Bioinformatics: Molecular Computational Tools (Module II)

Introduction: Computers have become indispensable tools for the modern molecular biologist. A number of computational programs are available that help search through genome databases, design primers for PCR, translate a DNA sequence into an amino acid sequence (and vice versa), identify conserved regions in the amino acid sequence (domains), etc. In this lab, you will see how genome databases are organized, annotated, searched, and used by a molecular biologist. You will also gain some familiarity with a range of other computational tools at the disposal of molecular biologists.

Sequence databases: When a researcher determines the sequence of a gene or protein, or the three-dimensional structure of a protein, the information is deposited in a database that serves as a resource to investigators around the world. Examples include GENBANK (nucleotide and protein sequence; NCBI), and SWISS-PROT (protein sequence; University of Geneva). All of these resources are accessible to scientists (including you) via the Internet.

Using the available databases and sequence analysis software, one can obtain a vast amount of information about a gene's structure, its encoded message, and the putative function of its encoded protein.

This lab is the beginning of a 13-week project in which you will engineer a construct to knockout a gene from the ciliate *Tetrahymena thermophila*. The first step is to identify an appropriate gene homolog in *Tetrahymena* using the databases. You will then design primers to amplify the gene by PCR and clone it into a plasmid vector. This lab will take you through the steps to do that, and will demonstrate the use of other bioinformatics tools.

Procedure:

I. Find the amino acid sequence of your protein.

a. Go to the NCBI homepage: http://www.ncbi.nlm.nih.gov/
b. At the top of the page search using only the protein database for your selected protein. (For example: XPC). You may want to also add the genus of the organism that your articles were studying in order to narrow the search criteria (For example: Homo XPC).

c. Click on the sequence that best describes your protein.
d. **Copy this report into a Word document to put in your bioinformatics report. Save it as follows: (Genus species GENE NAME; i.e. *Homo sapiens* XPC).**

II. Determine if there is a *Tetrahymena* homolog.

a. Go to the *Tetrahymena* Genome Database: http://www.ciliate.org/
b. Click on the BLAST button on the left menu.
c. In the "Query comment" box put the name of the protein you are using in the search (For example: XPC).
d. Next paste in the protein sequence found in step Id.
e. Choose the Search program that will search the protein against the 6 frame translated nucleotide database.
f. In the "Datasets" box select the "T. thermophila Mac Genome." Press the "START SEARCH" button, and wait for the program to display the alignment
g. **In the Word file from step Id write down the TGD gene number for your top three hits (if you have that many) including the e-values (For example: 1. TTHERM_00037210, 8.7e-45; etc).**
h. **Write your top 3 hits on the white board next to your protein name.**
i. Now paste the TTHERM_# into the quick search field (located at very top of TGD web page) for each of your top three hits.
j. For each one scroll down to the section marked homologs
k. **For each of the THERM_# listed in Step IIh write down the IPI (Human) and SGD protein homologs and their e-values in your Word file.**

III. Obtain the protein sequence of the *Tetrahymena* homolog.

a. Go to the *Tetrahymena* Genome Database: http://www.ciliate.org/
b. In the "QUICK SEARCH" box, type in the TGD gene name for your top hit (lowest e-value) from the BLAST search in step IIh and press enter.
c. Scroll down this page to the bottom where it says "retrieve sequences."

d. Click on the box marked "coding sequence" and change to "ORF TRANSLATION" and click on the "VIEW" button.
e. **Highlight the protein sequence and paste it into a Word document. Label it "Tt GENE NAME amino acid" as a header above the pasted sequence. YOU WILL BE USING THIS PROTEIN SEQUENCE IN LATER CLASSES SO MAKE SURE YOU SAVE IT!!**

IV. Obtain the nucleotide sequence for your protein homolog.

a. Go back to the page that describes the *Tetrahymena* homolog that had the best match (lowest e-value).
b. Scroll down to the section that says retrieve sequences and now click on view for the CODING SEQUENCE.
c. **Highlight the nucleotide sequence and paste it into a Word document. Label it "Tt GENE NAME CDS" as a header above the pasted sequence. Save this as (Tt GENE NAME nucleotide)**
d. Go back to the page that describes the *Tetrahymena* homolog that had the best match (lowest e-value).
e. Scroll down to the section labeled "GENOME BROWSER" and click on the image of the gene next to it.
f. Go to the section marked "DUMPS, SEARCHES, and OTHER OPERATIONS:" and click on the button that says "Annotate Restriction Sites" and change to "Dump Decorated FASTA File."
g. Then click on the "CONFIGURE" Button.
h. If your gene is on the opposite strand (look at direction arrow is pointing on "GENOME BROWSER" page; opposite strand points left) click on the "FLIP" button.
i. Then under the "PROTEIN GENE PREDICTION" heading click on the "FONT" button (this will make the coding sequence RED and the noncoding sequence will remain BLACK). Then click on the "GO" button.
j. **Highlight the sequence and paste it into the document that had the coding sequence. Label it "Tt GENE NAME Genomic Seq." as a header above the pasted sequence. Resave the file.**
k. Go back to the "GENOME BROWSER" page and now click on "SCROLL/ZOOM" button and move it up enough to see about

1500 base pairs on each side of the gene. The image should now show more sequence on both sides of the gene.

l. Click on the "CONFIGURE" button. Make sure "PROTEIN GENE PREDICTION" is selected on "FONT". Also, click "TETRAHYMENA ESTs" to be underlined (2 types). Then click on the GO button (make sure sequence is flipped if needed). It should now give you the coding sequence of your gene along with sequence upstream and downstream and underlined is the Expressed Sequence Tags (ESTs) found for your gene.

m. **Highlight the sequence and paste it into the document that has the coding and genomic sequences. Label it "Tt GENE NAME with NTS with underlined ESTs" as a header above the pasted sequence. Resave this file.**

V. Obtain the Expressed Sequence Tags (ESTs) for your protein homolog.

(If you do not have any listed for your gene, ask the instructor for a sample gene to obtain the EST sequences for this section).

a. Collect the sequence of the ESTs for your gene (if you have more than 5 ESTs pick the 5 largest sequences that do not overlap) by clicking on the picture of the EST on the GENE BROWSER page.
b. This will bring you to the NCBI nucleotide database.
c. Click on the blue number corresponding to the EST category.
d. Then click on the GenBank accession number to retrieve the sequence.
e. **Highlight the sequence and past it into the document that has the other sequences. Label it "Tt GENE NAME EST#".**
f. Repeat this step until you have all your ESTs pasted into your document (up to 5 ESTs only). REMEMBER TO SAVE YOUR DOCUMENT AT EVERY STEP!!!

VI. EST alignment to confirm correct coding sequence.

a. In order to align the Genomic DNA for your gene with the ESTs that you copied you will need to use a program that can align multiple sequences together. Go to the Kyoto University Bioinformatics Center website for multiple sequence

alignments using a program called MAFFT (version 5.0) at http://align.genome.jp/mafft/.

b. Type ">GENENAME:" then hit return and Paste in your genomic sequence with 100 base pairs on each side or up as much as was underlined from the step IVm above. Then return again.
c. Next repeat this process and paste in your ESTs (>EST1: through >EST5:). Then submit your data for alignment. This process will take a few minutes.
d. Scroll through the alignment and compare the gaps in sequence alignment (the true introns from the ESTs) to the red and black genomic sequence from step IVm above. If the annotation on TGD was mispredicted mark on the sequence the correct intron/exon sequences. **Then copy and paste into a new section titled "NEW Tt GENE NAME CDS" and remove the intron sequence from the sequence (black sections).**

VII. Translation of NEW coding sequence into NEW *Tetrahymena* protein sequence.

a. Paste CDS from step VId above into EMBOSS TRANSEQ (http://www.ebi.ac.uk/emboss/transeq/index.html) to translate the new coding sequence to the new *Tetrahymena* protein sequence.
b. Change the FRAME setting to F (this will give all three open reading frames) and the TABLE setting to the CILIATE codon usage. Then click on the RUN button.
c. **Highlight the translated sequence that give the largest open reading frame and paste into your Word document. Name it "NEW Tt GENE NAME AMINO ACID" and remember to save.**
d. Paste the translated protein sequence into the NCBI BLAST (Go to the NCBI homepage: http://www.ncbi.nlm.nih.gov/, click on BLAST at top of page, then click on protein BLAST)
e. Run the BLAST and confirm that it hits best with your homolog and contains the conserved domains (located at top of page).

VIII. Comparing NEW protein sequence to homolog.

a. Paste the original protein sequence of the organism that you started with into SEQUENCE 1 box in the NCBI SPECIALIZED BLAST (Go to the NCBI homepage:

http://www.ncbi.nlm.nih.gov/, click on BLAST at top of page, then click on BL2SEQ under SPECIALIZED BLAST)

b. Then paste in the sequence of the TGD predicted homolog of your protein (step IIIf) into SEQUENCE 2 box.
c. Change the PROGRAM box to BLASTP and click the ALIGN button.
d. **Copy and paste the alignment in your Word document and label the proteins being aligned.**
e. Repeat alignment using newly translated protein (step VIIc) obtained using the ESTs to determine the true coding sequence.
f. **Copy and paste the alignment in your Word document and label the proteins being aligned. MAKE A NOTE IF THE NEW ALIGNMENT IS MORE HIGHLY CONSERVED THAN THE TGD PREDICTED (e-value, %identity, %similarity).**

Now with all this data you are ready to design the primers to amplify the flanking sequences directly adjacent to the coding sequence of your gene of interest.

IX. Designing primers for PCR of gene flanking sequences.

a. Each PCR primer should be around 20-24 base pairs. You will design a total of 4 primers in order to amplify **at least 1000 base pairs** of flanking sequence on both sides of your gene. You may need to alter the length slightly to get a primer with a better T_m (T_m is more important than length). Ideally, all of your primers should have about the same T_m. To calculate the T_m use the following equation: T_m = [(#A+#T)*2]+[(#G+#C)*3]. You can also use the "oligo calculator step 1." (http://www.promega.com/biomath/calc11.htm#melt_results for help with primer design. These calculations must be included in your notebook.
b. **Include the 4 primers with the T_m for each in your bioinformatics report and highlight the region that corresponds to the primers in the sequence (step IVm)**

Good primer design: - 3' end ends with at least one C or G
- T_m of both primers is within 3˚C of each other

-T_m should be between 51-56°C (using calculation method listed above)
- If possible, avoid long stretches of self-complementary sequences within a single primer, or between both primers. 3' ends of both primers should not be complementary

c. Once you have tried to design the primers (if time permits) you can try to use the Yeast Genome Database Primer Design tool found at: http://seq.yeastgenome.org/cgi-bin/web-primer to check your primer design.
d. Paste 1200-1500 base pairs of flanking sequence into the box and click the SUBMIT button.
e. Then a screen will appear that contains a number of parameters do not change anything and hit the SUBMIT button. If there is an error in the primer design it will tell you what you may have to change. Try to change the parameter and try again.
f. If these primers look better than the once designed above **insert these into your report explaining how they were obtained and marking their location on the sequence (step IVm)**.

X. Adding restriction endonuclease sites to your primers.

a. In order to clone the flanking sequences into the knockout vector them must have restriction endonuclease cleavage sites inserted at the beginning of each of the primers.
b. In order to determine which cleavage sites will be used the flanking sequence that will be amplified using the above primer sets will have to be checked to determine if any of the cleavage sites already exist.
c. Paste the flanking sequence including the sequence of the primers into the box in NEBcutter 2.0 (http://tools.neb.com/NEBcutter2/index.php). Then click on the SUBMIT button.
d. Once the screen appears with showing a picture of the sites, click on "0 CUTTERS" under LIST in the bottom right corner.
e. Check the list and determine if KpnI, SpeI, XhoI, and ApaI are present in your sequence.
f. If these restriction enzymes are in the "0 CUTTER" list then you can add them to the ends of your primers as follows:

- Forward primer for 5'flanking has KpnI site added (5'-ATACGCGGTACC…. Your primer sequence-3')
- Reverse primer for 5'flanking has SpeI site added (5'-ATACGCACTAGT…Your primer sequence-3')
- Forward primer for 3'flanking has XhoI site added (5'-ATAAACCTCGAG…Your primer sequence-3')
- Reverse primer for 3'flanking has ApaI site added (5'-ATACGCGGGCCC…Your primer sequence-3')

g. If these restriction enzymes are not in the "0 CUTTER" list ask the instructor for assistance finding another restriction enzyme site to place on the end of the primer.
Put the sequence of the primers in the Word document and label each primer with the Gene Name, position from start codon (-) or from stop codon (+) and (F) Forward or (R) reverse.
For example: UBC13-1313F:
5'-ATACGCGGTACC**ACGACTCTCACTCACTTTAGC-3'**

-The total length of your primers should be 32-35 base pairs.

h. Once you have the primers completed **EMAIL THEM TO THE INSTRUCTOR**. I will look them over and pick the best sets of primers from each group and order them for the lab the next week.

Helpful websites:

T_m Calculators for PCR:
http://www.finnzymes.fi/tm_determination.html
http://www.promega.com/biomath/calc11.htm#melt_results
Protein Domains:
http://expasy.org/prosite/

BIOINFORMATICS MODULE II – INSTRUCTOR'S GUIDE

Prerequisite: Students should be familiar with basic gene structure and the purpose of introns/exons, translation START and STOP codons, and 5' and 3' nontranscribed sequences (NTS's). It is also helpful to have had an introduction to restriction enzymes and palindrome sequences. For the primer design, students should have an understanding of polymerase chain reaction (PCR) components and theory, the role of the primers, and the concept of melting temperature (T_m).

Objectives:

1. Familiarize students with databases and basic bioinformatics tools.
2. Teach concepts related to gene structure, transcription, translation, codon usage, and splicing.
3. Gain an understanding of how gene annotation is carried out and confirm gene annotation for a gene.
4. Design oligonucleotide primers for use in gene knockout plasmid construction modules.

Time: 4 hours for all parts (If some areas are not done as homework it can be expanded into two 3 hour sessions with Steps I. through Steps V. in the first session and Steps VI. through Steps X. in the second session).

Materials:

- -Computational lab-one computer with Internet connection for each student is ideal.
- -Master computer workstation in which computer can be projected on a screen for the whole class to follow demonstration of bioinformatics lab by the instructor at this station.

Instructor's Notes:

At the previous lab session give the students a packet containing two journal articles and one review on a specific protein of interest found in a certain model system (i.e. *Saccharomyces cerevisiae*). For homework prior to this lab session have them read the abstracts and introductions of each of the papers to understand the basic function of the gene that they will be working on for this lab session.

It is important to demonstrate the lab on the computer projected on the screen for the whole class to follow. I have found that demonstrating two sections at a time gives a good breaking point to not overwhelm them with information (i.e. Demo section I and II; section III and IV; section V and VI; section VII and VIII; and then section IX and X).

Section I: Finding the amino acid sequence of your protein.

Start by explaining the purpose of databases and point out the many tools that can be used from databases. Introduce them to the National Center for Biotechnology Information (NCBI) website.

When doing a search for a specific protein of interest it is best to add the model organism name. This will narrow the results. When the students are looking through the results have them view at least 3 of the entries in order to select the most recent submission and confirm the amino acid length of each is the same. Make sure that the students record the accession number of the entry that they choose so they can record this in their report.

Section II: Determining if there is a *Tetrahymena* homolog.

Begin this section with a definition of "homolog" and how that can be further divided into orthologs and paralogs (see Bioinformatics Module I Instructor's Guide for detailed description).

Introduce them to the *Tetrahymena* Genome Database (TGD) and the many features of the database. Then introduce them to searching the database for a homolog using BLAST (Basic Local Alignment Search Tool) and explain that this program can align nucleic acid and amino acid sequences to those entered in specific databases to determine if there is a matching sequence in the database. The students will search the TGD database. In particular they will search the entire macronuclear genome for homologous sequence using BLASTP (protein to protein).

Once the results are displayed explain the concept of an e-value (expect value), how it is used, and what you can determine from it (see Bioinformatics Module I Instructor's Guide for detailed description).

When pasting the sequence into the Word file the hard returns can be removed by selecting "REPLACE" under the "EDIT" menu and typing "^p" in the "FIND" field and leaving the "REPLACE" field blank. Then select "REPLACE ALL." This same process can be used in later steps throughout

the report and spaces can even be removed in this same manner by just clicking the space bar in the "FIND" field and selecting "REPLACE ALL."

At this point it is a good idea to have the students check their top 3 hits with the instructor to make sure that they are doing everything correctly.

Section III: Obtaining the protein sequence of the *Tetrahymena* homolog.

Describe the central dogma of molecular biology as it relates to a gene (DNA) the transcript (mRNA; coding vs. noncoding; introns/exons; codon; start codon-ATG; stop codon-TGA), and the translated sequence (amino acids).

At this point describe the features of the TGD gene annotation page while scrolling down to the bottom of the page where they will obtain the protein sequence.

Section IV: Obtaining the nucleotide sequence of your protein homolog.

Retrieving the coding sequence is similar to the section above but to retrieve the genomic sequence you will have to use the gene browser function. At time the picture shown on the genome browser page does not match the direction of the sequence when it is retrieved as a FASTA file. It is important that the students click the "FLIP" box if the sequence does not start with an "ATG.' Make sure that the students know that from that point on they will have to flip the sequence, if necessary, for the rest of the lab sections.

Section V: Obtain the ESTs for your protein.

If a student does not have any ESTs for their gene, they can use the 2nd or 3rd best hit from the BLASTP to go through this exercise (NOTE: these students should clearly mark that there gene does not contain any ESTs and the TTHERM_# of the gene that they are using for section V and VI.

If there are multiple ESTs shown only have them collect 3-5 ESTs and have them pick those that only slightly overlap or do not overlap at all. Also, have them note if the image on the genome browser confirms or refutes the intron predictions.

Section VI: EST alignment to confirm correct coding sequence.

When doing the alignment if the sequences do not align similar to the sequence that is underlined in the genomic sequence obtained in section V.

Have the student compare each EST to the genomic sequence alone to determine the EST that is the problem.

Once this section is section is done and the intron/exon junctions are determined this can be completed as homework to be finished for the report.

Section VII: Translation of NEW coding sequence into NEW *Tetrahymena* protein sequence.

The students should remove the introns from the genomic sequence (that contains the underlined sequence) that were confirmed to be introns by the ESTs. Also, if there were no ESTs for areas of the gene remove the introns as they were predicted. This should give a new coding sequence file that can then be translated into a protein sequence using EMBOSS TRANSEQ. Make sure that the students choose the Ciliate codon usage or some of the glutamines will show up as a stop codon.

If the NEW coding sequence contains an earlier stop codon than predicted it may indicate that introns that did not have ESTs were mispredicted and that should be noted so that primers are not made in that region of the predicted gene (This is important if you are planning on performing RT-PCR on this gene in future labs-see GE LABS).

Section VIII: Comparing NEW protein sequence to homolog.

This section confirms that the newly annotated protein sequence is now a better homolog to the protein from the model organism that was used that the beginning of the lab. During this section discuss the difference between identity and similarity and make sure that they record the e-value of the new alignment.

This section can be demonstrated and if there is time constraints it can be assigned as homework to be included in the report.

Section IX: Designing primers for PCR of gene flanking sequences.

If students are working in groups of two one student should work on the upstream flanking sequence and the other student should work on the downstream flanking sequence. Explain what "flanking sequence" refers to and the reason that it is needed to make the knockout construct. Review the concept of homologous recombination as it relates in the cell to repair and how it can be used to replace a gene with a drug resistance marker. Explain how selection of transformants is determined after homologous recombination. Make sure to stress to the students that they need to make

sets of primers that will yield a PCR product between 1,000 and 1,500 bp for efficient homologous recombination.

If time permits have the students do **steps c-f** otherwise they may skip this step.

Section X: Adding restriction endonuclease sites to your primers.

Discuss restriction endonucleases, what sequence are recognized, cleavage patterns (blunt, 3' overhang, and 5'overhang, compatibility of overhang sequence), and how the names originated. This section allows the student to determine what restriction sites are located within there flanking sequence (this is important for cloning confirmation in the knockout construction module) and what sites are absent (this is important for choosing the correct sites to put on the ends of the primers).

For the 5' flank primers if KpnI or SpeI sites are present in the flanking sequence: SacI or BamHI can be used in place of KpnI; XbaI, NheI, and AvrII can be used in place of SpeI.

For the 3' flank primers if XhoI or ApaI sites are present in the flanking sequence: SalI can be used in place of XhoI; NheI and AvrII can be used in place of ApaI (cloning H4NEO cassette will have to be performed differently-see KOC LAB5 Instructor's Guide for further information).

SAMPLE BIOINFORMATICS REPORT CRITERIA

What should be in the report . . .

Introduction/objective

-A few sentences describing what you are trying to accomplish by doing this lab (both general and specific)

Results of Bioinformatics (in this order)

From the Word document put the following:

- Amino acid sequence of your starting protein
- The top three hits for *Tetrahymena* homologs (include the TGD gene name, e-values, and the best

human and Saccharomyces homologs with their E-values)

- The amino acid sequence for your strongest homolog (for some of you this may not be your top hit)
- The coding sequence for that homolog from part iii
- The genomic sequence for that homolog from part iii
- The genomic sequence with flanking sequence and underlined EST hits for that homolog from part iii
- 5 ESTs found for your homolog from part iii or for a gene given to you by the instructor (if the gene is not the same as throughout the rest of the report make sure to note what the TGD gene name is that you are using)
- Alignment of the ESTs with the genomic sequence from part vi.
- The new *Tetrahymena* coding sequence (only if ESTs found that the sequence in part iv above was wrong).
- The new translation of the sequence in ix above (only if changed from part iv above).
- BLAST alignment of protein sequence in part i with part iii (and part i with part x if coding sequence not predicted correctly).
- The 2 sets of primers (forward and reverse) for the 5' flanking and 3' flanking regions with the T_m for each, labeled clearly, written 5'->3', and with the added restriction endonuclease site added to the end of each. (**SEND THIS TO BY EMAIL in order to order the primers**).
- Highlight the region that corresponds to the primers that you designed in the sequence from part vi. *NOTE: I will look at the primers designed in part xii from each member of your group and this will be used to make sure that you have designed good primers and then I will order the best pair from your group to be used in the lab for next week.

Discussion/Conclusions

-In a few sentences talk about the results that you found above and things that you found about your protein and its homolog in *Tetrahymena* through this lab exercise.

Send your Bioinformatics report as a Word document by email to the instructor so we have as electronic copy of your data

Paste all of this information as the first thing in your Lab notebook too!!