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!!