An Introduction to NCBI BLAST

Prerequisites:
Detecting and Interpreting Genetic Homology: Lecture Notes on Alignment

Resources:
The BLAST web server is available at https://blast.ncbi.nlm.nih.gov/Blast.cgi
Gene Record Finder is available through the GEP web site (http://gep.wustl.edu → Projects → Annotation Resources → Gene Record Finder)

Files for this Tutorial:
All of the files require for this tutorial is compressed into a single archive [Introduction_NCBI_BLAST.zip]

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 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 a sequence (unknown.fna in the tutorial package) from the Drosophila yakuba genome.

The NCBI BLAST web interface
Before we begin the analysis, 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 https://blast.ncbi.nlm.nih.gov/Blast.cgi. In this tutorial, we will only use a few of the tools available on the NCBI BLAST web site. To learn about the more advanced options available (such as setting up My NCBI accounts), click on the “Help” link on the main navigation bar to access the documentations for NCBI BLAST (Figure 1).

Figure 1. Click on the “Help” link to learn more about the NCBI BLAST web interface.
All of the NCBI BLAST pages have the same header with four links:

<table>
<thead>
<tr>
<th>Links</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home</td>
<td>Link to the NCBI BLAST home page</td>
</tr>
<tr>
<td>Recent Results</td>
<td>Link to results of the BLAST searches you have run previously</td>
</tr>
<tr>
<td>Saved Strategies</td>
<td>NCBI BLAST search parameters you have previously saved to your My NCBI account</td>
</tr>
<tr>
<td>Help</td>
<td>Documentation for NCBI BLAST</td>
</tr>
</tbody>
</table>

Besides the main toolbar, there are two other sections of the NCBI BLAST web interface that are of interests: the “Web BLAST” section contains links to the common BLAST programs and the “Specialized searches” section contains links to additional tools for performing sequence searches (e.g., use CD-search to identify conserved domains within a query sequence). The type of BLAST search you wish to perform will depend primarily on the type of query sequence and the database you would like to search.

Four of the five common BLAST programs are available through the “Web BLAST” section of the NCBI BLAST home page (Figure 2, top). The program tblastx, which translates the nucleotide query and nucleotide database when it performs the sequence comparisons, 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 (Figure 2, bottom).

The basic BLAST programs are summarized below:

<table>
<thead>
<tr>
<th>BLAST program</th>
<th>Query</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotide BLAST (blastn)</td>
<td>Nucleotide</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>Protein BLAST (blastp)</td>
<td>Protein</td>
<td>Protein</td>
</tr>
<tr>
<td>blastx</td>
<td>Translated Nucleotide</td>
<td>Protein</td>
</tr>
<tr>
<td>tblastn</td>
<td>Protein</td>
<td>Translated Nucleotide</td>
</tr>
<tr>
<td>tblastx</td>
<td>Translated Nucleotide</td>
<td>Translated Nucleotide</td>
</tr>
</tbody>
</table>
Instead of searching a query sequence against sequences in a database, you can also align two (or more) sequences by selecting the “Align two or more sequences” checkbox at the bottom of the “Enter Query Sequence” section (Figure 3). This feature is also known as BLAST 2 Sequences (bl2seq).

![Figure 3. Select the “Align two or more sequences” checkbox to compare a query sequence against a subject sequence instead of a BLAST database.](image)

**Detecting sequence homology to mRNA using blastn**

In this tutorial, we will characterize an unknown genomic sequence (*unknown.fna*) and determine if it has sequence similarity to any known genes. One strategy we can use is to search for sequence similarity to mRNA sequences in the NCBI Reference Sequence (RefSeq) database.

When we set up a BLAST search, there are three basic decisions we must make: the BLAST program we want to use, the query sequence we want to annotate, and the database we want to search. In addition, we can change several optional parameters (such as the Expect threshold and low complexity filters) in order to modify the behavior of BLAST.

In this case, we will set up our BLAST search using mostly default parameters. We will use the blastn program to search our sequence (query) against the NCBI Reference Sequence (RefSeq) RNA database (Figure 4).

1. Navigate to the NCBI BLAST home page and click on the “Nucleotide BLAST” image under the “Web BLAST” section
2. Click on the “Browse” or the “Choose File” button and select the file with the unknown sequence (*unknown.fna*)
3. Enter the Job Title “blastn search D. yakuba / RefSeq RNA”
4. In the “Choose Search Set” section, change the database to “Reference RNA sequences (refseq_rna)”
5. Under “Program Selection”, select “Somewhat similar sequences (blastn)”
6. Check the box “Show results in a new window” next to the “BLAST” button
7. Click “BLAST”
Figure 4. Setting up our blastn search of the unknown sequence against the NCBI RefSeq RNA database

Note: the blastn search may take a few minutes to complete when the NCBI web server is busy (Figure 5).

Figure 5. Waiting for the blastn search results

Once the search is complete, a new web page will appear with the BLAST report. For teaching purposes, the BLAST output is available in the file ‘blastnInitial.html’ in the tutorial package.
The default BLAST report consists of three main sections: a graphical overview, a table of BLAST hits, and the alignments for each BLAST hit. We will go through each of these sections in order to interpret our blastn output:

I. **Graphical Overview**
The BLAST output begins with a summary of the parameters used in the BLAST search (e.g., the lengths of the query and subject sequences). It then displays a graphical overview of the alignments of all of the BLAST matches to our query sequence (Figure 6a). The boxes correspond to regions in the query that have sequence similarity to the sequences in the database. The color of the box corresponds to the score, where hits with higher scores are more significant. When you move your mouse over a BLAST hit, the title of the subject sequence will appear in a tooltip. Click on the color box and then click on the “Alignment” link to jump to the alignments associated with that BLAST hit.

![Figure 6a. A graphical overview of all the Refseq mRNA blastn hits for our query sequence.](image)

II. **List of Significant BLAST Hits**
Scrolling further down the blastn output, we find a summary table in the “Descriptions” section that lists all the sequences in the database that have significant sequence homology with our sequence (Figure 6b). By default, the results are sorted by their Expect value (E value) in ascending order, where lower E values denote more significant hits. You can click on the column headers to sort the results by the other columns.
Figure 6b. List of blastn hits that produce significant alignments with our query sequence.

Clicking on the accession number in the table will bring up a new page with the GenBank record of the sequence. Clicking on the description of the hit will bring us to the corresponding alignment in the BLAST output. Alternatively, you can click on the “Alignments” button on the main toolbar to jump to the first alignment.

In addition to reviewing the records for individual sequences, you can also review multiple sequence records by selecting the checkbox next to each match. You can then use the buttons on the main toolbar to download all the corresponding GenBank records, FASTA sequences, and to see a Graphical overview of all the alignments (Figure 6c).

III. List of Alignments
Following the table of BLAST hits is a section that shows the alignments for each BLAST hit. The sequence alignments show us how well our query sequence match the subject sequence in the database. Because we will rely on sequence alignments heavily in our annotation efforts, we will examine this alignment section more closely.

Alignments to different subject sequences in the database are separated by a gray toolbar that contains options to manipulate the alignment results and to retrieve additional information for that specific BLAST hit (Figure 6d). For example, we can use the “Download” drop-down menu on this toolbar to obtain the FASTA sequence or the GenBank record for a specific hit.
In addition, we can click on the “Graphics” link to examine the location of each alignment block relative to the subject sequence (Figure 6e). We can use the navigation links at the right side of the toolbar to quickly navigate to the next or the previous BLAST hit.

As its name suggests, BLAST is designed to identify local regions of sequence similarity. This means that BLAST might report multiple distinct regions of sequence similarity when we align a query against a subject sequence in a database. For example, if we were to align a processed mRNA sequence to a genomic sequence, we would expect to see multiple alignment blocks (many of which correspond to transcribed exons) in our BLAST output. Each alignment block demarcates a local region of similarity between the query and the subject sequences. Regions of the genomic sequence without significant alignments that fall between these alignment blocks would likely correspond to intronic sequences.

The “Number of Matches” field beneath the name of the sequence shows the number of alignment blocks identified by blastn. For example, the fifth blastn hit contains 6 different alignment blocks to the subject sequence — the legless mRNA from D. melanogaster (Figure 6f). Each alignment block represents a region of the D. melanogaster legless gene that shows sequence homology with our genomic sequence from D. yakuba.
You can use the “Sort by” drop-down box (red arrow in Figure 6f) on the toolbar above each BLAST hit to sort the alignment blocks based on different criteria (e.g., by E value, query start position, subject start position). Each alignment block begins with a line that has the following format: “Range #:start to end” where # is the alignment block number. You can use the “Next Match” and “Previous Match” links to navigate to the different alignment blocks within the same BLAST hit.

Depending on the database you use, there might be additional links to other parts of NCBI listed under the “Related Information” panel next to the sequence alignments. For example, there are links to Entrez Gene, UniGene, GEO Profiles, and the Genome Data Viewer for the fifth hit “Drosophila melanogaster legless (lgs), mRNA” (NM_143665.4; Figure 6g). Entrez Gene provides us with an overview of the gene and links to literature references. UniGene provides information on the cDNAs and expressed sequence tags (ESTs) that are associated with the gene. GEO Profiles allow us to access expression data associated with the gene. The Genome Data Viewer allow us to view the BLAST alignments in a genome browser with other evidence tracks (e.g., gene annotations, RNA-Seq data, repeats).

![Figure 6g. You can learn more about the blastn match using the links under the “Related Information” section.](image)

What about the alignments themselves? Each alignment block begins with a summary, including the Expect value (i.e. 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 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. The key characteristics of a typical BLAST alignment.](image)
The - character in either the query or the subject sequence denotes a gap in the alignment (Figure 8).

By default, NCBI BLAST automatically masks low complexity sequences in the query sequence. Depending on your BLAST search settings, these masked bases may appear as either grey lowercase letters (Figure 9) or as X’s. The matching sequence consists of a combination of | and empty spaces, where | denotes a matching base between the query and subject sequences and the empty space denotes a mismatched base.

Now that we have a better understanding of how the BLAST report is organized, we are ready to interpret the blastn results. From the graphical overview and the description table (Figures 6a and 6b), we noticed that the top 23 hits are much more significant (with E values of 0.0) than the rest of the blastn hits. Most of these top hits contain regions of sequence similarity that span the entire length of the query sequence (Figure 6a). Looking at the descriptions and the corresponding GenBank records, it appears that these blastn hits correspond to the gene legless (also known as BCL9) in different Drosophila species. Among these significant matches, only the \textit{D. melanogaster} hit has an accession number that begins with the prefix “NM_” (whereas the accession numbers for the hits to the other Drosophila species have the prefix “XM_”). The main difference between these two prefixes is the type of information available to support the RefSeq mRNAs. The “NM_” prefix indicates that the RefSeq mRNA record is supported by experimental evidence, whereas the “XM_” prefix indicates that the record is based solely on computational predictions. Because we would prefer to base our inferences on a gene model that is supported by experimental evidence, we will use the \textit{D. melanogaster} model in this analysis.

From the blastn hit list, click on the description that corresponds to the \textit{D. melanogaster} hit to jump to the alignment section. Our analysis above has shown that there are six alignment blocks. We also notice that the \textit{D. melanogaster} mRNA has a total length of 5357 bases, so the first question we would like to address is whether the entire mRNA aligns to our sequence.
To address this question, we will examine the subject coordinates of the alignment blocks from the *D. melanogaster legless* mRNA. We find that these blocks span from 3209-1200, 5357-3394, 699-2, 987-699, 1204-990, and 3395-3198. Re-ordering the coordinates of the alignment blocks with respect to our subject sequence produces the following list of alignments (coordinates of the query sequence are in parenthesis): 2-699 (9853-9167), 699-987 (9107-8819), 990-1204 (8314-8100), 1200-3209 (5374-3359), 3198-3395 (2809-2606), and 3394-5357 (2552-586).

Despite some minor overlaps and missing bases, we can account for most of the mRNA sequence in this collection of alignments. Note that all the alignment blocks are collinear with respect to our query sequence (i.e. all the alignment blocks are in the reverse orientation relative to the subject mRNA) and show a high degree of sequence similarity (with sequence identity that ranges from 75–92% at the nucleotide level).

**Detecting Coding Regions Using blastx**

Because the RefSeq mRNA sequence consists of both translated and untranslated regions (i.e. 5’ and 3’ UTRs), the next step in our analysis is to identify the coding region in our sequence. 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 setup the blastx search.

1. Navigate to the NCBI BLAST home page and click on the “blastx” image
2. Click on the “Browse” or the “Choose File” button and select our sequence (unknown.fna).
3. Enter the 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” (Figure 10)

![Figure 10. Configure our blastx search of the unknown sequence against the NCBI RefSeq Protein database.](image-url)
For teaching purposes, the blastx search result is available in the file ‘blastxRefSeqProtein.html’ in the tutorial package. (Note that this blastx search can take several minutes to complete.)

The blastx report is similar to the blastn report. It consists of a graphical overview, a list of significant blastx hits, and a list of alignments. In addition to the highly significant hits to the legless protein in *D. melanogaster* and the homologous protein in the other Drosophila species, we also see a few significant hits to transposases in the region between 6000-8000 bp of our sequence (Figure 11).

![Figure 11. Multiple blastx hits in the region between 6000-8000 bp in our sequence.](image1)

These hits suggest our sequence contains a type of repetitious element called a transposable element. In future tutorials, we will learn how we can reduce the number of spurious hits in our BLAST reports by masking these elements prior to performing the BLAST search. For now, we will ignore these additional matches and focus on the best manually curated RefSeq hit — the *D. melanogaster* legless protein (NP_651922.1; Figure 12).

![Figure 12. The blastx result shows that our sequence is very similar to the *D. melanogaster* legless protein.](image2)
Figure 13. The key characteristics of a blastx alignment.

We can analyze the blastx alignments in the same way that we have previously analyzed the blastn report. However, because blastx translates the input sequence in all 6 reading frames before comparing our sequence with the protein database, there is an additional “Frame” field in each alignment block. The frame begins with either + or -, which corresponds to the relative orientation of our sequence compared to the protein. The number following the relative orientation in the frame field ranges from 1 to 3, which reflects the reading frame that produces the translated peptide sequence. Collectively, the relative orientation and the number can be used to represent all 6 reading frames. A frame shift between two alignment blocks in the blastx match often indicates that the two alignment blocks correspond to different coding exons. 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 (Figure 13).

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. However, the coordinates of the query sequence are still relative to the original nucleotide sequence. Like our blastn alignment, the grey lowercase residues in the query sequence correspond to low complexity sequences that were masked by BLAST.

There are some minor differences in the matching sequence of the blastn and blastx outputs. Residues in the matching sequence represent amino acids that are identical between the query and subject sequences. The “+” character denotes amino acids that are different between the query and subject but these different amino acids have similar chemical properties. A space indicates that the two aligned amino acid in the query and subject are different and they have different chemical properties.

When investigating the blastx alignment with the *D. melanogaster* legless protein, the first question is whether there are matches to the entire legless protein. We see from the “Length”
field underneath the sequence name that the *D. melanogaster* legless protein has 1469 residues. Sorting the alignment blocks by the subject start position, we see matches to the protein sequence at 1-158 (9344-8814), 148-229 (8332-8099), 228-897 (5374-3359), 888-959 (2827-2606), and 959-1469 (2553-1018). In addition, the coordinates relative to our query sequence (in parenthesis) are consistent with the results from our previous blastn search. 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 our *D. yakuba* sequence contains an ortholog of the *D. melanogaster* legless gene.

While the alignment generally looks good, there are a few problems with some of the blastx alignment blocks. Looking at the alignment block that corresponds to the first 158 amino acids of the protein sequence (9344-8814 in our query sequence), we noticed a large gap beginning at residue 61 (9167 in our query sequence) (Figure 14). Furthermore, the translation of the query in this region contains a stop codon (the * character). One possible explanation for the stop codon is that blastx might have combined two separate exons into the same alignment. If that were the case, the intron between the two exons would also be translated by blastx.

Another problem with the alignments is the substantial amount of overlap between two adjacent alignment blocks; this occurs with blocks 1-158 and 148-229, and with blocks 228-897 and 888-959. However, examination of the beginning of the alignment block that spans from 148-229 shows that the first 10 residues in the alignment block have much weaker sequence similarity than the rest of the residues in the alignment (Figure 15). We also see a similar pattern in the alignment block beginning at 888 compared to the block that ends at residue 897. Hence our observations suggest that blastx might have over-extended the alignments in both cases.
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 \textit{legless} gene in our \textit{D. yakuba} sequence. However, in order to construct a complete gene model, we must resolve the discrepancies in the alignments of our blastn and blastx output. Because the coding region is under strong selective pressure and is likely to be more conserved than other regions of the genome, our first step is to identify the coding regions of our putative gene.

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

1. Navigate to the Gene Record Finder (http://gep.wustl.edu \rightarrow Projects \rightarrow Annotation Resources \rightarrow Gene Record Finder)
2. Type “lgs” (the official FlyBase symbol for the \textit{legless} gene) in the textbox and click on the “Find Record” button

Figure 16. Access the Gene Record Finder from the GEP home page

In the “mRNA Details” section of the gene report, we notice that there is only one isoform of the \textit{legless} gene in \textit{D. melanogaster} (lgs-RA, the A isoform of lgs). We can access all the transcribed exons through the “Transcript Details” tab and all the translated exons through the “Polypeptide Details” tab (Figure 17).
To retrieve the amino acid sequence for each coding exon (CDS), click on the row that corresponds to the coding exon in the CDS table (Figure 18).

The first problem in our blastx results is the stop codon in the alignment block that spans from 1-158 of the translated protein sequence. To determine the locations of the coding exons, we will perform BLAST searches to compare the individual exons with our sequence. Because we are comparing a protein sequence against a nucleotide sequence, we will use the tblastn program for our search. 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.
1. Select the first CDS \((1\_9517\_0)\) from the Gene Record Finder CDS table and copy the sequence to the clipboard.
2. Open a new web browser tab and navigate to the NCBI BLAST home page; click on the “tblastn” image under the “Web BLAST” section.
3. Select the checkbox “Align two or more sequences” under the “Enter Query Sequence” section.
4. Paste the CDS sequence for \(1\_9517\_0\) into the “Enter Query Sequence” field.
5. For the “Subject Sequence”, click on the “Browse” or the “Choose File” button and select our unknown sequence (unknown.fna).
6. Click on the “Algorithm parameters” link to expand this section. **Verify that the “Word size” parameter is set to 3.**
7. Change the “Compositional adjustments” field to “No adjustment” under the “Scoring Parameters” section.
8. Uncheck the box “Low complexity regions” under “Filters and Masking.”
9. Click “BLAST” (Figure 19).

Figure 19. Use the “Align two or more sequences” feature with tblastn to align the first coding exon of \(lgs\) against our sequence with the low complexity filter turned off and no compositional adjustment.
For teaching purposes, the BLAST output is available in the file ‘bl2lgsExon1_tbn.html’ in the tutorial package.

From the bl2seq output, we see that the first coding exon has a length of 60 amino acids and corresponds to 9344-9168 in our query sequence (Figure 20). We can use the same strategy to map the rest of the coding exons:

<table>
<thead>
<tr>
<th>Exon # (Number of complete codons)</th>
<th>Protein Alignment (Start-End)</th>
<th>Our Sequence Alignment (Start-End)</th>
<th>Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (60)</td>
<td>1-60</td>
<td>9344-9168</td>
<td>-2</td>
</tr>
<tr>
<td>2 (96)</td>
<td>1-95</td>
<td>9104-8820</td>
<td>-2</td>
</tr>
<tr>
<td>3 (69)</td>
<td>1-69</td>
<td>8311-8105</td>
<td>-3</td>
</tr>
<tr>
<td>4 (667)</td>
<td>1-667</td>
<td>5371-3365</td>
<td>-3</td>
</tr>
<tr>
<td>5 (63)</td>
<td>1-63</td>
<td>2800-2606</td>
<td>-3</td>
</tr>
<tr>
<td>6 (511)</td>
<td>1-511</td>
<td>2550-1015</td>
<td>-1</td>
</tr>
</tbody>
</table>
The results of our exon-by-exon bl2seq analyses suggest we can account for all of the coding exons of the *D. melanogaster* legless gene in our sequence. Furthermore, we were able to resolve the problem with the first exon in our initial blastx search: the alignment block that spans from 9344-8814 actually consists of two separate exons, one that spans approximately from 9344-9168 and the other from 9104-8820.

For your own annotation projects, it will be advantageous to save the bl2seq outputs as you construct your gene model so that you can revisit the results later. Note that we have yet to generate a complete gene model for this putative gene in our *D. yakuba* sequence. In future tutorials, we will learn how we can use the UCSC Genome Browser to identify intron splice sites and to define the exact exon boundaries.

We could use the same strategy to map the locations of the putative UTRs using bl2seq with the blastn program. We can then compare the exon-by-exon search results with our initial blastn search against the RefSeq RNA database. To retrieve the exon sequences for each transcribed exon, select the “Transcript Details” tab in the Gene Record Finder and then click on a row in the exon table (Figure 21). Mapping of the untranslated exons is left as an exercise for the reader.

**Figure 21.** The “Transcript Details” tab of the Gene Record Finder shows the list of transcribed exons for the *D. melanogaster* gene legless

**Conclusion**

In this tutorial, we have used various BLAST programs to identify and characterize a putative gene in a genomic sequence from *D. yakuba*. You are now ready to tackle some of the more challenging problems in the BLAST homework exercises 1 and 2 (available on the GEP web site at [http://gep.wustl.edu](http://gep.wustl.edu)).