BIOL199 BDB

Fall 2022

Lab 4: Using BLAST to investigate an unknown sequence from *Drosophila yakuba*

*Adapted by Melinda A. Yang, from "*An Introduction to NCBI BLAST" by *Wilson Leung, in the GEP*

Table of Contents

Appendix Link (http://tiny.cc/blast_appendix)

(Learning) Objectives:

- Find putative genes of interest using the appropriate *BLAST* program
- Describe and analyze alignments and associated metrics to help support conclusions from *BLAST* searches
- Communicate evidence to clearly support a finding or describe discrepancies affecting support for a finding

Pre-lab Assignments:

1. Read the protocol and complete the pre-lab quiz before lab at http://tiny.cc/BLAST_prelabquiz
2. You can change your answers to the pre-lab quiz. I will indicate sometime during lab (will be announced) when I have graded your lab - you can then revise until you have answered all questions satisfactorily.
3. The pre-lab quiz is worth 1 Close Reading point for Week 4.

Introduction

The Basic Local Alignment Search Tool (BLAST) is a program that can detect sequence similarity between a query sequence and sequences within a database. *Similarity* is a metric we often use to infer *genetic homology*, or the state of having shared ancestry from common evolutionary descent. For instance, both of us have the *BRCA1* gene. The one in me is homologous to the one in you - they might differ at a few bases creating different versions of this gene, but the gene itself is the 'same'. We can extend this to different species - for example, there are also mouse and dog *BRCA1* genes that are homologous to the human *BRCA1* gene. It's not always the case, but we generally expect homologs to have more similarities with each other than with other genes that are not homologs. We will return to

the concept of homology in a couple of weeks, after we establish our understanding of evolution, but the above description should be enough to help you work through this lab.

If we want to find what genes are in a novel sequence or examine the relationship of a gene to genes in other organisms, similarity is often the first metric we use to begin determining what genes we're looking at (and maybe, what function we might expect for the associated protein!).

BLAST is popular because it can quickly identify regions of *local similarity* between two sequences. More importantly, BLAST uses a robust statistical framework that can determine if the alignment between two sequences is statistically significant. In this walkthrough, ***we will use the National Center for Biotechnology Information (NCBI) BLAST service to help us annotate a sequence from the Drosophila yakuba genome (unknown.fna). You will use multiple BLAST programs to identify and characterize a putative gene in a genomic sequence from D. yakuba***. This tool and the one we will be exploring in Weeks 5-6 (Genome Browser) are the main tools we need to begin our research project in Week 7, when we will attempt to annotate genes in different *Drosophila* species for further comparison.

Our goal in this lab is to ***characterize an unknown genomic sequence (unknown.fna) and determine if it has sequence similarity to any known genes. If so, then we want to understand where this gene is in our unknown genomic sequence***. The data was sequenced from a *D. yakuba* fruit fly. To complete the goal, you will do multiple searches to determine a putative gene, and in the process, provide thorough support for your conclusions.

A. Using *blastn* to compare our unknown sequence to a database of known RNA sequences

The strategy that we are first using is to search for sequence similarity of our unknown DNA sequence to mRNA sequences in the NCBI Reference Sequence (RefSeq) database. Thus, we are examining what RNA transcripts match our unknown DNA sequence, allowing us to pinpoint a potential gene of interest.

Instructions	**Comments**
1. Open a new web browser window and navigate to the NCBI BLAST main page (http://tiny.cc/blastpage).	
2. Examine the graphic titled Web BLAST. Four of the five common BLAST programs are available through the "Web BLAST" section of the NCBI BLAST home page. The program *tblastx* is not listed under the "Web BLAST" section. However, you can access this program by clicking on any of the BLAST programs in the "Web BLAST" section and then click on the "*tblastx*" tab in the NCBI BLAST search form.	The five tools are used to perform sequence searches for sequences of high similarity to your **query sequence**. The type of BLAST search you need to use will depend primarily on the type of query sequence you have and the database you would like to search.

BLAST program	Query	Database
Nucleotide BLAST (*blastn*)	Nucleotide	Nucleotide
Protein BLAST (*blastp*)	Protein	Protein
blastx	Translated Nucleotide	Protein
tblastn	Protein	Translated Nucleotide
tblastx	Translated Nucleotide	Translated Nucleotide

The table above shows each of the five BLAST tools and the appropriate situation to use them. The type of data in your query sequence, and the type of data in the databases or sequences you search are the two things you want to consider.

3. To begin investigating the unknown sequence, download the file from today's lab on BB and put it in your Downloads/ folder.

4. Then, use your Mac Terminal or Windows MobaXTerm to navigate into your Downloads folder (throwback to Weeks 2 and 3!).

- Use `ls -lrth` to check if your unknown sequence file downloaded (called unknown.fna)
- Use `less unknown.fna` look inside the file (and use 'q' to exit when you're done)

For understanding a FASTA file, return to our previous Week 2 reading, http://tiny.cc/fasta_fastq, and read the first two paragraphs.

DISCUSS & CHECK-IN: As a group, determine the name of the sequence in the FNA file, as well as the type of sequence (protein, DNA, or RNA). Then, share your findings with your instructor or TA.
Name of sequence:_____

Type of sequence: _____

5. In your browser, return to the NCBI BLAST home page and click on the "Nucleotide BLAST" image under the "Web BLAST" section (i.e. *blastn*)

6. You must set up the search you want to do on the new page. Complete the following steps:

a. Under the "Enter Query Sequence" section, click on the "Browse" or the "Choose File" button and select the file with the unknown sequence from your Downloads/ folder
b. Enter the Job Title "*blastn* search *D. yakuba* / RefSeq RNA"
c. In the "Choose Search Set" section, change the database to "Reference RNA sequences (refseq_rna)"
d. Under "Program Selection", select "Somewhat similar sequences (*blastn*)"
e. Check the box "Show results in a new window" next to the "BLAST" button
f. Click "BLAST"

See image on next page for more details.

The set of steps shown are to take the unknown sequence and compare this sequence to a database of known RNA sequences that were previously studied.

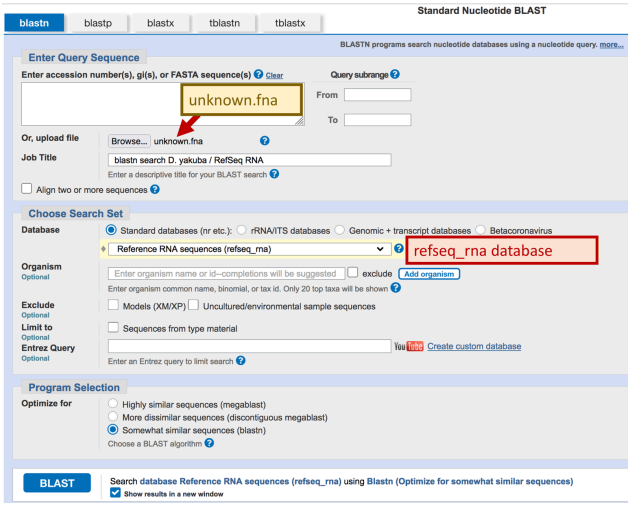

Image on left shows the search parameters we set up for our *blastn* search of the unknown sequence against the NCBI RefSeq RNA database.

Note that search may take a few minutes to complete if the NCBI web server is busy.

7. Once the search is complete, a new web page will appear with the BLAST report. Use the BLAST Appendix (http://tiny.cc/blast_appendix) to learn more about the important sections of the BLAST report.

DISCUSS & CHECK-IN: As a group, determine which hit to use and confirm this with your instructor or TA before moving to the next step.

Accession ID of hit used: _____

Answer question A1

8. Go to the alignment for your chosen hit and analyze it to answer our first question, on ***whether the entire mRNA aligns to our sequence.*** To do so, complete the following steps:

a. In the Descriptions tab, uncheck 'select all' at the top
b. Check the box next to your chosen hit
c. Click on either 'Alignments' or the link connected to your chosen hit → this will bring you to the 'Alignments' tab
d. <u>Re-order the coordinates of the alignment blocks with respect to your subject sequence (i.e., Sort by: Subject start position)</u>
e. Note all coordinates for the query sequence and the subject sequence in each alignment (Keep order from left to right, even when positions are reversed)
f. Assess whether your alignment blocks are **<u>collinear</u>** with respect to your sequence. This means that all the alignment blocks show consistent orientation for the query or the subject (the query and subject can have reverse orientations) and there is a high degree of sequence similarity (somewhat subjective, but 70% or above is a good starting point)

Answer questions A2-A3

9. With the information above, you now have a pretty good guess of what gene might be in your unknown sequence. Construct a hypothesis for the

Answer question A4

question of what gene the unknown sequence in *D. yakuba* contains. Make sure the format follows the structure we've discussed for hypotheses.

B. Using *blastx* to compare our unknown sequence to a database of known protein sequences

Above, you used *blastn* to determine the gene that your *D. yakuba* unknown sequence might contain. However, we often want to compare directly against the protein itself, to confirm that the coding sequences of the gene is actually present in the unknown sequence. An mRNA search will also include the untranslated regions (i.e., **5' and 3' UTRs**), and potentially, the match is only to the UTRs, rather than sequences that actually code for the protein (i.e. **the CDS's**).

We will set up a blastx search in order to compare a nucleotide genomic sequence against a protein database. Because every mRNA in the RefSeq RNA database has a corresponding sequence in the RefSeq Protein database, we will search our *D. yakuba* sequence against the RefSeq Protein (refseq_protein) database. We now have all the information we need to set up the *blastx* search.

Instructions

Comments

1. Navigate to the NCBI BLAST main page (http://tiny.cc/blastpage) and click on the "*blastx*" image

2. On the search page, adjust the following settings for your search:
   a. Under the "Enter Query Sequence" section, click on the "Browse" or the "Choose File" button and select our sequence (*unknown.fna*) from the appropriate folder on your computer.
   b. Enter the Job Title "*blastx* search *D. yakuba* / RefSeq Protein"
   c. In the "Choose Search Set" section, change the database to "Reference proteins (refseq_protein)".
   d. Check the box "Show results in a new window" next to the "BLAST" button
   e. Click "BLAST"

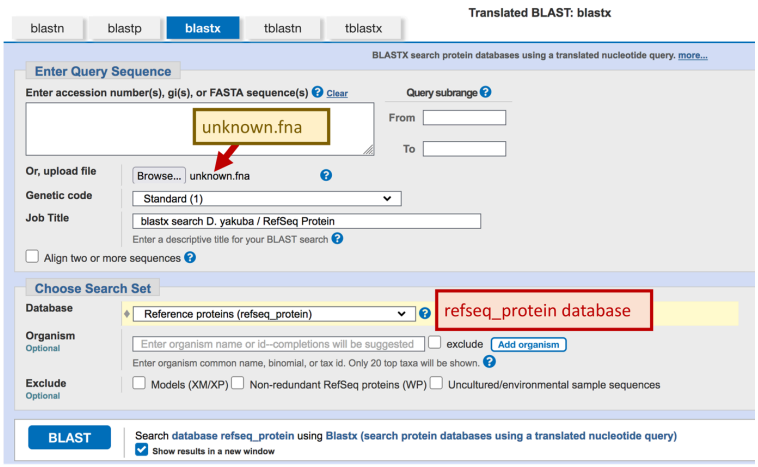

Image showing how to configure your *blastx* search of the unknown sequence against the NCBI RefSeq Protein database. Note that this blastx search can take several minutes to complete.

3. This *blastx* report is similar to the *blastn* report, only instead of nucleotide bases, we will see amino acids (or a * to represent a stop codon). As a result, these are some differences:

- Accession IDs should end with 'P' to indicate a protein, rather than an 'M' to indicate mRNA.
- The 'Frame' field for each alignment block will be either -3, -2, -1, 1, 2, or 3. This refers to the reading frame to match base pairs to amino acids. Thus, we are shown the reading frame and corresponding amino acids for the query (our unknown sequence), rather than the nucleotides of our unknown sequence.
- Multiple alignment blocks (or 'Ranges') often indicate that the two alignment blocks correspond to different coding sequences (CDS), especially if the frame changes.
- The "Positives" field corresponds to the number of amino acids that are either identical or have similar chemical properties between the translated query and the subject sequences.
- Between the query and subject sequences, a letter indicates the amino acid matches, a + indicates the amino acids do not match but have similar properties (behave similarly in cellular environment), and a gap indicates the amino acids do not match and do not have similar properties. Thus, more gaps indicates low similarities between the query and subject sequences.

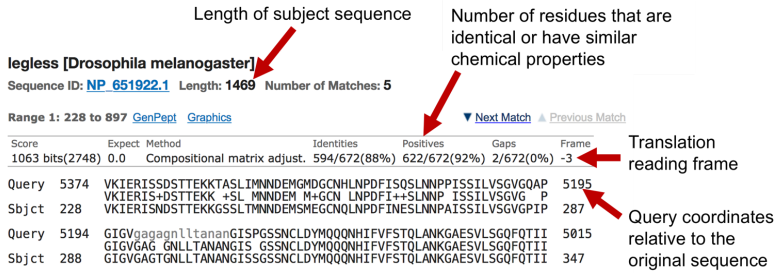

Image depicting the key characteristics of a *blastx* alignment.

Translating the code from nucleotides to amino acids corresponds to 3 bases for every one amino acid. *This results in six possible amino acid sequences for any one nucleotide sequence.* The 'Frame' field indicates which of these six matches with the amino acid sequence specified in the search.

A codon refers to a triplet of nucleotide bases that ultimately corresponds to an amino acid or stop codon. A stop codon indicates the end of an amino acid sequence and typically you would not expect to find it in the middle of a DNA sequence that codes for a protein.

Answer question B1

4. Find and **<u>sort the alignment blocks by the subject start position</u>**, following similar instructions to step A8. Examine your results based on the worksheet questions.

DISCUSS & CHECK-IN: Look with your group to see if there are any results that seem odd - this includes things that don't match your *blastn* search in Part A and large gaps/mutations in one or more of your alignment blocks. Confirm any odd results with your instructor or TA.

Answer questions B2-B3

Discuss with your group if you want to take a 3 minute break to rest your eyes. If anyone says yes, take a quick break!

C. Using the coding sequences of *D. melanogaster*'s *legless* gene to resolve discrepancies and develop a gene model for *legless* gene in *D. yakuba*

In Parts A and B, you performed searches comparing against an RNA and a protein database, and in both cases, you found that the best hit was for a *legless* RNA and protein, respectively (if not, go back and check through your work!). However, in Part B, you also noticed some discrepancies between the two searches, which need to be resolved.

The regions of a gene coding for a protein are usually under strong selective pressure to maintain the same amino acid sequence, so that the protein can perform its typical function. Thus, it is usually more likely to be ***conserved*** (i.e. similarity is maintained more than expected) compared to other regions of the genome and thus show higher similarity between species. Thus, to resolve discrepancies, we will now perform a separate *BLAST* search for each coding sequence (CDS) of the *legless* gene in *D. melanogaster*, as that information is readily available in the Gene Record Finder (http://tiny.cc/generecordfinder), a website that searches FlyBase (https://flybase.org/), a database of *Drosophila* genes and genomes, and pulls out information on genes in the *D. melanogaster* reference genome. By scanning CDS by CDS, we can confirm whether the sequences that compose the *legless* gene is present in our unknown sequence from *D. yakuba*, and potentially resolve why the *blastn* and *blastx* searches differed slightly.

Instructions	**Comments**
1. Navigate to the *Gene Record Finder* (http://tiny.cc/generecordfinder).	
2. Type "lgs" (the official FlyBase symbol for the *legless* gene) in the textbox and click on the "Find Record" button.	
3. Look at the 'mRNA Details' section of the gene report to note the number of isoforms for this gene.	
4. Scroll down and click on the 'Polypeptide Details' tab. The 'CDS usage map:' lists each CDS, and the blue and white table at the bottom shows details on each CDS.	**Answer question C1, the first two columns**
5. In the table showing information on each CDS, click on the first entry. Note this is in the Table where the first column is for 'FlyBase ID'. Upon clicking, you should see a text box open showing a sequence entry in FASTA file format. Copy everything in the textbox. **DISCUSS & CHECK-IN: Discuss with each other whether the sequence is for amino acids or nucleotides. If we are now comparing this sequence to the unknown sequence (i.e. this is the query and the unknown sequence is the subject), what *BLAST* software do you think we should use (see step A2 for list)? Share your answer with your instructor or TA.** ***BLAST* software: _____**	
6. Navigate to the NCBI BLAST main page (http://tiny.cc/blastpage).	

7. Click on the *BLAST* program you determined in step 5 for your search. This is to compare the copied CDS sequence back against the unknown sequence for *D. yakuba*.

We are not comparing a sequence to a database, but a smaller sequence against a larger sequence.

8. On the search page, adjust the following settings for your search:
   a. Select the checkbox "Align two or more sequences" under the "Enter Query Sequence" section
   b. Paste the CDS sequence for 1_9487_0 into the "Enter Query Sequence" field
   c. Click the 'Job Title' box - if you copied the entire textbox entry, the job title should automatically fill
   d. For the "Subject Sequence", click on the "Browse" or the "Choose File" button and select our unknown sequence (unknown.fna)
   f. Click on the "Algorithm parameters" link to expand this section.
   g. Verify that the "Word size" parameter is set to 3
   h. Change the "Compositional adjustments" field to "No adjustment" under the "Scoring Parameters" section
   i. Uncheck the box "Low complexity regions" under "Filters and Masking"
   j. Check the box "Show results in a new window" next to the "BLAST" button
   k. Click "BLAST"

In order to prevent BLAST from masking low complexity regions in our protein, we will turn off the low complexity filter. In addition, because we are only comparing two sequences, we will also turn off compositional adjustments under scoring parameters.

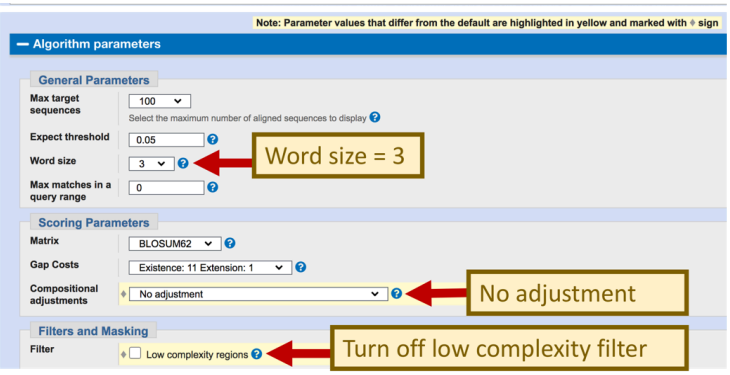

Image shows some of the settings that must be adjusted for the *tblastn* search

9. When the search is done, go to the "Alignments" tab - the format should be similar to that of the *blastx* search for amino acid sequence comparison.

Answer question C1 for the first empty row

10. Repeat steps 5-9 for each of the CDS entries on the Gene Record Finder page. Note that you can speed things up by returning to your original search window and adjusting only the query sequence and job title.

Answer question C1 for the remaining rows

11. Answer remaining worksheet questions.

Answer question C2

If you finished early, make sure you've completed all parts of the worksheet. Then, check in with your instructor, to have one worksheet chosen at random for submission in your group. Finalize your notes in your lab notebook on your findings, and then you're free to leave.