

BIOL 419/519 Bioinformatics Research - June 6, 2013

Practice Annotation Problem 1

Introduction

In this exercise, you will download the sequence and preliminary analysis of a *Drosophila ananassae* fosmid from Washington University in St. Louis. You will work through the initial characterization of the fosmid in order to come up with a list of genes on the fosmid. The characterization includes:

1. Using the six-frame translation of the entire fosmid to conduct a BLASTX search of all protein sequences from *Drosophila melanogaster*. This will allow the identification of homologous *D. ananassae* genes on the fosmid.
2. Using hypothetical proteins predicted by GENSCAN, a gene prediction program, to conduct a BLASTP search of all known protein sequences. This will either confirm the gene prediction or invalidate it.

When you have completed this exercise, you will be ready to carry out the initial analysis of the fosmid that will be your semester research project.

Resources

Course website:

<http://www.discoveryandinnovation.com/bioinformatics/>

From the main page, click the **Tools** link at the top of the page for access to links to various tools:

<http://www.discoveryandinnovation.com/bioinformatics/tools.html>

The **Class Projects** link gives access to this document and the required data file:

http://www.discoveryandinnovation.com/bioinformatics/class_projects2.html

The **Glossary** defines many terms in genomics and bioinformatics.

1. Download the fosmid sequence and associated files

Use the link on the **Class Projects** page to locate the 3L control fasmids from *Drosophila ananassae*. Download the folder **dananassae_3Lcontrol_Jan2013_fosmid_1475K17**. This is already on the desktop of the classroom Mac laptops.

The folder contains another folder named *src* that contains the sequence of the fosmid. It also contains a folder named *analysis* that has summaries of various kinds of analysis.

2. BLASTX search of the entire fosmid sequence against *D. melanogaster* proteins

We would like to learn what genes are present on the fosmid. *D. ananassae* is closely related to *D. melanogaster*. While it is possible that there are genes in *D. ananassae* that are not present in *D. melanogaster*, for a small segment of genomic DNA (a 44 kb fosmid), it is unlikely that there will be any genes that lack easily identified orthologs in *D. melanogaster*.

We will use a six-frame translation of the entire fosmid as the query sequence. The BLASTX program will generate this automatically. Imagine six long protein sequences, each almost 15,000 amino acids in length (44kb/3), containing multiple stop codons. We expect these query sequences to produce significant matches to protein sequences from *D. melanogaster* for those parts of the long protein sequences that correspond to conserved exons of reasonable length. While we might not be able to match short exons that are poorly conserved, we hope to match at least part of the coding sequence for each gene present on the fosmid.

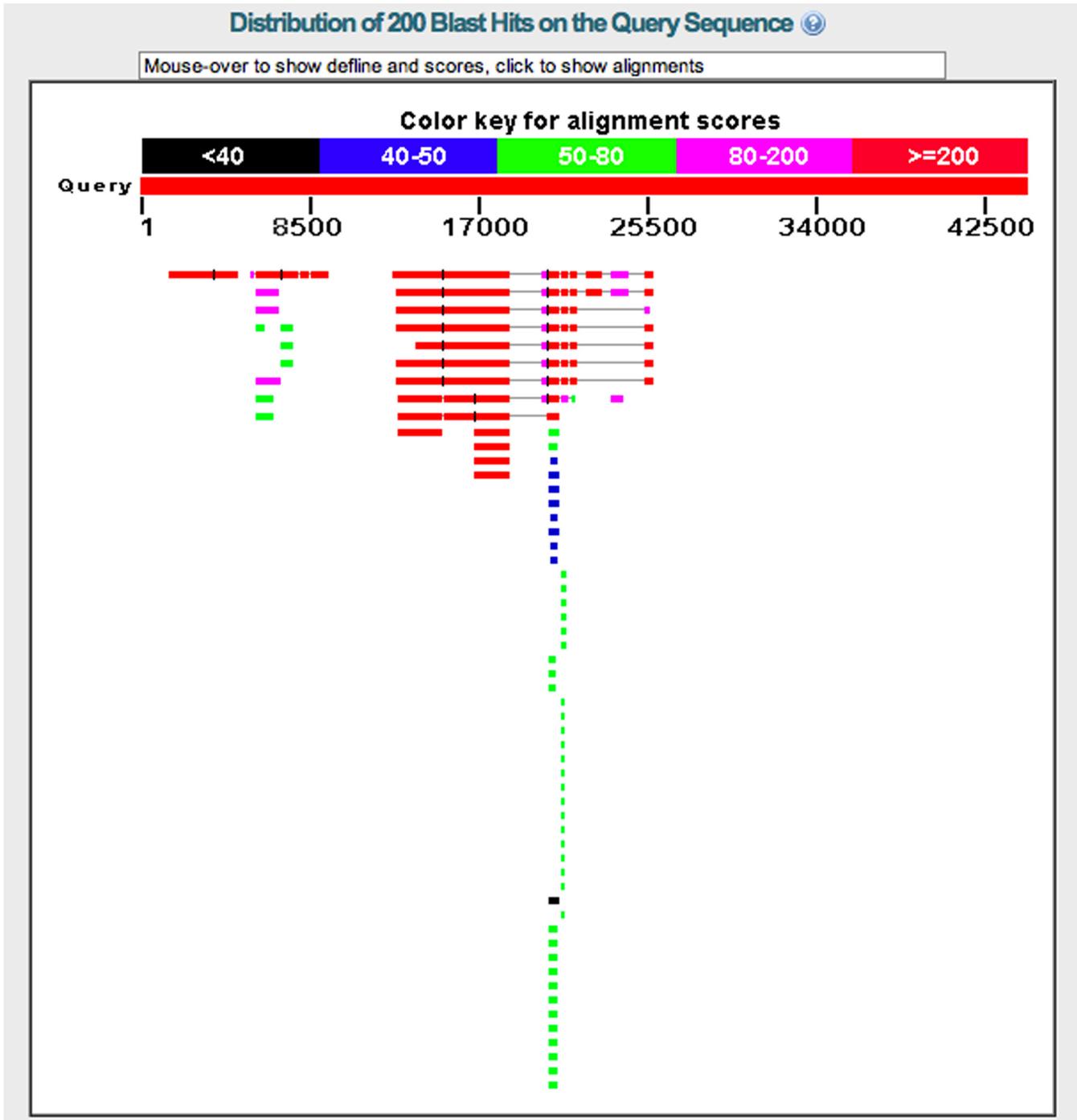
Use the link on the **Tools** page of the course website to navigate to blastx at NCBI.

The screenshot shows the NCBI BLASTX search interface. The page title is "Translated BLAST: blastx". The "Enter Query Sequence" section has a file upload button labeled "Choose File" with "fosmid_14...17.fasta" selected, indicated by a red arrow. The "Choose Search Set" section has "Database" set to "Non-redundant protein sequences (nr)" and "Organism" set to "Drosophila melanogaster (taxid:7227)", both indicated by red arrows. The "BLAST" button is visible, and the "Show results in a new window" checkbox is checked, also indicated by a red arrow.

1. Click the button to upload the query sequence and load *fosmid_1475K17.fasta* from the *src* folder.
2. Set the database to **Non-redundant protein sequences (nr)**.
3. Restrict the **Organism** to **Drosophila melanogaster**.
4. Check the box to **Show results in a new window** and click **BLAST**.

Results will resemble the screenshot on the next page.

Here is the graphic summary of the BLASTX search using query sequence *fosmid_1475K17.fasta* against the *D. melanogaster* nr protein database.



The thick lines indicate regions of significant alignments. In some cases, there are multiple regions of significant alignment to the same protein (the top alignment covering 17000-25500, for example).

CLASS DISCUSSION: Why do parts of the query sequence match proteins in broken segments? Why do parts of the query sequence match many proteins with lower scores?

Scroll down to the alignments.

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

[Alignments](#) [Download](#) [GenPept](#) [Graphics](#) 

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
<input type="checkbox"/>	tenascin major, isoform E [Drosophila melanogaster] >gb AGB94904.1 tenascin major, isc	2246	5781	21%	0.0	99%	NP_001262211.1
<input type="checkbox"/>	tenascin major, isoform D [Drosophila melanogaster] >gb ABW08579.2 tenascin major, is	2244	5738	20%	0.0	99%	NP_001097661.2
<input type="checkbox"/>	tenascin-like protein [Drosophila melanogaster]	2240	5204	16%	0.0	99%	CAA51678.1
<input type="checkbox"/>	odd Oz protein [Drosophila melanogaster]	2238	5298	16%	0.0	99%	AAB88281.1
<input type="checkbox"/>	odz pair rule gene product=tenascin homolog [Drosophila melanogaster, 9- to 12-hour-old	2238	4690	14%	0.0	99%	AAB30821.1

Click the checkbox next to the top alignment. Click the gear icon in the top right and deselect **Max Score**, **Total Score**, and **Coverage**. The display changes as shown below.

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:1 [Show all columns](#)

[Alignments](#) [Download](#) [GenPept](#) [Graphics](#) 

	Description	E value	Max ident	Accession
<input checked="" type="checkbox"/>	tenascin major, isoform E [Drosophila melanogaster] >gb AGB94904.1 tenascin major, isoform E [Drosophila mel	0.0	99%	NP_001262211.1
<input type="checkbox"/>	tenascin major, isoform D [Drosophila melanogaster] >gb ABW08579.2 tenascin major, isoform D [Drosophila me	0.0	99%	NP_001097661.2
<input type="checkbox"/>	tenascin-like protein [Drosophila melanogaster]	0.0	99%	CAA51678.1
<input type="checkbox"/>	odd Oz protein [Drosophila melanogaster]	0.0	99%	AAB88281.1
<input type="checkbox"/>	odz pair rule gene product=tenascin homolog [Drosophila melanogaster, 9- to 12-hour-old embryos, Peptide, 240	0.0	99%	AAB30821.1

Click the Download icon and select FASTA (aligned sequences) to show only those portions of the tenascin sequence that align with the query sequence. Why is the protein sequence broken into segments?

Click the GenPept link to go to the entry for tenascin major, isoform E. Click the FASTA link in the upper left of the page to go to the sequence of the protein. Note that the GenBank ID for this sequence is NP_001262211.1. Copy this using the copy function on your computer (Command-c on the Mac).

Return to the BLASTX page at NCBI. Follow these instructions. After the third step, your screen will resemble the screenshot on the next page.

1. Click the button to upload the query sequence and load *fosmid_1475K17.fasta* from the *src* folder.
2. Click the checkbox to **Align two or more sequences**. A new window appears.
3. Enter the accession ID *NP_001262211.1*.
4. Check the box to **Show results in a new window** and click **BLAST**.

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

NCBI/BLAST/blastx **Align Sequences Translated BLAST: blastx**

blastn blastp **blastx** tblastn tblastx

Enter Query Sequence BLASTX search protein subjects using a translated nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

Query subrange [Clear](#)

From

To

Or, upload file fosmid_14...17.fasta [Clear](#)

Genetic code [Clear](#)

Job Title

Enter a descriptive title for your BLAST search [Clear](#)

Align two or more sequences [Clear](#)

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence [Clear](#)

NP_001262211.1 [Clear](#)

Subject subrange [Clear](#)

From

To

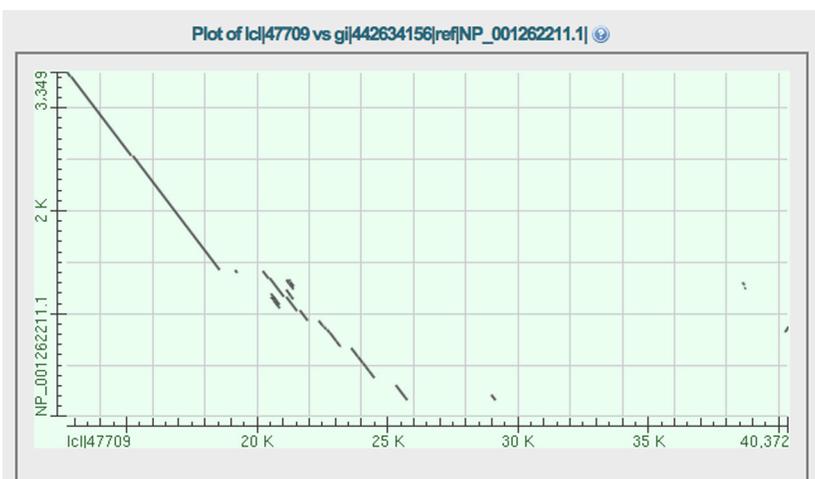
Or, upload file no file selected [Clear](#)

BLAST Search protein sequence using **Blastx** (search protein subjects using a translated nucleotide query)

Show results in a new window [Clear](#)

[+ Algorithm parameters](#)

In the page that appears when you click BLAST, click the + sign next to **Dot Matrix View** to see the image shown below.



In a dot matrix, two sequences are arranged as the axes of a graph. In our example, the *D. melanogaster* protein sequence is the Y axis and the six-frame translation of the fosmid is the X axis. When a window of sequence on the X axis matches a window of sequence on the Y axis, a dot is placed on the graph.

CLASS DISCUSSION: We will have a class discussion at this point to address the following questions:

1. What would a dot plot look like if the same sequence was used for both X and Y?
2. What would a dot plot look like if a sequence was used for one axis and the reverse sequence was used for the other axis?
3. What would the dot plots in the first two questions look like if a non-matching sequence was inserted in the middle of the sequence on the X axis?
4. What would the dot plots in the first two questions look like if a non-matching sequence was inserted in the middle of the sequence on the Y axis?
5. How do you explain the results of our **Dot Matrix View** comparing the six-frame translation of the fosmid to the protein sequence that we identified?

Have a look at the alignments below the **Dot Matrix View**. At the top of each segment of the alignment, the last entry in the header is **Frame**. Make a table of each of the aligned segments. In the first column, enter the first and last coordinates of the fosmid (the query sequence) for each segment. In the second column, enter the **Frame**. What can you learn from this table?

Use the **Tools** menu on the course website to navigate to **FlyBase**. In the **Jump to Gene** box in the upper right, enter *Ten-m* and click **Go**. In the page that appears, go to the **Genomic Location** section and click **View in GBrowse**. What can you learn about the gene model for *Ten-m*? Does this help you to understand the results of using the six-frame translation of the fosmid against all *D. melanogaster* proteins in a BLASTX search?

Repeat the search outlined on page 2. Are there any other genes besides *Ten-m* on the fosmid?

The *analysis* folder in the fosmid folder contains a folder called *BLASTresults*. Open the file inside this folder. You may have to duplicate the file (Command-d on the Mac) and rename it with a .txt extension in order to be able to open it. How do the entries in this file correspond to your results?

CLASS DISCUSSION: How many genes has the BLASTX search identified on the fosmid? When considering this question, multiple isoforms of the same protein should be considered to be products of a single gene.

3. Evaluating GENSCAN predictions

GENSCAN is a gene prediction program that searches genomic DNA for regions that might be genes. GENSCAN's predictions are based on machine learning. GENSCAN was provided with a set of "training" genes that have very solid experimental evidence for the gene models. By studying the training set, GENSCAN has learned the properties of sequences encoding genes, and can predict genes when presented with novel genomic sequence.

GENSCAN is one of several gene prediction programs that we will use. GENSCAN tends to over-predict genes, spotting genes that are not real. This is not very much of a disadvantage, because it is relatively easy to check whether GENSCAN's predictions are valid.

The *analysis* folder in the fosmid folder contains a folder called *GeneFinder*. Inside the *GeneFinder* folder is another folder called *Genscan*. Find the document called *fosmid_1475K17.fasta.masked.genscan*. In order to open this file with a text editor, you should duplicate it (Command-d on the Mac) and rename the duplicate with a .txt extension. There is also a graphic representation of the GENSCAN predictions aligned to the fosmid (*fosmid_1475K17.fasta.masked.genscan.pdf*).

GENSCAN predicts five genes on the fosmid, numbered 1 through 5. For each prediction, you can read the coordinates of the gene model on the fosmid. This makes it possible to see whether any of the GENSCAN predictions corresponds to the genes that you found in the BLASTX search.

The file also contains the predicted peptide sequence for each of the five genes. You can use each of the peptide sequences as a query sequence for a BLASTP search using the following steps.

Go to the BLASTP page at NCBI using the link from the **TOOLS** page.

1. Paste the sequence of a predicted peptide in the query box.
2. Set the database to **Non-redundant protein sequences (nr)**.
3. Do not restrict the organism (the default setting).
4. In the **Exclude** section, click both checkboxes to exclude **Models (XM/XP)** and **Uncultured/environmental sample sequences**.
5. Check the box to **Show results in a new window** and click **BLAST**.

This will search the database of all known proteins for any protein that matches the GENSCAN prediction. For each predicted peptide, record the identity of any significant alignment and the E value. Remember that E values larger than e^{-5} are extremely questionable.

CLASS DISCUSSION: Did GENSCAN find any new genes that we missed using BLASTX?