Lab Week 8 – An In-Depth Introduction to NCBI BLAST

(document created by Wilson Leung, Washington University)

Resources:
The BLAST web server is available at http://blast.ncbi.nlm.nih.gov/Blast.cgi. The Gene Record Finder is available at http://gander.wustl.edu/~wilson/dmelgenerecord/index.html

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. The ability to detect sequence homology allows us to identify putative genes in a novel sequence. It also allows us to determine if a gene or a protein is related to other known genes or proteins.

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 tutorial, we will use the BLAST web interface at the National Center for Biotechnology Information (NCBI) to help us annotate an unknown sequence from the *Drosophila yakuba* genome.

The NCBI BLAST web interface:
Before we begin analyzing any unknown sequence, we should first familiarize ourselves with the NCBI BLAST web interface. Open a new web browser window and navigate to the BLAST main page at http://blast.ncbi.nlm.nih.gov/Blast.cgi. In this tutorial, we will only use a few of the tools available. If you wish to learn more about the advanced options available (such as My NCBI accounts) on the BLAST interface, click on the 'Help' button at the top of the page at any time (Figure 1).

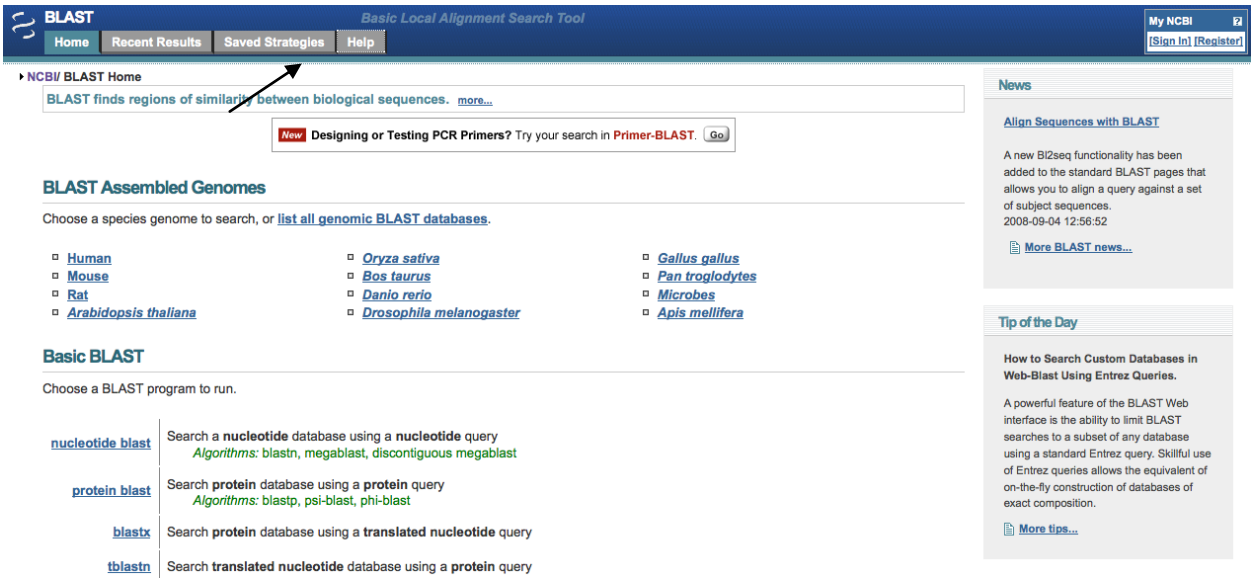


Figure 1. Click on the "Help" button to learn more about the BLAST web interface at NCBI.

All NCBI BLAST pages have the same header with four tabs:

Tab	***Explanation***
Home	Link to the BLAST home page
Recent Results	Link to results of the BLAST searches you have performed in your current browser session
Saved Strategies	BLAST input forms with the parameters you have saved to your MyNCBI account
Help	List of all BLAST help documentations

Besides the page header, there are two other sections that are of interest: the "Basic BLAST" section contains links to common BLAST programs. The type of BLAST search you should use will depend on the type of query sequence and the database you wish to search. The different BLAST programs are summarized below (Figure 2):

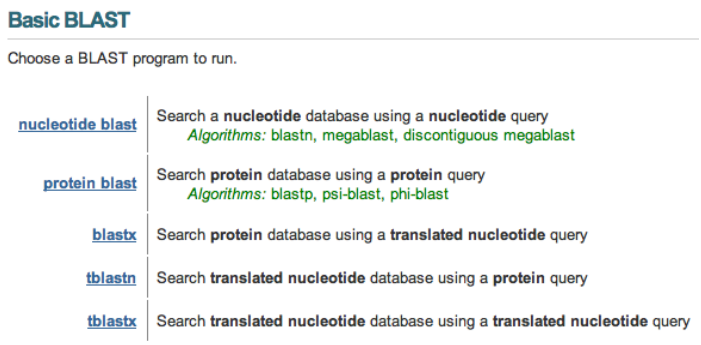


Figure 2. The different BLAST programs available on the NCBI web server.

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

You can also align two (or more) sequences using 'blast 2 sequences' (bl2seq) service under the 'Specialized BLAST' section of the NCBI BLAST main page (Figure 3).

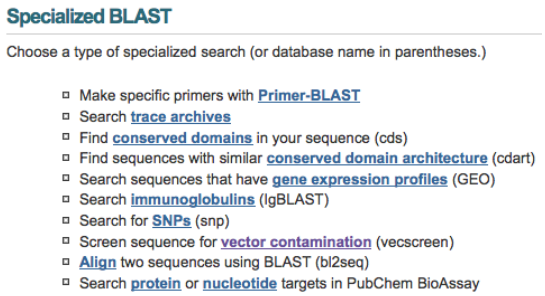


Figure 3. Specialized NCBI BLAST searches include searching for vector contamination or aligning two sequences

Objective of Tutorial:
Our goal is to determine if the unknown genomic sequence from *Drosophila yakuba* (a relative of the model fruit fly *Drosophila melanogaster*) contains region(s) with sequence similarity to any known genes. The unknown sequence is an 11,000 base pair (bp) fragment of genomic DNA, and the objective of **gene annotation** is to find and precisely map the coding regions of any genes in this part of the genome.

When we design a BLAST search, there are three basic decisions we must make: the BLAST program we wish to use, the query sequence we want to annotate, and the database we want to search. In addition, there are several optional parameters (such as the 'expect' threshold and other scoring parameters) that we can use to modify the behavior of BLAST.

Detecting sequence homology to mRNA using blastn:
One strategy to finding protein-coding genes is to search for sequence similarity to mRNA sequences. Thus, we will first perform a blastn search using our unknown genomic sequence from *D. yakuba* as the Query input, to search the Reference mRNA Sequences (Refseq) nucleotide database. The Reference Sequence (RefSeq) database contains sequences that have been reviewed by scientists at NCBI, to provide an integrated, non-redundant, well-annotated set of sequences. We will set up our BLAST search using mostly default parameters (Figure 4).

1. Download the *unknown.txt* sequence from 'Assignments' on Blackboard by right clicking on it and saving the file onto the desktop.
2. Navigate to the NCBI BLAST web server and click on "nucleotide blast".
3. Click on 'Browse' and select our sequence (unknown.txt); you can also paste the copied sequence directly into the Query box
4. Enter a Job Title "blastn search D. yakuba / Refseq RNA"
5. In the "Choose Search Set" section, change the database to "Reference mRNA sequences (refseq_rna)".
6. Under "Program Selection", select "Somewhat similar sequences (blastn)"
7. Check the box "Show results in a new window" next to the "BLAST" button
8. *Click "BLAST"*

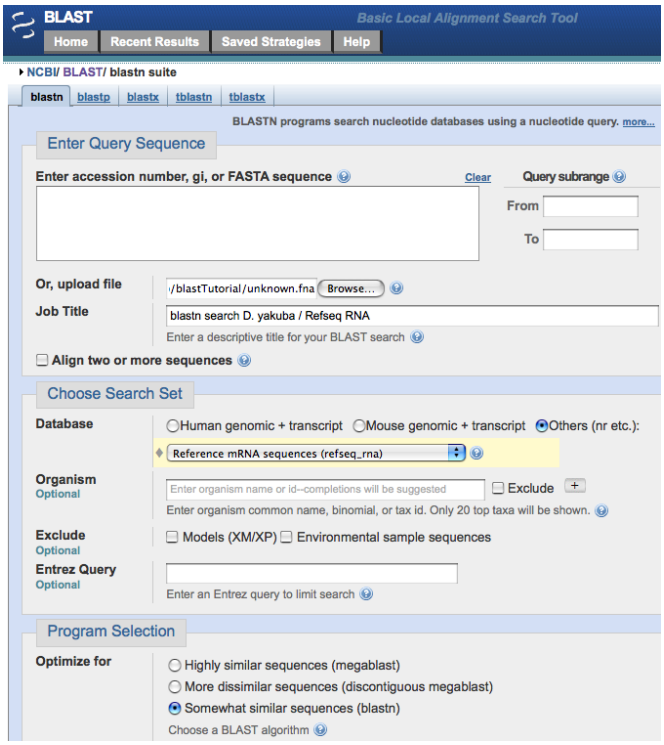


Figure 4. Setting up our blastn search of our unknown sequence against the NCBI Refseq RNA database

When the NCBI web server is busy, the search may take 5 minutes or more (Figure 5).

Figure 5. Waiting for our blastn results to arrive.

Once the search is complete, a new web page will appear with the BLAST report. The BLAST output begins with a description of the version of BLAST used, and some details on the database and the query sequence used in the search. The rest of the default BLAST report consists of three main sections: a graphical summary (Figure 6a), a list of blast hits, and the corresponding alignments. We will go through each of these sections in order to interpret our blastn output.

I. Graphic Summary

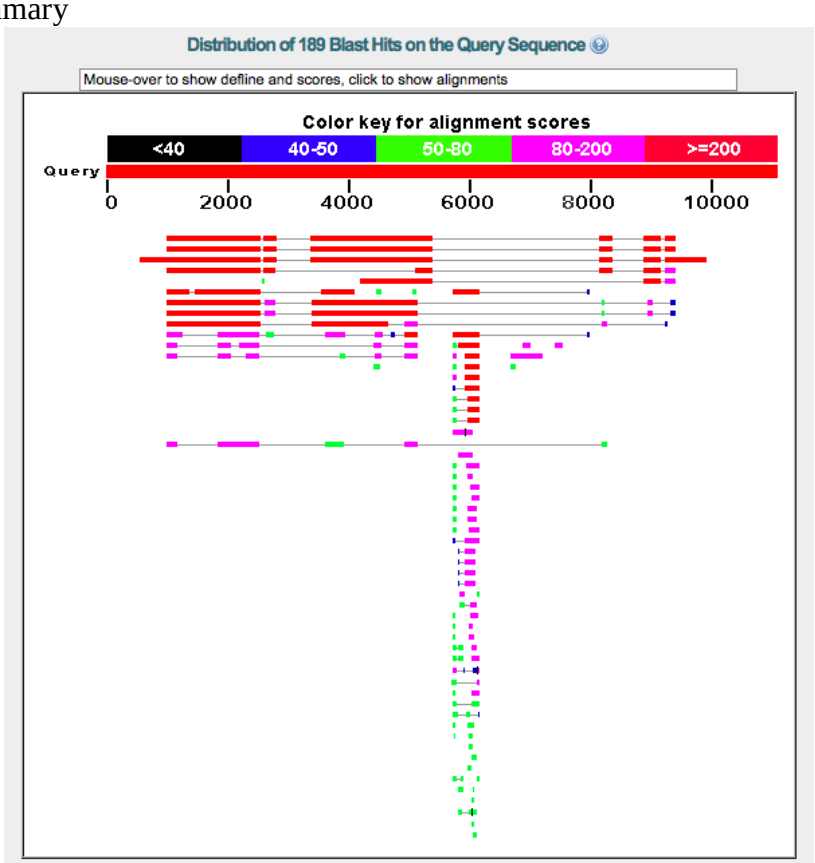


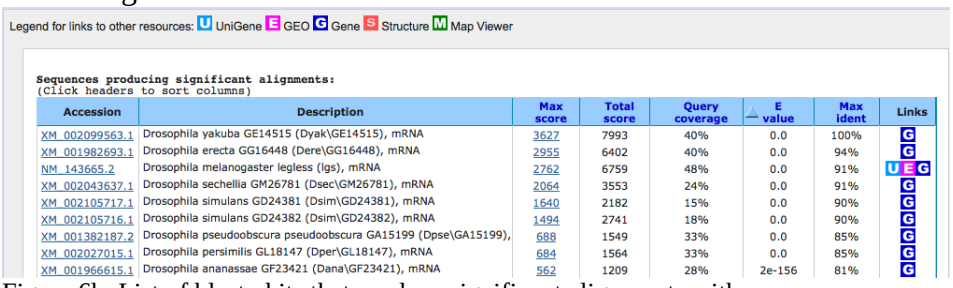
Figure 6a. A graphical overview of all the Refseq mRNA blastn hits for our query sequence

The Graphic Summary shows alignments (as colored boxes) of database matches to our Query sequence (solid red bar under the color key). As its name suggests, BLAST is designed to identify *local* regions of sequence similarity. This means that BLAST may report multiple discrete regions of sequence similarity between a query sequence and a subject sequence in a database. For example, if a spliced (mature) mRNA sequence is aligned to the unknown genomic sequence, we would expect to see multiple alignment blocks (many of which likely correspond to transcribed exons) in our BLAST output. Regions of the genomic sequence without significant alignment that fall between these exons are likely to be introns.

The color of the boxes corresponds to the score (S) of the alignment, with red representing the highest alignment scores. (See slide #4 in the previous Power Point handout for how the value of S and E are computed.) Generally, the higher the alignment score, the more significant the hit. When you move your mouse over a BLAST hit, the definition and score of the hit will be shown in the text box above the graphical overview. When you click on a box, you will jump to the actual DNA alignment associated with that BLAST hit (see III below).

In this case, we notice that the top three hits match much better to our sequence than the remaining BLAST hits. We also see that these three database matches span almost the entire length of our Query sequence.

II. List of Significant BLAST Hits



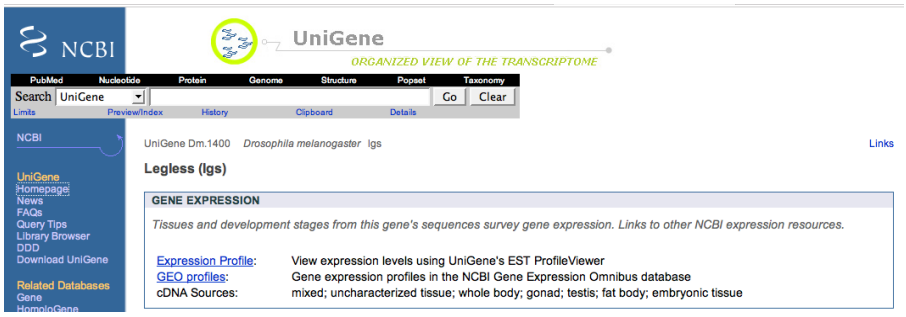
Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
XM_002099563.1	Drosophila yakuba GE14515 (Dyak\GE14515), mRNA	3627	7993	40%	0.0	100%	G
XM_001982693.1	Drosophila erecta GG16448 (Dere\GG16448), mRNA	2955	6402	40%	0.0	94%	G
NM_143665.2	Drosophila melanogaster legless (lgs), mRNA	2762	6759	48%	0.0	91%	U E G
XM_002043637.1	Drosophila sechellia GM26781 (Dsec\GM26781), mRNA	2064	3553	24%	0.0	91%	G
XM_002105717.1	Drosophila simulans GD24381 (Dsim\GD24381), mRNA	1640	2182	15%	0.0	90%	G
XM_002105716.1	Drosophila simulans GD24382 (Dsim\GD24382), mRNA	1494	2741	18%	0.0	90%	G
XM_001382187.2	Drosophila pseudoobscura pseudoobscura GA15199 (Dpse\GA15199),	688	1549	33%	0.0	85%	G
XM_002027015.1	Drosophila persimilis GL18147 (Dper\GL18147), mRNA	684	1564	33%	0.0	85%	G
XM_001966615.1	Drosophila ananassae GF23421 (Dana\GF23421), mRNA	562	1209	28%	2e-156	81%	G

Figure 6b. List of blastn hits that produce significant alignments with our query sequence

Scrolling further down the output, we find a summary table that shows all the sequences in the Refseq database that show significant sequence homology to our sequence (Figure 6b). By default, the results are sorted according to the Expect value (E-value) in ascending order. We can click on the column headers to sort the results by different categories.



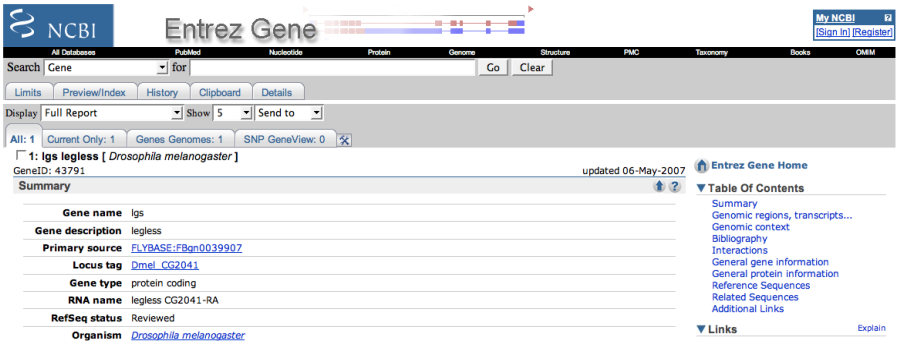


Figure 6c. The blastn hit list contains links to the NCBI UniGene (previous page) and Entrez Gene databases.

Clicking on the Accession number in the table will bring up a new page with the Genbank record for the BLAST hit. Clicking on the 'Max score' will bring us to the corresponding alignment in the BLAST output. Depending on the database you use, there may also be additional links to other parts of NCBI. For example, there are links to UniGene (U) and Entrez Gene (G) sections of NCBI for the third hit "*Drosophila melanogaster* legless (*lgs*), mRNA (NM_143665.2)" (Figure 6c). UniGene allow us to examine expression data for a gene while Entrez Gene provides us with an overview of the gene and links to additional literature references.

III. List of Alignments

Following the table of BLAST hits is a section showing all of the alignment blocks for each BLAST hit (Figure 6d). The sequence alignments show us how well our query sequence matches with the subject sequence in the database. Since we will rely heavily on sequence alignments in our annotation efforts, we will examine the alignment view of the *Drosophila melanogaster* legless mRNA sequence more closely.

```
>ref|NM_143665.2| UEG Drosophila melanogaster legless (lgs), mRNA
Length=5359

 GENE ID: 43791 lgs | legless [Drosophila melanogaster] (Over 10 PubMed links)

                                                    Sort alignments for this subject sequence by:
                                                      E value  Score  Percent identity
                                                      Query start position  Subject start position

 Score = 2762 bits (3062),  Expect = 0.0
 Identities = 1822/2016 (90%), Gaps = 6/2016 (0%)
 Strand=Plus/Minus

Query  3359  ATTACCAGCAGAGGACTGACCGAATGAGTCCAATTCCTTTGGTGATAGATGGGATAGAGG  3418
             || |||||||||||||||||| || || |||||||||||| ||||||| ||||   ||
Sbjct  3205  ATCACCAGCAGAGGACTGACCGAAAGACTCAAATTCCTTTGGGGATAGATGAGATAAGGG  3146

Query  3419  GGTACTTGGGTTGCTGTTTAAGTTATGCGTAAGAACGCTTGATGATGCGGTATTTCTACT  3478
             ||||||||||||||||| |||||||||||||||||||||||| ||| ||||||||||||||
Sbjct  3145  GGTACTTGGGTTGCTGCTTAAGTTATGCGTAAGAACGCTTGACGATCCGGTATTTCTACT  3086

Query  3479  TCGATTTTGATTTGACGGCGATGGGGTGTCTGCCTGAAAACAGTTTTTGTTTGCTGAGAG  3538
              |||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||||
Sbjct  3085  ACGATTTTGATTTGACGGCGATGGGGTGTCTGCCTGAAAACAGTTCTTGTTTGCTGAGAG  3026

Query  3539  CACTGTTGTTGTGCCAGCCTGAGCCGCCGACGTATTAGCTTGTGGAGCAGATCCAGATAA  3598
             |||||||||||||||| |||||| ||| ||| |||||||||||| ||| ||||||||||||||
Sbjct  3025  CACTGTTGTTGTGCCTGCCTGAACCGTCGAAGTATTAGCTTGCGGAACAGATCCAGATAA  2966
```

Figure 6d. Part of the blastn alignment between the unknown Query sequence and the subject Refseq mRNA sequence for legless.

The list of alignments for *different* BLAST hits (subject sequences in the database) is separated by definition lines that begin with a '>' character, followed by the Accession number and name of the subject sequence (see Figure 6d). One or more blocks of alignments will be listed

following the definition line. Each alignment block demarcates a local region of similarity between the query sequence and the subject sequence identified by the definition line. For example, blastn reported six different alignment blocks to the subject sequence – the *legless* mRNA from *D. melanogaster*. Each alignment block represents a region of the *legless* gene that shows sequence homology with our unknown genomic sequence from *D. yakuba*. Note that by default, these alignment blocks are vertically arranged by decreasing S values, which often does not match their physical arrangement (e.g., from left to right) along the Query DNA molecule in the Graphic Summary. You can use the 'Sort' links in the upper right of the alignment view to sort the blocks in a variety of different ways.

What about the alignments themselves? Each alignment block begins with a summary that includes the Max score and Expect value (the statistical significance of the alignment), sequence identity (number of identical bases between the query and the subject sequence), the number of gaps in the alignment, and the orientation of the query sequence relative to the subject sequence. The alignment consists of three lines: the query sequence, the matching sequence, and the subject sequence (Figure 7).

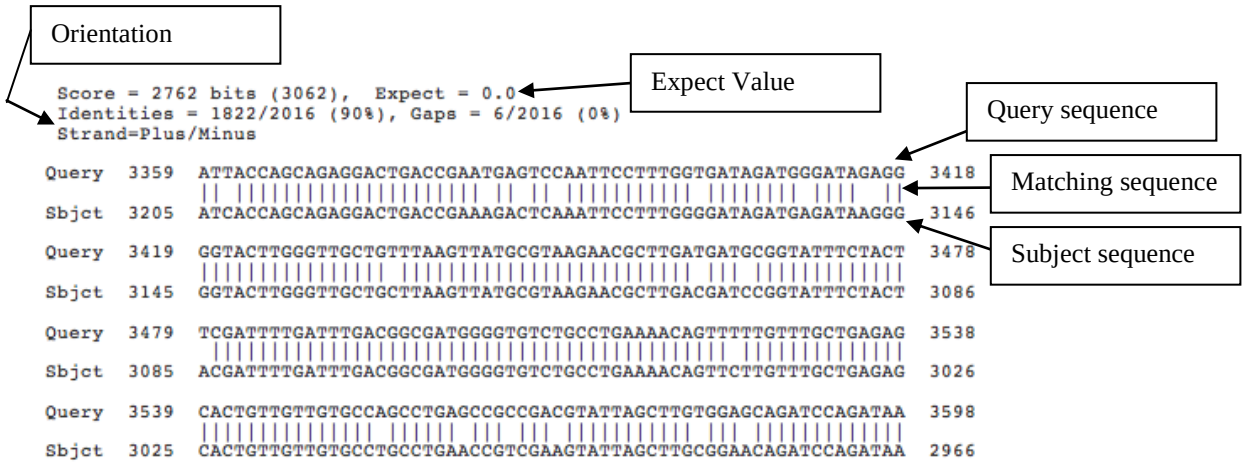


Figure 7. Important features of a typical BLAST alignment

The matching sequence consists of a combination of '|' characters and empty spaces. The '|' character denotes identical bases between the query and the subject sequences. An empty space in the matching sequence denotes a mismatched base. The '-' character in either the query or the subject sequence denotes a gap in the alignment (Figure 8).

Figure 8. Gaps in the alignment are represented by the '-' character

By default, NCBI BLAST automatically masks *low complexity sequences* (sequences with lots of the same bases) in the query sequence, e.g. BLAST will not try to match these regions to sequences in the database. These masked bases may appear as grey lower case letters (Figure 9) or as X's depending on the BLAST search settings.

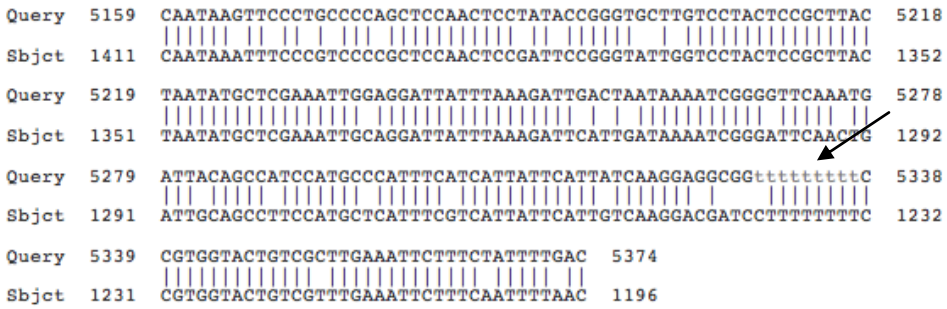


Figure 9. Bases masked by the low complexity filter appear as lower case grey letters by default

Now that we have a better understanding on the organization of a BLAST report, we are ready to interpret our blastn results. From the graphical overview and the description list (Figures 6a and 6b), we notice the top eight hits show much more significant alignments (with E-values of 0.0) compared to the other hits. These hits also have regions of homology across our entire query sequence (Figure 6a). Looking at the descriptions, it appears that our unknown *D. yakuba* genomic DNA has sequence homology to the gene *legless* in *D. melanogaster*. Comparing the descriptions for the top eight hits, we noticed that only the *D. melanogaster* hit has an accession number that begins with 'NM' while the accession number of the other hits begin with 'XM'. The primary difference between the two prefixes is the type of information available to support each of the Refseq mRNAs. Refseq mRNA records confirmed by experimental evidence have the prefix NM, while Refseq entries based solely on computational predictions have the prefix XM. Since we would generally prefer to base our inference on a gene model that has supporting experimental evidence, we will use the *D. melanogaster* model for our subsequent analysis.

From the blastn hit list, click on the score that corresponds to the *D. melanogaster* hit to navigate to the alignment section. We first notice that the *D. melanogaster* mRNA has a total length of 5359 bases (see top left of Fig. 6d), so the first question we would like to address is whether the entire mRNA aligns to our sequence. Looking at the alignments, we see that there are six alignment blocks relative to the *D. melanogaster* mRNA at postions 3205-1196, 5359-3390, 695-1, 983-695, 1200-986, and 3391-3194. Therefore, despite some minor overlaps and missing residues in the alignments, we can account for the majority of the mRNA in our genomic *D. yakuba* sequence. Note that all the alignments are also consistent with our query sequence: all alignment blocks are in the reverse orientation relative to the subject mRNA and show a high degree of sequence homology (sequence identity ranges from 75-92% at the nucleotide level). Reordering the coordinates of the alignment blocks with respect to our subject mRNA sequence (coordinates of the query sequence in parenthesis) produces the following list of alignment blocks: 1-695 (9850-9167), 695-983 (9107-8819), 986-1200 (8314-8100), 1196-3205 (5374-3359), 3194-3391 (2809-2606), and 3390-5359 (2552-580).

Detecting Coding Regions Using blastx

Since the Refseq mRNA sequence consists of both translated and untranslated regions (5' and 3' UTRs), the next step of our analysis is to identify the coding region in our sequence. In this case, we would like to search a *nucleotide* sequence against a *protein* database so we will set up a blastx search. Every mRNA in the Refseq RNA database has a corresponding sequence in the Refseq Protein database, so we will search our *D. yakuba* sequence against this Refseq Protein database. We now have all the information we need to setup our BLAST search (Figure 10):

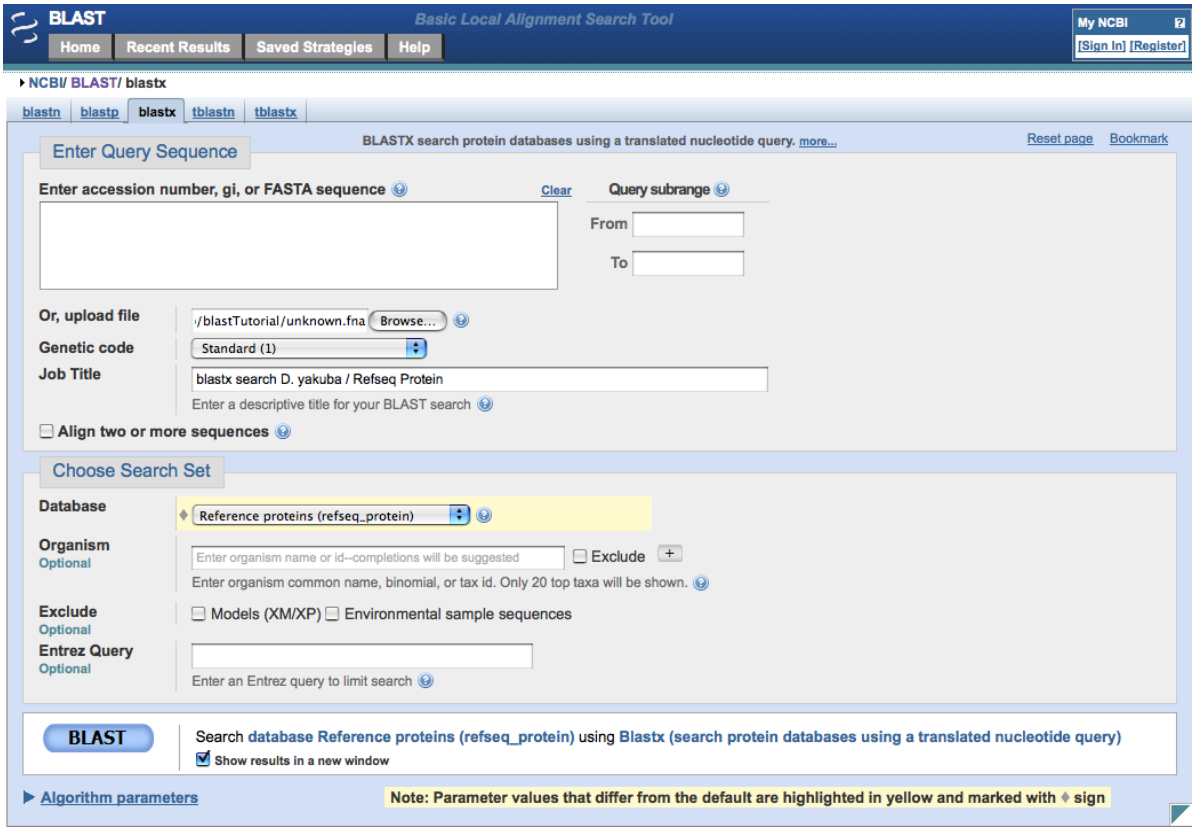


Figure 10. Setting up our blastx search against the NCBI Refseq Protein database

1. Navigate to the NCBI BLAST web server and click on "blastx".
2. Click on 'Browse' and select our sequence (unknown.fna).
3. Enter a Job Title "blastx search D. yakuba / Refseq Protein"
4. In the "Choose Search Set" section, change the database to "Reference proteins (refseq_protein)".
5. Check the box "Show results in a new window" next to the "BLAST" button
6. Click "BLAST". The graphical results of this search are shown on the next page in Figure 11.

The blastx report is similar to the blastn report. It has a graphical overview (Figure 11, next page), a list of significant blastx hits, and the alignments themselves. In addition to the two significant protein hits to the *D. melanogaster* and *D. yakuba* protein, we also see a number of significant hits to transposases in the region between 6000-8000 bp of our sequence. This is a clear indication that our sequence contains a type of repetitious element called a *transposable element*.

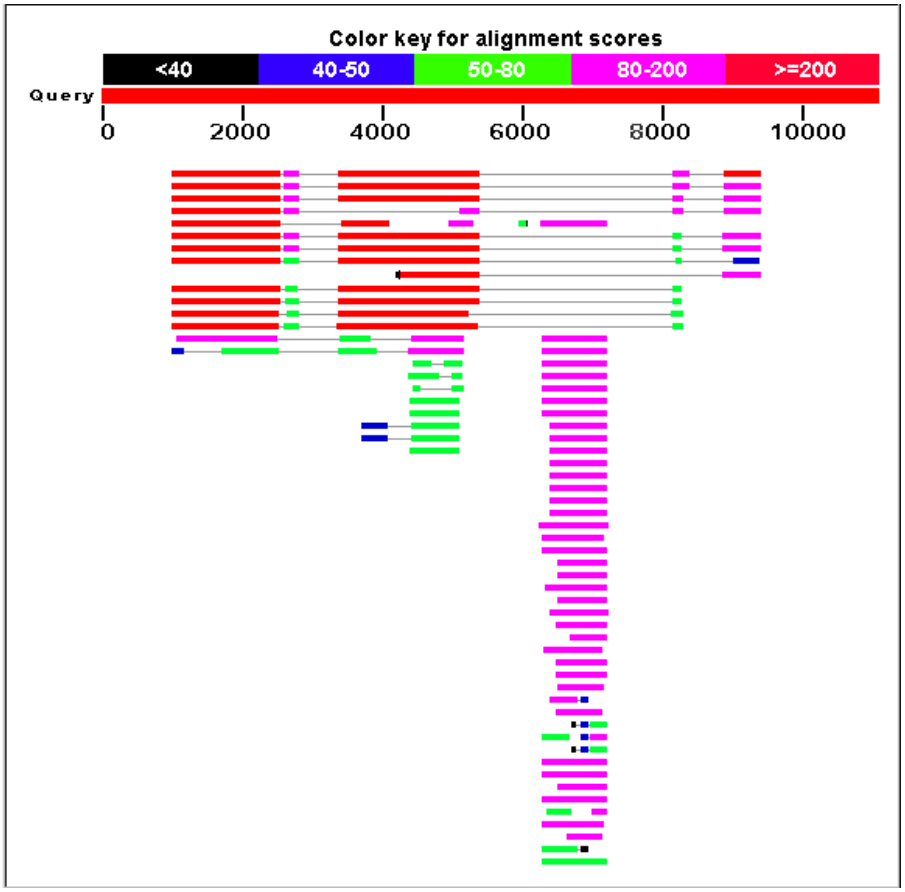


Figure 11. Additional blastx hits in the region between 6000-8000 bp in our sequence

It is possible to decrease the number of spurious hits in our BLAST reports by masking these elements prior to our BLAST search using a program called *RepeatMasker*. For now, we will simply focus on the hit to the *D. melanogaster* legless protein (Figure 12).

Sequences producing significant alignments:		Score (Bits)	E Value
ref|XP_002099599.1|	GE14515 [Drosophila yakuba]	1191	0.0
ref|NP_651922.1|	legless CG2041-PA [Drosophila melanogaster]	1063	0.0
ref|XP_001982729.1|	GG16448 [Drosophila erecta]	1060	0.0
ref|XP_002043673.1|	GM26781 [Drosophila sechellia]	739	0.0
ref|XP_002105752.1|	GD24382 [Drosophila simulans]	685	0.0
ref|XP_002027051.1|	GL18147 [Drosophila persimilis]	650	0.0
ref|XP_001382224.1|	GA15199 [Drosophila pseudoobscura pseudoo...	648	0.0
ref|XP_001966651.1|	GF23421 [Drosophila ananassae]	620	5e-175
ref|XP_002105753.1|	GD24381 [Drosophila simulans]	608	8e-172
ref|XP_002072634.1|	GK13708 [Drosophila willistoni]	508	5e-141

Figure 12. The top "NP" blastx hit shows that our sequence is similar to the legless protein.

We will analyze our blastx alignments in the same way we have analyzed the blastn report previously. Since blastx translates our input sequence in all 6 reading frames before comparing

our sequence with the protein database, there is an additional field 'Frame' for each alignment block. The frame can either be + or – and it corresponds to the relative orientation of our sequence compared to the protein. The frame also has a value that ranges from 1 to 3, which reflects the reading frame that produced the translated peptide sequence. Together, the relative orientation and the frame represent all 6 reading frames. In our protein alignments, a frame shift between alignment blocks would be a good indication of separate exons. There is also a new field called 'Positives' which corresponds to the number of amino acids that are either identical between the query and the subject sequence or have similar chemical properties (Figure 13).

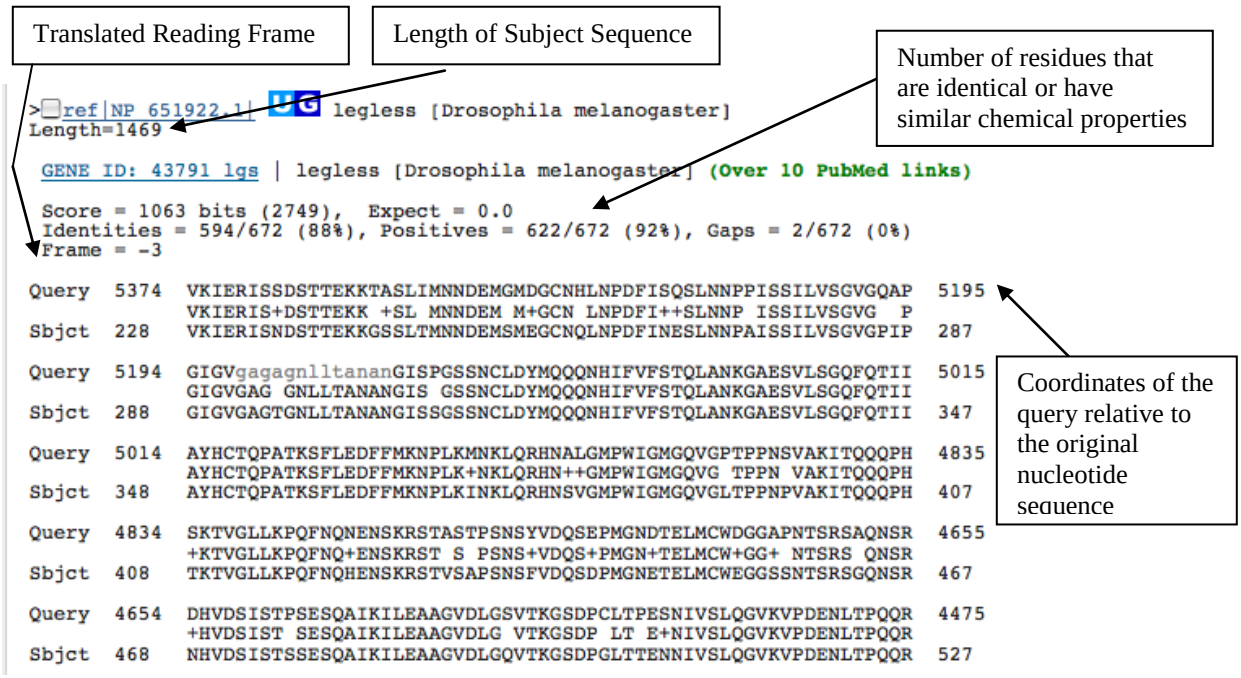


Figure 13. The key characteristics of a blastx alignment

Similar to the blastn alignment, each alignment block in our blastx report also consists of three lines: the query sequence, the matching sequence, and the subject sequence. Note that the query sequence has been translated into the corresponding amino acid sequence in the reading frame specified by the "Frame" field. The numerical coordinates at the ends of the Subject lines refer to the amino acid positions in the polypeptide from the database. However, the numerical coordinates of the Query sequence are still relative to our original (unknown) nucleotide sequence. Like our blastn alignment, the grey lower case residues in the query sequence indicate bases that are masked as low complexity.

There are some minor differences in the matching sequence of the blastn and blastx outputs. Similar to the blastn matching sequence, residues in the matching sequence represent amino acids that are identical in both the query and subject sequences. A space in the matching sequence represents amino acids that are completely different. In addition, the '+' character denotes amino acids that are different between the query and subject sequences but the two residues have similar chemical properties.

Our first question when investigating the blastx alignment with the *D. melanogaster legless* protein is whether we have matches to the full-length protein. We see from the "Length" field underneath the definition line that the *D. melanogaster* legless protein consists of 1469 amino acid residues. Looking at the alignment blocks, we see matches to the protein sequence at 1-158 (9344-8814), 148-229 (8332-8099), 228-897 (5374-3359), 890-959 (2821-2606), 959-1469 (2553-1018). In addition, the coordinates relative to our query sequence (in parentheses) are consistent with the results we have previously obtained with blastn. Based on both the blastn and blastx results, we can determine the approximate coordinates of the UTRs and the coding regions in our *D. yakuba* sequence. Hence it appears that we have a likely *D. yakuba* homolog of the *D. melanogaster legless* gene in our sequence.

However, there are a few problems with the alignment blocks. For example, looking at the alignment block that corresponds to the first 158 amino acids of the protein sequence (9344-8814 in our query sequence), we notice a large gap beginning at residue 61 (9167 in our query sequence) (Figure 14). Furthermore, the translation of the query in this region appears to contain a stop codon (the * character). A possible explanation for this observation is that blastx may have combined two exons in the same alignment and translated the intervening intron in frame.

```
 Score =  187 bits (475),  Expect = 9e-45
 Identities = 124/178 (69%), Positives = 132/178 (74%), Gaps = 21/178 (11%)
 Frame = -2

Query  9344  MLSTTMPRSPPQAQPQQNSDAS-TSASGSNPGVGIGNGISATNISSPKNLKNELFSTMSP  9168
             MLSTTMPRSP Q QPQ NSDAS TSASGSNPG  IGNG SA + SSPK L +E FST+SP
Sbjct  1     MLSTTMPRSPTQQQPQPNSDASSTSASGSNPGAAIGNGDSAASRSSPKTLNSEPFSTLSP  60

Query  9167  GKCYVLIFHCAEI*QLSMFTDQIKVTPDEGTEKSGLSTSDKaggvavgggGNISSEGPTM  8988
                                DQIK+TP+EGTEKSGLSTSDKA      G GN   EG TM
Sbjct  61    --------------------DQIKLTPEEGTEKSGLSTSDKAATGGAPGSGNNLPEGQTM  100

Query  8987  LRQNSSSSINSCLVAspqnssehsnssnvSGTVGLTQMVDCDEQSKKKKCSVKDEEGK  8814
             LRQNS+S+INSCLVASPQNSSEHSNSSNVS TVGLTQMVDCDEQSKK KCSVKDEE +
Sbjct  101   LRQNSTSTINSCLVASPQNSSEHSNSSNVSATVGLTQMVDCDEQSKKNKCSVKDEEAE  158
```

Figure 14. Large gap and stop codon in the blastx alignment block

Another potential problem with the alignments is the significant overlap of two adjacent alignment blocks; this occurs with blocks 1-158 and 148-229, and with blocks 228-897 and 890-959. However, looking at the beginning of the alignment block that spans 148-229 we notice the first 10 residues in the alignment block have a significantly worse alignment than the remaining residues (Figure 15). We also see a similar pattern in the block beginning at 890 compared to the block that ends with residue 897. Hence it is likely that blastx has overextended the alignments in both cases.

```
 Score = 94.4 bits (233),  Expect = 1e-16
 Identities = 69/82 (84%), Positives = 72/82 (87%), Gaps = 4/82 (4%)
 Frame = -3

Query  8332  HKCPI----SEICSNkakglaagggcgtgstsslTVKEEPTDVLGSLVNMKKEERENHSP  8165
             +KC +    +EI SNKAKG AAGGGC TGSTSSLTVKEEPTDVLGSLVNMKKEERENHSP
Sbjct  148   NKCSVKDEEAEISSNKAKGQAAGGGCETGSTSSLTVKEEPTDVLGSLVNMKKEERENHSP  207

Query  8164  TMSPVGFGSIGNAQDLSATPGK  8099
             TMSPVGFGSIGNAQD SATP K
Sbjct  208   TMSPVGFGSIGNAQDNSATPVK  229
```

Figure 15. The beginning of the alignment shows a much lower degree of sequence homology

Elucidating the Intron-Exon Boundaries with Gene Record Finder and bl2seq

Based on our previous blastn and blastx analyses, our current hypothesis is that we have identified the putative ortholog of the *legless* gene in our *D. yakuba* sequence. However, in order to build a complete gene model, we must resolve the apparent discrepancies in the alignments of our blastn and blastx output. Since the coding region is under strong selective pressure and is likely to be more conserved, our first step is to identify the coding regions of our putative gene.

To begin our more detailed analysis, we will perform a series of BLAST searches using the amino acid sequence of each coding exon in the *D. melanogaster* version of the *legless* gene. It will be extremely helpful to our annotation efforts if we can obtain the amino acid sequence that corresponds to each exon individually. Fortunately, individual exon sequences for *Drosophila melanogaster* can be easily obtained from the Gene Record Finder (Figure 16).

1. Navigate to the Gene Record Finder at http://gander.wustl.edu/~wilson/dmelgenerecord/index.html (GEP Home Page -> Projects -> Annotation Resources -> Gene Record Finder).
2. Type the FlyBase gene symbol "lgs" for legless in *D. melanogaster*, then press "Enter"

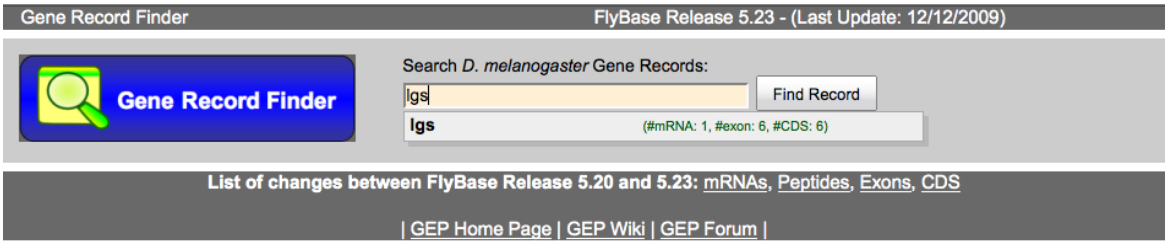


Figure 16. Gene Record Finder is available at http://gander.wustl.edu/~wilson/dmelgenerecord/index.html

From the autocomplete box, we noticed that there is only one isoform (#mRNA: 1) for the *legless* gene in *D. melanogaster*. The Gene Record Finder search result page is divided into three main sections (Figure 17). The 'Gene Details' section provides basic information about the gene (viewable by clicking on the 'FlyBase ID' link). The 'mRNA Details' section lists all the mRNA isoforms for that gene. The 'Transcript Details' and 'Polypeptide Details' sections contain the exon and coding sequences (CDS) usage tables, respectively. The 'Transcript Details' table lists the locations of each of the exons along the corresponding *D. melanogaster* chromosome, and the number of nucleotides in each ('Length'). Clicking on a row in the 'Transcript Details' table will allow you to retrieve the corresponding nucleotide sequence of that exon. The 'Polypeptide Details' table lists the number of amino acids encoded by each CDS, and clicking on a row in this table will display the corresponding amino acid sequence for that CDS. For multi-isoform genes (which *lgs* is not), the tables will change when particular isoforms are selected. You can also obtain all the possible exons or coding sequences for a gene or all the sequences for a particular isoform by using the 'Export All…' tabs next to 'Options.'

Before going any further, click on the 'View in GBrowse' link, which will take you directly to FlyBase. This window will provide you with helpful information about your gene, including the physical location of your gene on a particular chromosome in *D. melanogaster* and graphical overviews of the gene location, mRNA (Transcript) structure and the CDS. The thin lines between the colored boxes in the 'Transcript' and 'CDS'

views indicate the position of the introns. Note that for most eukaryotic genes, the CDS are a subset of the 'Transcript' blocks. The gray blocks at the ends of the Transcripts represent the 5-′untranslated and 3′-untranslated regions (UTRs) of the mature mRNA (e.g., exon sequences at the beginning and end, respectively, of a mature mRNA molecule that do not contain any protein-coding information).

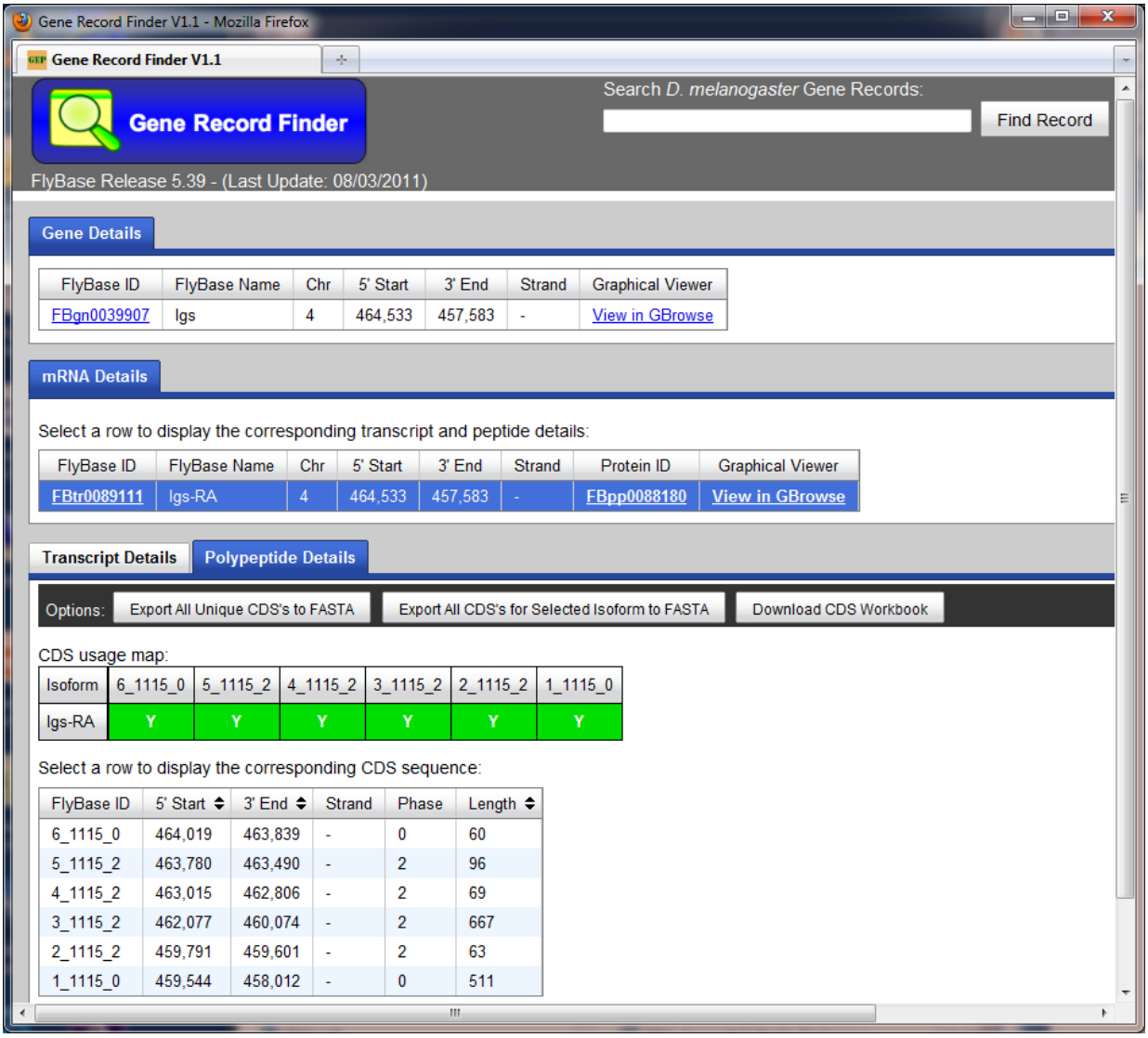


Figure 17. The Gene Record Finder results page allows you to retrieve the sequences for individual coding sequences.

The first problem in our blastx results was the stop codon in the alignment block that spans from 1-158 of the translated protein sequence (see Fig. 14). To determine the locations of the coding exons, we should compare the individual exons with our sequence. We will use the program *bl2seq* to align the two sequences. Since we are using a protein sequence (i.e. the *D. melanogaster* CDS) as the Query in a comparison to a translated nucleotide sequence (our unknown *D. yakuba* DNA sequence), we will use the tblastn program for our search. In order to prevent BLAST from masking low complexity regions in our protein, we should also turn off the low complexity filter.

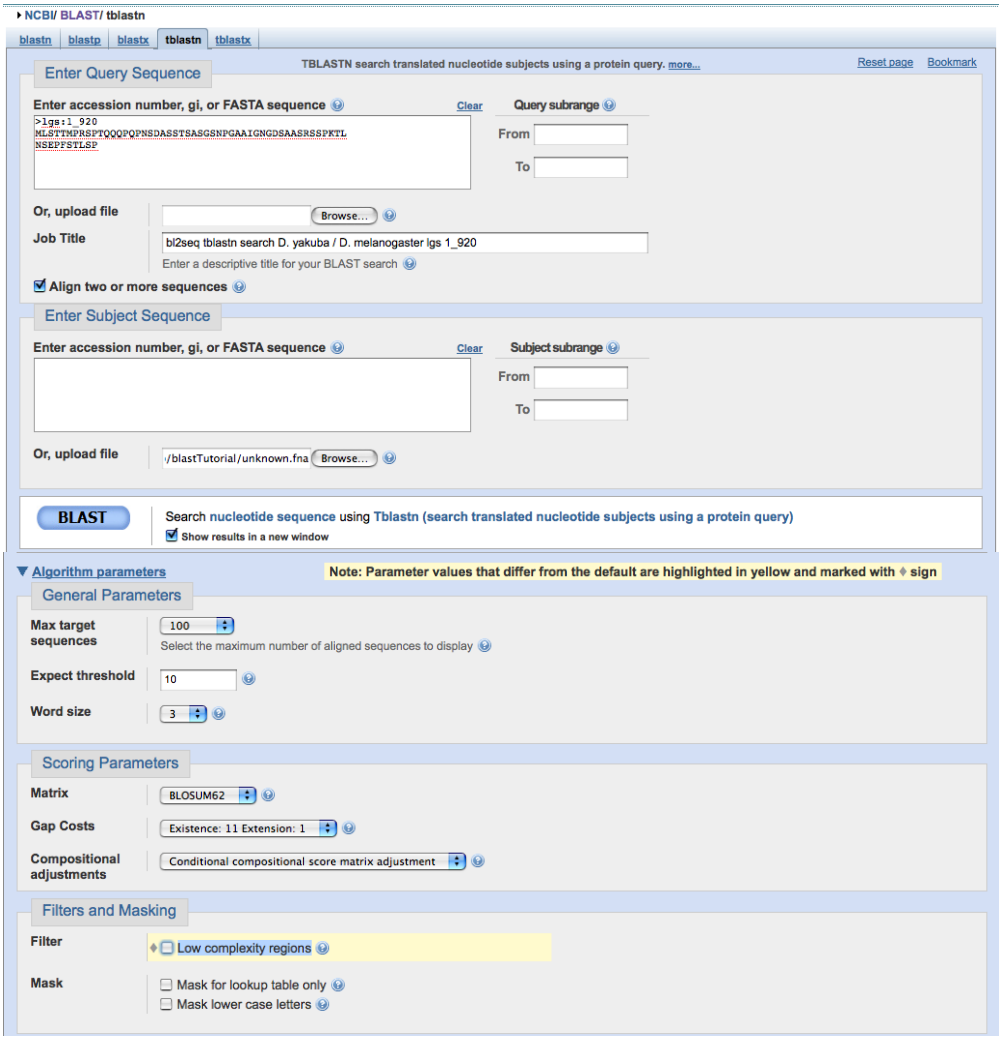


Figure 18. BLAST 2 sequences (bl2seq) using the tblastn program, aligning the first coding exon with our sequence with the low complexity filter turned off.

Proceed as follows to do the tblastn search:

1. From the Gene Record Finder record for lgs, select the first row of the CDS sequence table with the FlyBase ID "6_1115_0".
2. Select the content of the sequence window and copy it to the clipboard
3. Open a new browser window and navigate to the NCBI BLAST web server
4. Select "Align two (or more) sequences' using BLAST (bl2seq)" under Specialized BLAST. Click on the tab labeled "tblastn" in the next window.
5. Paste the amino acid sequence for the 6_1115_0 CDS into the "Query Sequence" text box (Figure 18)
6. For the "Subject Sequence" section, click on "Browse" and select the unknown nucleotide sequence (unknown.fna).
7. Enter a Job Title "tblastn search D. yakuba / D. melanogaster lgs 6_1115_0"
8. Expand the "Algorithm parameters" section and **uncheck** "Low complexity regions" under "Filters and Masking"
9. Check the box "Show results in a new window"
10. Click "BLAST".

From the bl2seq output, we see that the first coding sequence has a length of 60 amino acids and corresponds to 9344-9168 in our unknown (Subject) sequence (Figure 19).

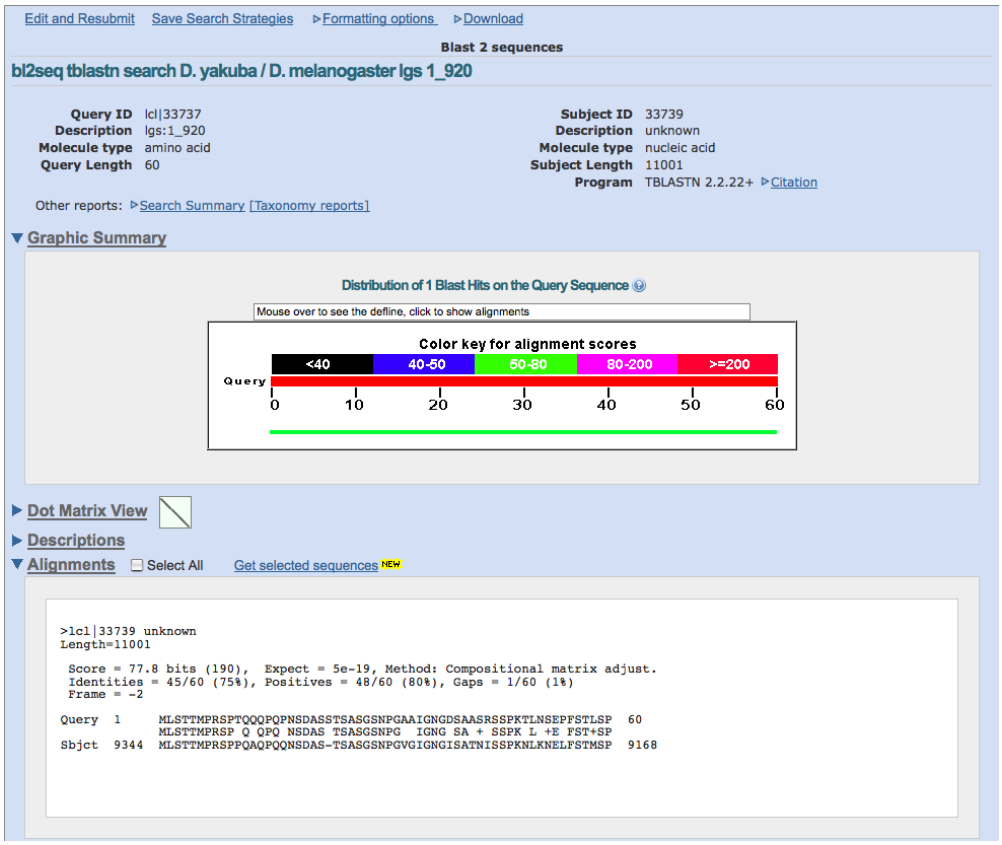


Figure 19. bl2seq results that shows the tblastn alignment of the first coding exon with our sequence

Use the same strategy to map the remaining coding sequences, record the results in the table below and answer the questions on the next page.

CDS # (Number of complete codons)	Protein Alignment (Start-End)	Our Sequence Alignment (Start-End)
6 (60)	1-60	9344-9168
5 (96)	1-96	
4 (69)	1-69	
3 (667)	1-667	
2 (63)	1-63	
1 (510)	1-510	

The results of this bl2seq analysis suggest that you are able to account for all six of the coding exons of the *D. melanogaster legless* gene in the *D. yakuba* unknown sequence. For your own annotation projects, you should save the bl2seq alignments for each CDS **(by pasting them into a Word document)** as an early step in constructing your gene model, so that you can revisit the results later.

Questions

1. Return to p. 110 and recall the problem with the first exon, which spanned from 9344-8814, in the initial blastx search (which compared the translated *D. yakuba* unknown DNA sequence to the RefSeq protein database). What do the results in the table on p. 114 indicate about a possible solution to this problem?

2. What other problems did you encounter with the blastx search results on p. 110, which were resolved by the results in the table on p. 114?

3. Examine the table on p. 114. Were all of the alignments full-length, e.g. were all of the amino acids at the beginning and end of each *D. melanogster* (Query) CDS accounted for in the tblastn alignments?

4. Which of the CDS were missing amino acids (aa) at the end(s) of the tblastn alignments and which amino acids were missing in each (by number order, such as aa 1-5, 665-667, etc.)?

   Note: it will be very important to keep track of whether or not each CDS alignment is full length when doing your individual annotation projects.

Conclusion

In this tutorial, you have used various BLAST programs and the Gene Record Finder to identify and characterize the coding regions of a putative gene in a piece of recently-sequenced genomic DNA from *D. yakuba*. Although the CDS-by-CDS tblastn search approach provides a finer level of resolution than the blastx/RefSeq protein search, your answers to #3 and #4 above suggest that you have not yet generated a *complete* gene model for this putative *D. yakuba* gene. In the next lab, you will learn how to *precisely* map the exact intron and exon boundaries using the UCSC Genome Browser. To check your understanding to this point, there will be an *on-line quiz for you to complete prior to the start of your next scheduled lab.* ☺

Blank Page