BIOINFORMATICS MODULE I – INSTRUCTOR'S GUIDE

Prerequisite information: Students should have some familiarity with concepts of coding versus noncoding DNA in genomes and basic gene structure [introns/exons, translation START and STOP codons, 5' and 3' untranslated regions (UTR's)]. It also helps to have had some introduction to what a restriction enzyme is. For primer design, students should have a basic understanding of T_m, polymerase chain reaction, and role of primers in PCR.

Objectives:

1. Familiarize student with basic bioinformatics tools
2. Utilize tools to research the structure of a predicted gene in *Tetrahymena*
3. Reinforce understanding of concepts related to genome, gene structure/organization, and codon usage.
4. Teach concept of functional domains
5. Design oligonucleotide primers for use in subsequent cloning modules

Time: ~3 hours for all parts

Materials:

A computational lab – one computer (internet connected) per student is ideal. Two students on one computer will work, but is not ideal.

A master computer workstation that can be projected on a screen for the entire class to see. The instructor will demonstrate from this station.

PART I: Searching databases to find gene homologs

A. It is best to demonstrate use of NCBI, OMIM, and examples of individual databases. Even if students will be using only one of the described methods, it is educational for them to experience all three if time permits.

B. Explanations:

1. Begin with a definition of "gene homolog" (usually include ortholog vs. paralog)

Homolog: A gene related to a second gene by descent from a common ancestral DNA sequence. The term, homolog, may apply to the relationship between genes separated by the event of speciation (ortholog) or to the relationship between genes separated by the event of genetic duplication (paralog).

Orthologs: Genes in different species that evolved from a common ancestral gene. Based on sequence, they appear to be the same protein (performing the same job) in different organisms. Normally, orthologs retain the same function in the course of evolution.

Paralogs: Genes related by duplication within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.

2. Explain what BLAST is ("Basic Local Alignment Search Tool"). This program aligns nucleic acid or amino acid (protein) sequences to sequences in a selected database to find the closest matches.
3. Students will be searching the Tetrahymena Genome Database (TGD) specifically for *Tetrahymena* genes with your gene of interest from another organism. The particular database to search within TGD is "TIGR Gene Predictions", a database of gene coding sequences (only

60% of the entire genome) predicted by "The Institute for Genome Research" (TIGR) the same company that sequenced the genome. You will also need to select what kind of sequence it should align to. BLASTP is comparing protein (amino acid) sequences, whereas BLASTN compares nucleotide sequences.

4. E-value: Definition, how it is used, what is a "good" e-value?

EXPECT VALUE (E-VALUE) A parameter that describes the number of hits that would be 'expected' to occur by chance when searching a sequence database of a particular size. An e-value of 1 means that it would be expected to find a match with a similar score simply by chance. The lower the e-value, the more significant the match.
Basis of e-value (identical vs. similar amino acids). Show what a BLAST alignment looks like.

5. Describe differences between genomic sequence (introns + exons), gene coding sequence (exons only, start at ATG, end at TGA), and the translated sequence (amino acids).

6. If pressed for time, students may do this as homework. Often it is easiest to work on a hard copy of the sequence. It is valuable to have them go through by hand and find all of the predicted introns. The first exon should begin with ATG, the last exon should end with TGA. It may be necessary to go back and acquire a larger window of sequence (5kb). If you cannot find the ATG or TGA, the gene sequence is probably on the opposite strand. Going back one window and selecting "Flip Sequence" will give the reverse complement.

Part II: Functional Domains

This exercise will teach the concept of protein domains, and how they are used to predict the function of putative proteins. Each student will find the functional domains in their protein of interest and use these to predict their protein's activities.

A **protein domain** is a part of protein sequence and structure that can evolve, function, and exist independently of the rest of the protein chain. Each domain forms a compact three-dimensional structure and often can be independently stable and folded. Many proteins consist of several structural domains. One domain may appear in a variety of evolutionarily related proteins. Domains vary in length from between about 25 amino acids up to 500 amino acids in length. Domains often form functional units, such as the calcium-binding EF hand domain of calmodulin. These are often referred to as "functional domains". Finding domains in your protein is the first step toward understanding the function of your protein.

Students will be using the NCBI BLAST to recognize protein domains. The domains appear at the top of the search result page.

Not all sequences will have recognizable domains. If this is the case with a given protein, a message stating that no domains were found will appear at the top of the BLAST page. Students who do not get domain results should try a different domain search program. The 'SMART modular domain database' is recommended.

Part III: Confirming the open reading frame

In this exercise, students will not learn anything new about their protein, but it will illustrate what an open reading frame is, and differences in codon usage between organisms.

When 'Transeq' translates their coding sequence using the ciliate codon table, the first reading frame will be correct, naturally. Students will see a peptide beginning with methionine and ending with an asterisk (STOP). They can compare this with the next two reading frames, which should be littered with very frequent stops, illustrating how true open reading frames can easily be distinguished.

If students are unable to produce an open reading frame, this is usually because they have forgotten to select the ciliate codon table. In this case, many of the glutamines (Q) will be deciphered as STOPs (for Q, they use what is a STOP codon in most other organisms).

Part IV: Restriction enzyme map of coding and genomic DNA sequences

This exercise illustrates a basic computational cloning tool that is useful for subsequent cloning projects. It also illustrates the concept of enzyme recognition sequences, their palindromic nature, the types of cuts made (staggered or blunt), and resources to learn about these enzyme characteristics.

Students are encouraged to explore the mapping programs to see what kind of information they can obtain on enzymes.

It is most useful for students to print graphical maps of their cut sites instead of the actual sequence with cut sites (this is too long, hard to use, and wastes paper). They will be using these in subsequent cloning modules for analysis of their cloned gene.

Part V: Designing primers

For directional cloning into the pENTR-D/TOPO vector, the forward primer for gene amplification must have CACC at the 5' terminus. Students should design their forward primer beginning **one codon down** from the ATG (start) of their gene.

5' CACC…….. 3'

The reverse primer should begin (at the 5' end) with the reverse complement of the TGA (stop) codon at the end of their coding sequence.

5' TCA……. 3'

Hints for good primer design are included in the student handout.

REPORT SUGGESTIONS

It will be necessary for students to have the following in their notebooks and/or organized into a report that is collected:

1. The name of the starting gene from a different organism identified in **Part IA**

2. TTHERM numbers and E-values for the top three *Tetrahymena* orthologs **(Part IB3)**

3. Coding sequence, genomic sequence, and translated (ORF) sequence of the *Tetrahymena* gene that will be investigated further **(Parts IB4&5)**

4. Names of domains in protein and their relative locations **(Part II3 or II5)**

5. Prediction of gene function based on domains **(Part II4)**

6. Translations of the coding sequence in all 3 reading frames **(Part III3)**

7. Restriction enzyme map of the genomic and coding sequences **(Part IV4)**

8. Primer sequences including their calculated Tms, %G/C content, and number of bases. Regions where primers will anneal should be highlighted on the coding strand (assuming it is double stranded). **(Part V)**