**DAY ONE – BIOINFORMATIC EXERCISES**

**NS&B – STG Integrative Molecular Module**

EXERCISE ONE – BLAST+ and TRANSCRIPTOME EXPLORATION

**Learning Objectives**

* Become comfortable manipulating BLAST to compare nucleic acid sequences
* Engage in gene discovery from a *de novo* assembled transcriptome via homology searches
* Prepare sequences for downstream analysis

**Rationale**

The ability to rapidly search a large database of sequence information is a common skill in molecular biology. Many different approaches can be used to accomplish this. One of the most common is to use the BLAST algorithm to compare sequences to identify homologous sequences at the protein or nucleic acid level. Especially with a *de novo* assembly, the ability to find genes of interest in the transcriptome assembly is very useful.

**What is BLAST?**

BLAST stands for Basic Local Alignment Search Tool. This tool allows for FASTA formatted sequences to be compared against one another based on their nucleotide or amino acid sequence information. This is useful for determining the most likely identity of an unknown sequence by aligning it against a large database of known sequences. Moreover, it also allows for quantitation of sequence similarity between species, giving a rough idea of species relatedness.

**FASTA sequence format**

The most common format for sequence data is the FASTA format. **FASTA format** is a text-based format for representing either nucleotide sequences or peptide sequences, in which base pairs or amino acids are represented using single-letter codes. A sequence in FASTA format begins with a single-line description, demarcated with the “>” followed by lines of sequence data.

**General Overview**

Electrical coupling is a critical aspect of coordination of activity in the stomatogastric ganglion. There is widespread electrical coupling among STG neurons, to varying degrees of connectivity and strength. The proteins responsible for coupling in invertebrates are the Innexin family of proteins, of which at least 6 have been identified in crustaceans.

The goal of our first week of qPCR exercises is to use electrical coupling and Innexin expression n the STG to investigate the relationship, if any, between the functional strength and impact of electrical coupling for given neurons in the STG to the expression of proteins responsible for coupling.

We will begin with an exploration of the transcriptome of the STG to search for Innexin coding genes via bioinformatic comparison of crab sequence with innexins identified from other species. Following gene identification, we will design primers that will allow of the quantitation of innexin transcripts via qRT-PCR.

BLAST+ Instructions

1. To download blast+, follow this link: <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>
2. Download the version that is compatible with your computer.
3. Open the file, agree to the license, and install it.
4. Open your command prompt (whatever that may be for your operating system).
5. Navigate to C:\NCBI\blast-2.4.0+
   1. Type the command cd to change folders (directories)
6. Make 3 new folders called Input, Output, and Databases
   1. Either use your file explorer, or use the command mkdir <folder\_name>
      * 1. Example: C:\NCBI\blast-2.4.0+>mkdir Input
7. Now that we are organized with our folders, let’s add some files to our Input and Databases folders.
   1. Copy (using file explorer or the command copy, cp for linux/termianl) the file that will be your query into Input.
   2. Copy the file that will be your subject/database into Databases
8. We need to now tell blast to format the FASTA file in our Databases folder into a BLASTable database. While in your bin directory, run the command below:
   1. makeblastdb –in C:\NCBI\blast-2.4.0+\Databases\myDB.fsa –dbtype nucl –parse\_seqids -out C:\NCBI\blast-2.4.0+\Databases\myNucleotideDB
9. Now we can take our input sequences and align them against the database we made.
   1. tblastx –db C:\NCBI\blast-2.4.0+\Databases\myNucleotideDB –query C:\NCBI\blast-2.4.0+\Inputs\nucleotide.fsa -evalue 1e-30 –out C:\NCBI\blast-2.4.0+\Outputs\myblast.html -html
      1. E-value is the expected value that a match could be made by chance, so the closer to zero this is, the more we trust our alignment to not be random.
      2. tblastx is useful because we expect there to be more sequence conservation at the protein level for a given gene than the nucleotide level due to the fact that the same amino acid can arise from many different codon triplicates.
10. Go to your Output folder, and you should see a newly created html file there. Open it.
    1. Here will be a list of sequences you used as a query and all the sequences it found matches for in the database, with the order based on the E-value.
       1. This is why having an E-value threshold can make your downstream analysis cleaner.
11. Now you have sequence targets that you can find in your database file.
    1. I recommend using wordpad, notepad, or a similar text editor to view the database and find your newly found sequence.
    2. A good final check is to blast the sequence online at NCBI to double check its alignment.

For more information on blast+, see the manual: <http://computing.bio.cam.ac.uk/local/doc/blast+_user_manual.pdf>

Definitions and Commands

Query – the sequence(s) that you want to align against a database. Query sequences are used as inputs.

Subject – any sequence(s) that you are aligning your queries against; usually a database, but can be individual sequences as well.

BLAST – Basic Local Alignment Search Tool

blastn – aligns a nucleotide sequence against another nucleotide sequence with similar sequences

Can change to -megablast to increase stringency on similarity

blastp – aligns a protein amino acid sequence (query) against other protein sequences (subject)

Has other features for comparing protein domains, but are much less often used.

blastx – translates a nucleotide sequence into protein (using all six frames possible) and aligns it against protein sequences.

tblastn – aligns a protein sequence query against a translated nucleotide subject/database.

tblastn is probably the least used blast tool

tblastx – aligns translated nucleotide sequence(s) against a translated nucleotide subject/database.

tblastx takes the longest, by far, as it is aligning 36x the information (6 translations for query, 6 translations for subject) compared to a blastn (no translation to amino acid).

makeblastdb -in \PATH\TO\FASTA\FILE\Protein.fa\OR\Nucleotide.fa -dbtype nucl OR prot -parse\_seqids -out \PATH\TO\DATABASES\FOLDER\DatabaseName

tblastx -query \PATH\TO\INPUT\FOLDER\nucleotide.fa -db \PATH\TO|DATABASES\NucleotideDB -evalue 1e-100 -out \PATH\TO\OUTPUT\Genus\_species-geneX.html -html

cd – change directory (folder)

cd .. will go up one directory.

copy – copy a file (this is cp in linux/unix/terminal)

cp <PATH1> <PATH2>

Your path is the list of directories to where you or a file currently are.

mkdir – make a folder/directory

move – move a file/folder (this is mv in linux/unix/terminal) from one directory or another. Also used to rename files/folders.

help – gives a list of available commands. When in doubt, ask for help!

BONUS:

color 2 – allows you to enter the matrix