Bioinformatics: Molecular Computational Tools (Module I)

Introduction: Computers have become an indispensable tool for the modern molecular biologist. A number of computational programs are available that help with the design of primers for PCR, show restriction enzyme cut sites on any piece of DNA, and translate DNA sequence into amino acid sequence (and vice versa), among many other features. In this lab, you will gain some familiarity with the range of molecular biology computational tools available to the public. As well, you will see how genome databases are organized, annotated, searched, and used by a molecular biologist.

Sequence databases: When a researcher determines the sequence of a gene or protein, or the three-dimensional structure of a protein, the results are deposited in databases which serve as resources to investigators around the world. Examples include GENBANK, a database of gene sequences maintained by the U.S. National Center for Biotechnology Information (NCBI), and SWISS-PROT, a collection of protein sequences maintained by the University of Geneva, and the Protein Data Bank, a repository of coordinates of three-dimensional structures of proteins. All of these resources are accessible to scientists (including you) via the Internet.

Using the available databases and sequence analysis software, one can obtain an array 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 12-week project during which you will clone a gene from the ciliate *Tetrahymena thermophila*, fuse it to the gene encoding GFP, put the engineered gene back into *Tetrahymena*, and induce its expression. The gene will be cloned by PCR, so the first step is to design primers to allow the amplification of your desired gene. This lab will take you through the steps to do that, and will also demonstrate the use of other bioinformatics tools.

Procedure:

I. Find a *Tetrahymena* homolog for your gene of interest.

A. First find the gene and protein sequence you are interested in from another organism.

Three ways:

1. NCBI site: (http://www.ncbi.nlm.nih.gov)

a) From the homepage, select the "Gene" database and search for "histone deacetylase", for example. If you want the gene from a particular organism, you should include the organism name in the search box. For example, "histone deacetylase human".

b) Click on a gene name. This will take you to a gene page that has various information on the gene.

c) From the gene page for your selected gene, select "Nucleotide" on the right hand menu panel.

d) Scan through the first few gene entries. Select one that represents "mRNA" or "cDNA".

e) Scroll down the page. You will see both nucleotide and translated amino acid sequence. Copy the translated sequence into a Word file. Keep this file handy - you will keep adding to it.

2. **Online Mendelian Inheritance of Man (OMIM):**
   a) You can use this site for finding human genes. Type your search word in the search box on the homepage.

   b) The page that comes up will describe the genes involved. Select 'Genbank' from the menu on the left. The page that comes up has links to the gene and protein information. You probably want the first gene on the list.

3. **Individual organism databases** (Saccharomyces Genome Database, Flybase, etc.)**:** Each database has a homepage where you can search for a keyword(s). The search will bring up the appropriate genes in that particular database. By selecting one, it should lead you to the coding sequence and protein sequence for that gene.

B. Search the *Tetrahymena* database for a homolog of your gene (http://www.ciliate.org)

1. Open the Tetrahymena database. Choose BLAST from the left hand menu column.
2. Copy/paste the translated sequence from your Word file into the query box. Select "BLASTP" as the search program. Select "TIGR Gene Predictions" as the data set to search. Start the search.
3. Write down the TTHERM numbers and e-values for your first 3 'hits'. Save for your notebook. For example, **TTHERM_00037210; 8.7e-45**

Consult an instructor to select one of these top three genes to proceed with.

4. Click on the TTHERM number for your selected gene. Scroll to the bottom of the gene page to obtain the gene coding sequence and translated (ORF) sequence. Copy/paste both of these into your Word document.

5. Obtain the genomic sequence (exons + introns) of your gene.
   - On the gene page, click anywhere in the "Genome Browser" window.
   - Under "Reports and Analysis" select "Download Decorated FASTA File" from the menu.
   - Below this, you may select how much of the genomic sequence you would like shown. If you want flanking sequences upstream or downstream of the START and STOP codons, then select a number of kb that is larger than the default shown.
   - Select "Configure". Under Protein Gene Predictions, select "CAPS", then hit "Go" at the bottom.

The resulting sequence will contain both exons and introns. Exons are in capital letters. Find the predicted gene coding region in this sequence and highlight it (both introns and exons).

If you cannot find it, you may need to go back one page and flip the sequence.
Copy the entire region with all introns and exons into your Word document.

II. Find the predicted functional domains of your protein

Knowing the domains in your protein will inform you of its putative function. Domains may be found through many different programs available on the internet. We will use one on the NCBI site.

1. Open the NCBI homepage. Select BLAST from the menu bar at the top.
2. Under Basic BLAST, choose "protein blast", and copy/paste the translated amino acid sequence of your *Tetrahymena* gene into the sequence query box. Hit 'BLAST' at the bottom left of the page to start the search.
3. The first screen to come up should have a diagram of the domains on your protein. Write down the domains and their relative locations (near the middle, at the C-terminus, N-terminus, etc.). If there are no domains identified, go to step 5.
4. **Predict your gene's function:** The functional domains in your protein are the best predictors of its function in *Tetrahymena*. Click on a domain. On the page that comes up, you may see results of alignments to similar domains in other proteins. Mouse over each domain schematic. Information about the function of each will pop up. Note the activities described. For example, what processes do these protein homologs affect? Do they function alone or in a complex? If they help catalyze a reaction, what is that reaction?

5. If no domains were identified through the NCBI BLAST, try searching with a different program and database (any one database can miss some information). Use the **SMART modular domain database**
   http://smart.embl-heidelberg.de/
   Paste in your sequence and search for: outlier homologues, PFAM domains, internal repeats, intrinsic protein disorder.

III. Confirm the open reading frame translation for the coding sequence

There are many programs available through the internet that will translate nucleotides into amino acid sequence. You will use a program that allows one to select the ciliate codon table for translation. Codon usage in ciliates is slightly different from that in most other organisms.

1. Access the translation program called "Transeq"(**http://www.ebi.ac.uk/emboss/transeq**).
2. Select translation in all "forward" reading frames ("F"), set the codon table for ciliates.
3. Paste your **coding sequence** into the search box. Run the program and review the output.

Identify the longest open reading frame (ORF) starting with a methionine (M) and ending with a STOP, designated by an asterisk. Copy all 3 reading frame translations into your Word document.

IV. Obtain a restriction map.

There are many restriction mapping programs available on the internet. Using one listed below, obtain a map of restriction endonuclease cut sites across each of your selected genes.

1. Select the option for cut sites on a LINEAR (not circular) piece of DNA.
2. Select only the 6 bp cutters (enzymes that recognize and cut 6 bp long sequences).
3. Obtain a graphical (image) representation of the map, instead of the sequence letters.
4. Obtain maps for both the coding and genomic DNA for your gene

This information will be used later to confirm that you have cloned the correct gene. Include a copy in your notebook.

Programs to try: 1) Webcutter (**http://users.unimi.it/~camelot/tools/cut2.html**)
2) NEBCutter (**http://tools.neb.com/NEBcutter2/index.php**)

V. Design primers for PCR. Each PCR primer should be 20-25 bases long. 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.

Use this formula to calculate T_m: **$T_m = 2(A+T) + 4(G+C)$**

Primer T_m calculations must be included in your notebook. Remember, you must also keep the reading frames in mind. For expression cloning, design primers using the **coding sequence**. Begin your forward primer one codon down from the ATG. Begin your reverse primer with the TGA stop codon.

Good primer design:
- 3' end ends with at least one C or G, but avoid more than 2 C/G in a row
- T_ms of both primers are within 5˚C of each other
- 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.
- Avoid long runs of one kind of base if possible
- Aim for ~30% G/C content if possible (or get as close as you can)

***Create a separate Word document with your two primer sequences:**

1. Write each in the 5' → 3' direction. Write 5' and 3' on the respective ends.
2. Designate which primer is 'forward', which is 'reverse'.
3. Write their respective T_m's that you calculated.
4. Copy the coding sequence of the gene into the document. Highlight on the sequence where your primers will bind.
5. For expression cloning modules: add a CACC to the 5' end of your forward primer. Do not recalculate the primer T_m with this addition (keep the original calculation).

*** Email this file to your instructor, who will order your primers.**

BIOINFORMATICS MODULE – 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)**