Practical Application #4

Name: Tony Kabilan Okeke

Section: 062

Group Number: 3

For this assignment, you will be learning how to search for gene sequences on NCBI, and how to use the online **B**asic **L**ocal **A**lignment **S**earch **T**ool (BLAST). For the context of this tutorial we will be interested in the *LuxI* gene sequence and comparing between different organisms.

1. For review, what protein product does *LuxI* encode for? What does this protein (and any enzymatic product of this protein) do?

The product from *LuxI* is N-acyl-l-homoserine lactone (AHL). AHL is an autoinducer synthesized by I proteins or AHL synthases and is sensed by transcriptional regulators called R proteins. It is used as an autoinducer in gram-negative bacteria like *A. fischeri* for quorum sensing and sense population density in the extracellular matrix to induce bioluminescence.

2. Outside of *Allivibrio fischeri*, what other organism(s), either species or genera, would you anticipate homologous *LuxI* sequences to be present in? Why?

Gram-negative bacteria and other organisms in the Vibrio genus produce AHL predominantly for survival. AHL is produced by the *LuxI* gene and therefore we would expect to see homologous *LuxI* sequences in the majority of the gram-negative bacteria and Vibrio genus organisms. One of the examples is S. Typhimurium.

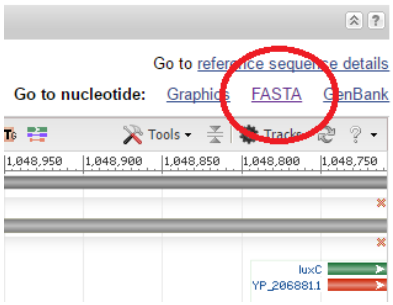
Before we can BLAST, the reference sequence is needed. Go to the NCBI website (https://www.ncbi.nlm.nih.gov/) and under the dropdown menu that says "All Databases" select "Gene." In the text box, type "LuxI" and hit search.

Select the first gene result, the autoinducer synthesis protein for "*Vibrio*[sic] *fischeri*." You may also use a different search result for an autoinducer synthesis protein if you like, for additional practice. Frequently you will find unusual genetic associations between organisms you would not have anticipated!

3. Why would a similar gene sequence be seen in two organisms that are not closely related?

Because this gene sequence has been conserved from some common ancestor. Also, between two distinct bacteria, horizontal gene transfer occurs when a bacteria non-sexually replications using binary fission. In addition, vertical gene transfer can be associated with the transfer of gene material to offspring.

Scroll down to “Genomic regions, transcripts, and products.” This is a convenient genome browser that lets you visualize the organism’s genome, and provides useful links. Currently it is centered on your autoinducer of choice, however you can zoom out or drag the window to see neighboring regions of the genome.



Next click FASTA (pronounced “Fast Ayyyyyyy” *thumbs up*). This directs you to the gene sequence in FASTA format that can be used in a variety of ways, including many offline tools. For now, simply copy and paste the sequence into a text pad (or leave this window open and proceed in a new tab).

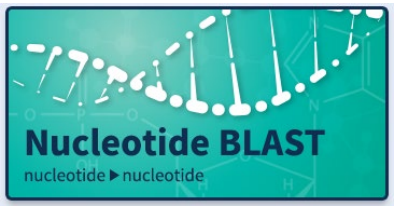
BLASTN

In the new tab, go to the NCBI website (https://www.ncbi.nlm.nih.gov/) and then click “BLAST” on the right-hand side. Notice that there are several ways of performing a BLAST search, and we will be covering two of them.

4. While we won’t cover the reverse situation of performing a BLAST search when starting with a protein sequence, describe a potential scenario when you would have a protein sequence, but want to BLAST for a nucleotide sequence.

> If an experiment were run to identify the amino acid sequence of a protein from an unknown organism (via mass spectrometry), the researcher would have to run a protein blast since they do not have any information about the genetic material of the unknown organism.

Click nucleotide BLAST (nucleotide -> nucleotide) button. This type of search expects a nucleotide reference sequence, and looks for similar nucleotide sequences in the database.



Input your sequence for the *luxI* gene.

Sometimes BLAST searches can return a lot of unwanted results. Since *luxI* is frequently used in plasmids, many cloning vectors are going to be listed, however for the context of this assignment we’re not interested in those. Under “Choose Search Set,” “Organism,” type in “taxid:29278” and from the list select “cloning vectors.” Then tick the “exclude” checkbox. This will remove them from the BLAST search.

Under “Optimize,” click “somewhat similar sequences.” Just ‘cuz.

5. Before searching, what do you predict to be the main BLAST result? Why?

We expect BLAST to return the *A. fischeri* genome sequence as the top hit. Since the *luxI* gene sequence we are running on BLAST is taken from the *A. fischeri* genome, we expect that the complete *A. fischeri* genome will contain an exact match for our query sequence; as such, it will most likely be the top result. We also expect to see matches from closely related strains of bioluminescent bacteria.

Click BLAST, and wait for results. Sometimes this can take several minutes, and certain times of the day during high traffic can also affect this.

6. What is the best result (Species, E-value, coverage)?

Description: Vibrio fischeri ES114 chromosome II, complete sequence
Species: Alivibrio fischeri
E-value: 0.0
Coverage: 100%
Score: 1034

7. Are there any other interesting results? (You may have to scroll down a bit.) Write some here.

Description: Synthetic construct gene for transcriptional activator protein LuxI, complete cds
Species: synthetic construct
E-value:4e-76
Coverage: 98%

Description: Shewanella hanedai lux operon (luxC, luxD, luxA, luxB, luxE, luxG) genes and flanking regions, strain: NCIMB 2157
Species: Shewanella hanedai
E-value: 4e-63
Coverage:78%

Description:Photobacterium phosphoreum LuxR (luxR) and LuxI (luxI) genes, partial cds
Species: Photobacterium Phosphoreum
E-value: 1e-56
Coverage: 33%

8. Which of these would you expect? Why?

We would expect *Shewanella Haneda*i because it is a gram-negative marine bacteria. In addition, *Shewanella Hanedai* is a bioluminescent bacteria such as *A. fischeri.*

Scroll down past “Descriptions” and look at “Alignments.” This is where each individual result from the list in “Descriptions” is shown in a visual alignment. The top line is the reference gene used to query the database, while the bottom line is what it was found to match up with.

BLASTX

Now return to the root BLAST page and this time click blastx (translated nucleotide -> protein).

9. What are the advantages of performing a BLAST based on the amino acid sequence encoded by the nucleotide sequence?

BLAST is more sensitive to patterns when comparing amino acid sequences since they are composed of 20 characters in comparison to nucleotide sequences which only utilize 4 characters (ACGT). Since amino acid sequences use more characters, the statistical significance calculated is more accurate when performing BLAST on them.

Again paste your gene sequence into the query field. It will automatically be converted to an amino acid sequence by the program. Keep in mind that there are situations where the genetic code will be different, such as mitochondria, however for our purposes the standard genetic code is correct.

Make sure to filter out cloning vectors from our search by typing in "taxid:29278" and clicking on "cloning vectors" and checking the exclude checkbox.

10. What is the best candidate gene and organism that BLAST returns now?

Description: GNAT family N-acetyltransferase [Aliivibrio fischeri]
Species: Alivibrio fischeri
E-value: 4e-139
Coverage: 99%

11. Note other BLAST hits in the results and compare with those alternate hits seen in the nucleotide – nucleotide BLAST from earlier. Have any shifted up in score? Down?

The results contain several matches with "GNAT family N-acetyltransferase" in *A. fischeri*, as well as a match with "3-oxo-C6-HSL autoinducer synthesis protein LuxI" in *A. fischeri*. Overall, the scores for the nucleotide-protein BLAST are lower than the nucleotide nucleotide BLAST. One reason for the lower alignment scores is that the protein sequences are shorter than the corresponding nucleotide sequences (in our case, the nucleotide sequence was 573 bases long, therefore, the protein queried by blastx was 191 amino acids long).

12. Other than similarities between two different amino acid's structures, give a reason for the shift in score.

The shift in scores results from having nucleotide to nucleotide matches that are more specific than protein to protein matches, which is reflected in the score being higher for nucleotide to nucleotide BLAST. This can be explained by the fact that multiple codons can code for the same amino acid which results in multiple arrangements that code for a

particular protein. This lowers the specificity of the BLASTx and therefore results in lower score.

13. As suggested in 12, if two different amino acids are "similar" enough, then sometimes a substitution won't affect protein function. Scroll through the alignments and list three commonly seen substitutions (illustrated with a "+" symbol between the query and subject line), and give a reason those two amino acids might be interchangeable.

Common substitutions:
$Q \Rightarrow E$ Glutamic acid (E) replaces Glutamine (Q) because they are both hydrophilic amino acids.
$R \Rightarrow Q$ Arginint (R) replaces Glutamic acid (E) because both are hydrophilic amino acids
$N \Rightarrow S$ Serine (S) replaces Asparagine (N) because both are hydrophilic amino acids

14. Do you see any high scores that are listed as "hypothetical protein?" Is this BLAST search sufficient to claim that those hypothetical proteins are indeed a functional *luxI* homolog? Why or why not? Include the species name associated with the hypothetical protein.

None of the high scores were listed as a "hypothetical protein." A hypothetical protein is a protein whose existence is predicted but lacks evidence of their actual existence. Therefore, there were no hypothetical proteins listed because the gene might be encoding for a different protein or pathway which results in no scores for hypothetical proteins.

15. If you wanted to verify the function of one of these hypothetical proteins, what would be the next step?

The next step would be to perform *in vivo* testing to analyze the behavior of the genes and their regulation. By upregulating the target gene for analysis in some populations of the organism and comparing it to the rest of the populations, it will aid in observing different phenotypes and determine the true effect of the target gene in question.

Bonus: List one interesting result you saw during your searches, and why it was interesting to you. 10,000 word minimum