence will be listed on the left and to the right of this a “score” will be listed. A higher score indicates a greater percentage of identical bases between the two segments or a longer stretch of similarity. The E value is an indication of how likely you would be to find a sequence in the database with this much similarity purely by chance. A very low E value will thus indicate a high probability that two

sequences are actually related. Once you receive the results of your query, look it over closely to see *what sequences from the database, if any, are similar to your sequence*. If so, *do any of these genes or cDNAs have known functions that can help you to determine what the COT1 protein might do?*

### **Analysis of the COT1 cDNA using Gene Inspector.**

**1. Getting started.** Before you can do analyses with the Gene Inspector program you will need to create both an electronic “notebook” where you will store the results of your analyses and a sequence file for the COT1 cDNA. Start by opening up the Gene Inspector program. Choose **File** → **New** from the menus at the top of the screen. Create a new notebook, which can be named “Joe’s COT1 notebook” or whatever other name you like. Put an appropriate heading at the top of your notebook by choosing **Notebook** → **Tools** and the text icon, clicking at the top of the notebook page, and typing in the text of your heading. Normal text (not headings) can be put into your notebook by simply typing onto the page.

Choose **File** → **New** again and create a new nucleic acid sequence file called “COT1 cDNA”. Choose **Sequence** → **New Sequence** from the menus and type in an appropriate sequence name. Now you are ready to put the sequence into the file. Note that if you now start typing the sequence in, the computer’s voice will repeat back to you each of the bases as you enter them (if **Sequence** → **Speak Typing** is selected). This allows for much more accurate entry of sequence data. However, since you already have the COT1 cDNA sequence in a Word file on your desktop, all you have to do is copy it and paste it into your Gene Inspector sequence file.

It is highly recommended to Save your notebook and sequence files frequently as you continue to make changes to them.

**2. Analysis of open reading frames.** The first thing to do is to check to see what (if any) open reading frames exist in your cDNA. If you do not find an open reading frame, it will mean one of two things; either your cDNA does not really code for any protein or you made a mistake in your sequencing. If you do find an open reading frame then this will help you to determine what protein is encoded by your cDNA. Open up your COT1 cDNA sequence file and choose **Analysis** → **New Analysis**. On the left side of the window that appears choose Open reading frames and then click the OK button. In the Analysis Setup window you will have to make several choices about how to do the analyses. In the Method box choose Start and Stop codons. In the Display box choose only ORFs. Click on Browse Tables and choose the Xenopus laevis codon table, since this is the organism on the list that is most closely related to mock turtles. On the left side of the Analysis Setup window click on input sequences and confirm that the sequence chosen for analysis is indeed the COT1 cDNA. Then click on output location to confirm that the results will be sent to the correct notebook. Once everything is in order, click on the RUN button.

Once the analysis is completed, you should be able to find the results at the beginning of your notebook. By clicking on the object once, you will be able to move it around and adjust its size and shape. By clicking on the object twice, you will be able to access individual parts of the object separately.

The arrows in the figure indicate where open coding regions are found in each of the six possible reading frames of your sequence. *Which one do you think is the one that actually encodes the COT1 protein? Why?* Once you have decided where the COT1 coding region is located, you are ready to proceed to the next step.

**3. Determining the protein sequence.** Go to your sequence file and mark your presumed start and stop codons. You can do this by making them bold, underlined, or a different color. Select the entire *COT1* cDNA sequence (including notations, numbers etc.) and **Copy** and **Paste** this sequence into your notebook. By dragging one of the corners outward you can have your sequence displayed with 60 (or 70) bases per line instead of only 40. You may also need to expand the size of the object's box at the bottom so that all of the sequence will fit into it. Now double-click on the sequence "object" so that you can manipulate individual aspects of it. Select the coding region of the cDNA and choose **Features → Translate**. Make sure that the *Xenopus* translation table is being used and click on OK. Choose **Features → Display → One Letter AA Code**. Choose **Features → Display → Line Spacing** and instruct the computer to put 7 pixels of extra space between each of the lines. Choose **Features → Adjust Size to Contents**. Now you should have (in your notebook) a nice printout of your cDNA sequence together with the deduced protein sequence. You can of course manipulate the display in your notebook for easier viewing by making the coding region blue, the protein sequence bold, etc., etc.

*note: If you have found the correct amino acid sequence for the COT 1 protein, some of the letters in the sequence will spell out a special "cold tolerance motif".*

**4. Other nucleic acid analyses (Dot Matrix etc.).** There are quite a few additional nucleotide analyses that Gene Inspector can do on your sequence. One type of analysis that is often quite useful is the Dot Matrix analysis, which allows you to compare your sequence to another sequence to see where regions of similarity are located. Choose **Analysis → New Analysis**, click the button for Nucleic Acid Analysis, and on the left side of the box choose Dot Matrix. Click on the OK button. In the Analysis Setup box (with Dot Matrix selected in the left panel), click on the Table selection button and under Standard tables, choose Nucleotide Identity. Type 7 into the Threshold value box and click on Add Threshold. With these threshold parameters, you will get a black dot if at least 7 out of 10 bases are identical when comparing the two sequences.

Click on Input Sequence in the left panel. The Chosen Sequence box will display what sequences are to be compared. You will want one of them to be the *COT1* cDNA and the other to be the sequence that you wish to compare it to. Click on the Add button to add a sequence to the window. Then find the sequence file(s) that you want to use. Then click on the actual sequence that you want to use and hit the Add button at the bottom. When you are finished, click on the Done button. Now you should have the two desired sequences listed in the Chosen Sequence box. Now click on Output Location in the left box to make sure that the output will be going to your notebook. When you are all ready to run the analyses, hit Run. Manipulate the "object" that appears in your notebook as desired to make it look nice and easily readable. *What information does the Dot Matrix give you about the similarity between the COT1 cDNA and the other sequence that you compared it to?*

In addition to making comparisons between two different sequences, it can also be informative to compare a sequence to itself using a Dot Matrix analysis. This will tell you whether there are repeated elements in your sequence. *Do a Dot Matrix comparing the COT1 cDNA to itself and put the results into your notebook. What information can you obtain from this analysis?*

Other analyses that you might wish to carry out on your cDNA sequence include Restriction enzyme digest, Codon preference, Find repeats, etc. *Carry out at least one of these other analyses on your cDNA sequence and put the results into your notebook. What information can you obtain from this analysis?*

**5. Protein analyses.** The first thing that you will need to do is to make a protein sequence file for the COT1 protein. In your *COT1* cDNA sequence file, go to the *COT1* cDNA sequence and select only the open reading frame. Choose **Sequence → Manipulate → Translate**. Make sure that the correct translation table is being used and click OK. A new (untitled) peptide sequence file will be created with the COT1 peptide sequence. Save this file with an appropriate name.

It would be helpful to be able to predict some of the characteristics of the secondary and tertiary structure of a protein from its amino acid sequence. This could tell us some useful information about what functions the protein might perform. There are a number of methods for doing this, but it is important to remember that they are not completely accurate. There are so many complicating factors that affect the way a protein folds (some of which are not completely understood) that it is difficult to predict it with a high degree of certainty. This is especially true when using the relatively simple folding prediction software included in Gene Inspector. You should take the results of any such prediction with a grain of salt.

Do a GOR Structure Prediction on your COR1 peptide. The output of this analysis will be a graph indicating which portions of the peptide are likely to form alpha helices, beta sheets, etc. In order to look at a representation of how the peptide might actually look, double-click on the graph and choose **Object → View As Squiggles**.

Do a Prosite motif search and at least two other analyses on the COT1 peptide. *What useful information did you find by doing these analyses?*

*Taking all of this information into account, help Elena answer the questions posed at the top of p 29?*