What is BLAST?

- BLAST stands for Basic Local Alignment Search Tool
- Why is BLAST popular?
  - Good balance of sensitivity and speed
  - Reliability
  - Flexibility

Where Can I run BLAST?

- We will use two sites:
    databases updated constantly (daily); very slow at times
  - FlyBase (http://flybase.org/blast/)
    databases for many Drosophila species

BLAST output

- Graphical overview, showing alignment blocks as bars = regions of sequence similarity between Query (top) and database sequences
- List of Sequences with scores (see next slide)
  - Raw score, higher is better (length dependent)
  - Expect value, smaller is better
    (length and database size independent)
- List of alignments

Calculating alignment scores

The raw score $S$ for an alignment is calculated by summing the scores for each aligned position and the scores for gaps. In this figure, a DNA alignment is shown.

$$\text{score} = \text{Max}(S)$$

$E$ value (Expectation value). The number of different alignments with scores equivalent to or better than $S$ that are expected to occur in a database search by chance. The lower the $E$ value, the more significant the score.

The Databases

- Genbank m/n (protein and nucleotide versions)
  Non-redundant large databases (compile & remove duplicates)
  Anyone can submit, you can call your sequence anything
  Quality low; names can be meaningless
- EST (Expressed Sequence Tags) databases
  Short single reads of cDNA clones
  Other short single reads
  High error rates
- Swissprot
  Curated from literature
  REAL proteins; REAL functions; small;
  Genomic Databases
  Human, Mouse, Drosophila, Arabidopsis, etc
  NCBI, species-specific web pages

BLAST Protocols

- The most common BLAST search includes five protocols:

<table>
<thead>
<tr>
<th>Program</th>
<th>Database</th>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLASTN</td>
<td>Nucleotide</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>BLASTP</td>
<td>Protein</td>
<td>Protein</td>
</tr>
<tr>
<td>BLASTX</td>
<td>Protein</td>
<td>Nt $\rightarrow$ Protein</td>
</tr>
<tr>
<td>TBLASTN</td>
<td>Nt $\rightarrow$ Protein</td>
<td>Protein</td>
</tr>
<tr>
<td>TBLASTX</td>
<td>Nt $\rightarrow$ Protein</td>
<td>Nt $\rightarrow$ Protein</td>
</tr>
</tbody>
</table>
### BLASTN
- The query is a nucleotide sequence.
- The database is a nucleotide database.
- No conversion is done on the query or database.
- DNA vs. DNA homology.
- Mapping oligos to a genome.
- Cross-species sequence exploration.
- Annotating genomic DNA with ESTs.

### BLASTP
- The query is an amino acid sequence.
- The database is an amino acid database.
- No conversion is done on the query or database.
- Protein vs. Protein homology.
- Protein function exploration.
- Novel gene \( \rightarrow \) makes parameters more sensitive.

### BLASTX
- The query is a nucleotide sequence.
- The database is an amino acid database.
- All six reading frames are translated on the query and used to search the database.
- Coding nucleotide seq vs. Protein homology.
- Gene finding in genomic DNA.
- Annotating ESTs (and Shotgun Sequence).

### TBLASTN
- The query is an amino acid sequence.
- The database is a nucleotide database.
- All six frames are translated in the database and searched with the protein sequence.
- Protein vs. Coding Nucleotide DB homology.
- Mapping a protein to a genome.
- Mining ESTs (Shotgun DNA) for protein similarities.

### TBLASTX
- The query is a nucleotide sequence.
- The database is a nucleotide database.
- All six frames are translated on the query and on the database.
- Coding vs. Coding homology.
- For searching distantly-related species.
- Sensitive but expensive.
Appendix D: Completed Annotation Report for the Spinophilin G Isoform of Drosophila erecta

Annotation report

Student Name: xxxxxxx & xxxxxxxxx  
Student E-mail: xxxxxxx@amherst.edu & xxxxxx@amherst.edu  
Faculty Advisor: Dr. Julie Emerson  
College/University: Amherst College  

Project name: derecta_2nd3Lcontrol_Nov2011_fosmid52  
Project species: Drosophila erecta  
Date of submission: Dec. 7, 2011  
Size of project in base pairs: 28069  
Number of genes in project: 1  

Complete Report

For each gene complete the following Gene Report Form (copy and paste to create as many copies as needed, be sure to create enough isoform reports within your gene form for all isoforms):

==========Gene Report Form==========

Gene name: D erecta Spinophilin  
Gene symbol: dere_Spn  
Approximate location in project (from start codon to stop codon): 37067-8998  
Number of isoforms in D. melanogaster: 10  
Number of unique isoforms based on coding sequence: 8  
List names of unique isoforms (i.e. PA, PC etc): PB/PC/PH, PI, PJ, PG, PF, PD, PE, PK  
Number of unique isoforms found in this project: 1  

==========Isoform report==========

For each (protein) isoform complete the following (copy and paste to create as many copies as needed):

Gene-isoform name: dere_Spn-PG  
Is the 5’ end of this isoform missing off the end of the project: no  
Is the 3’ end of this isoform missing off the end of the project: no
Enter the coordinates of your final gene model for this isoform into the gene model checker and paste a screen shot of the results below:

Using the custom track feature from the Gene Model Checker, capture a screen shot of your gene model shown on the browser for your project; zoom in on only your isoform. **If available**, also show these tracks: other ref seq; all relevant Gene Prediction Tracks, 3-way and 5-way multi 2). If you need help, see lab instructor and/or read the bottom of page 9 in The Gene Model Checker User Guide, on the “Documentations” page under the Help menu at gep.wustl.edu. (Type comments about your model below the screen shot.)
Each of the gene predictor tracks is very consistent with our model. As can be seen from the screen shot there are no major differences between my model of the Spn-PG gene and the various gene predictor tracks, thus supporting our model.

Do an alignment using blast2seq of the predicted gene model protein coding sequence compared to the protein sequence from *D. melanogaster*. Copy and paste the complete results below. (Type comments about the quality of the alignment below it):

```
>lcl|50093 unnamed protein product
Length=2122

Score = 3671 bits (9520), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 2047/2102 (97%), Positives = 2072/2102 (99%), Gaps = 5/2102 (0%)

Query  1     MEKPMHAPAPVGVKSQIANIFQRKPIEIQPVEQPSAVAHAHAAAAHHAAHVVQGA  60
Sbjct  1     MEKPMHAPAPVGVKSQIANIFQRKPIEIQPVEQPSAVAHAHAAAAHHAAH-VQGA  59

Query  61    PAVRTESHSARFNNARALFEKLGVESNSNVSSRLRSGSREDNLCDGSDRRSSSSDSSQ  120
Sbjct  60    PAVRTESHSARFNNARALFEKLGVESNSNVSSRLRSGSREDNLCDGSDRRSSSSDSSQ  119

Query  121   SPPKRRTPPFPPSGLVHNNNAAIQAQNGVPPQRSLSKFIIVEPAQQVPVPTTVKYPQH  180
Sbjct  120   SPPKRRTPPFPPSGLVHNNNAAIQAQNGVPPQRSLSKFIIVEPAQQVPVPTTVKYPQH  179

Query  181   NISRLKSEEPSVPPASGSVSLPASSGGDKPKEFKPERKFSRELIEKQKKWTSHTFTKTK  240
Sbjct  180   NISRLKSEEPSVPPASGSVSLPASSGGDKPKEFKPERKFSRELIEKQKKWTSHTFTKTK  239
```
<table>
<thead>
<tr>
<th>Query</th>
<th>Sbjct</th>
</tr>
</thead>
<tbody>
<tr>
<td>241</td>
<td>240</td>
</tr>
<tr>
<td>299</td>
<td>298</td>
</tr>
<tr>
<td>359</td>
<td>358</td>
</tr>
<tr>
<td>419</td>
<td>418</td>
</tr>
<tr>
<td>479</td>
<td>478</td>
</tr>
<tr>
<td>539</td>
<td>538</td>
</tr>
<tr>
<td>599</td>
<td>598</td>
</tr>
<tr>
<td>659</td>
<td>658</td>
</tr>
<tr>
<td>719</td>
<td>718</td>
</tr>
<tr>
<td>779</td>
<td>778</td>
</tr>
<tr>
<td>839</td>
<td>838</td>
</tr>
<tr>
<td>899</td>
<td>898</td>
</tr>
<tr>
<td>959</td>
<td>958</td>
</tr>
<tr>
<td>1018</td>
<td>1017</td>
</tr>
<tr>
<td>1077</td>
<td>1076</td>
</tr>
</tbody>
</table>

- Query and Sbjct columns represent the sequences compared in the alignment.
- The numbers refer to the positions in the sequences.
As can be seen from the excellent correlation between our predicted protein in *D. erecta* against the equivalent protein in *D. melanogaster* in the Blastp, our prediction is most likely very accurate.
Appendix E
A Sample Annotation Problem – Instructor Answer Key

Question 1
Which gene predictor matches the best to the BLASTX output?

The Nscan Gene Predictor matches up best to the BLASTX output, showing three features that line up well to the three different features (putative protein-coding genes) in the BLASTX track.

Question 2
What D. melanogaster protein-coding gene in the BLASTX track appears to match well to the left-most D. grimshawi feature, as predicted by Genscan or Nscan? Write the symbol for this gene below.

PMCA

Instructor’s Note: If one clicks on any of the alignment blocks in the BLASTX track for each of the features, a BLAST Summary Viewer window appears with details of the alignments, including the E values. The PMCA feature appears to match much better (larger, more numerous blocks and lower E values) than the other ‘hits’ and suggests that this region of the fosmid contains the ortholog of the Drosophila melanogaster PMCA gene.

Type this gene symbol (case sensitive) into the http://flybase.org/ Quick Search box. What is the full name of this gene?

Plasma membrane calcium ATPase I

Which chromosome is it on in D. melanogaster? 4
Why does the fact that the \textit{D. melanogaster} gene is on this particular chromosome (and not on a different one) strengthen the case for this gene being an ortholog to the \textit{D. grimshawi} gene?

Twenty-seven of 28 genes are shared between \textit{D. melanogaster} and \textit{D. virilis} in a region of the fourth (dot) chromosome that is shown in the introductory Power Point slides (figure from 2006 Genome Biology paper). This shows that there has been little movement of genes off of the 4\textsuperscript{th} chromosome since the two species diverged. The fosmid we are analyzing for this exercise contains dot chromosome DNA from \textit{D. grimshawi}. Since the evolutionary distance between \textit{D. grimshawi} and \textit{D. melanogaster} is similar to that between \textit{D. virilis} and \textit{D. melanogaster} (see the 12 \textit{Drosophila} genomes phylogenetic tree), we would expect that there would be similar levels of gene conservation on the 4\textsuperscript{th} chromosome of \textit{D. grimshawi} and \textit{D. melanogaster}. Thus, the fact that the PMCA gene is on the 4\textsuperscript{th} chromosome in \textit{D. melanogaster} is additional evidence that this region of the \textit{D. grimshawi} dot chromosome contains the PMCA ortholog.

**Question 3**

Examine the list of the top 25 hits. How do the Scores and E-values of the PMCA isoforms compare to the other hits?

The results (see screen shot below) show hits to various isoforms of PMCA with Scores of over 2000 and E-values of 0 (which indicates that there is zero percent probability that we could have gotten those high S scores by alignment of any two random sequences).

![Screen shot of BLAST results](image)

The next best hits are to isoforms of the SPoCk and Atpalpha genes, with Scores about 10-fold lower and E-values ranging from $10^{-67}$ to $10^{-52}$. Looking at the alignments, one sees the reason for these
lower scores and higher E-values, as the aligned sequences are much shorter and much less similar. Turns out that the SPoCk and Atpalpha genes are also NOT on the dot chromosome in *D. melanogaster*, which further decreases our confidence that they are present in this fosmid (which contains DNA from the 4th chromosome of *D. grimshawi*). Thus, the evidence indicates that it is the PMCA ortholog that is found in this region of *D. grimshawi* DNA.

**Question 4**  
**Hint**: Examine both sides of the Polypeptide Details window to answer these questions.  

*How many different protein isoforms exist for this gene?*

There are 11 different mRNA isoforms, but only seven different protein isoforms. From Figure 4 on p. 27 of the Sample Annotation Problem and the table below, one can see that all but one (the P isoform) of the protein isoforms have CDS #22... through #8... in sequential order, with CDS #22... being the first (5'-end) coding exon of the gene). However, as seen below, the other end of the protein varies across several isoforms.

![CDS usage map](image)

*Instructor’s Note*: The isoform numbers after the first underscore in the above screenshot and in the text and Figures 4 and 5 of the Sample Annotation Problem change concomitant with updates to FlyBase and the Gene Record Finder. These differences can be ignored. (The above screenshot was taken on May 29, 2012.)

*Which mRNA isoforms code for identical protein isoforms?*

What might the different protein isoforms tell you about the (minimum) number of stop codons that are used in the expression of this gene?

Isoforms O, Q, K, I, R, P and L include the 3'-most exon (#1_...), so there must be a stop codon somewhere in this exon. However, there must also be a stop codon in exon #5_..., since this is the final exon of isoforms J, S and M; isoform N ends with exon #6_..., which therefore must also have a stop codon. So, the minimum number of stop codons is three.

Given the above answer, why do you think there is not a protein isoform that includes all of the CDS?

All isoforms that include the final exon (#1_...) also omit CDS #6_... and #5_.... Since we know from the above answer that exons #6_... and #5_... have stop codons in them, the only way that a protein can be made that includes more 3' coding sequences (e.g., #3_... and/or #1_...) is to omit (by alternative splicing) exons #6_... and #5_....

Question 5.
Repeat the same blastx search with the next two CDS’s (#21_... and 20_...); copy and paste the best alignments into a Word document (when copying alignments, be sure to include the Score, etc. header information and shrink the margins and/or font to keep the sequences in alignment). What are the DNA base coordinates of the beginning and end of each alignment? What frame was translated to generate the amino acid sequence for each alignment?

Answers highlighted in red below

Second exon (CDS #21_...):
Length=64

Score = 129 bits (324), Expect = 4e-34
Identities = 64/64 (100%), Positives = 64/64 (100%), Gaps = 0/64 (0%)
Frame = +2

Query 3257  LSGKADDEHRETFSVNPVPKPFKTLTLVWEALQDVTLLILEVAALVSLGLSFKPA  3436
Sbjct  1     LSGKADDEHRETFSVNPVPKPFKTLTLVWEALQDVTLLILEVAALVSLGLSFKPA  60
Query  3437  DEDA
Sbjct  61    DEDA  64

Third exon (CDS #20_...):
Length=95

Score = 189 bits (479), Expect = 9e-52
Identities = 92/95 (96%), Positives = 95/95 (100%), Gaps = 0/95 (0%)
Frame = +1

Alignment on next page:
Question 6

Look around the region where the alignment to CDS 21...*(the second exon) begins. How many acceptor sites can you find? *See Instructor’s Note on page 3.

The UCSC Genome Browser window for this area of the fosmid is shown below. The second exon begins at around base 3257. There are two acceptor sites in this region: the “AG” at 3253-4 and the “AG” at 3260-1.

Considering the frame of the conserved amino acids you found in question 5, what is the phase of each putative acceptor site you find?

From the exon by exon blastx search results on page 4, we know that the reading frame with the conserved amino acid sequence for exon 21 is Frame +2 of the fosmid. The Frame +2 amino acid sequence is the second row of light and dark gray boxes in the above screenshot. Note that in Frame +2, the 3253-4 acceptor results in the exon beginning two bases (GC) before the first complete codon (for the L amino acid); this we denote “phase 2.” The acceptor at 3260-1 has one base (C) before the first complete codon (the codon for the G) in frame +2 (phase 1).
Using just phase information, which if any of these acceptor sites is/are usable to maintain the proper translation frame throughout the first two exons?

Since the intron donor sequence after the end of the first exon creates a phase 1 exon (see page 31 of the Sample Annotation Problem), we must find an acceptor that results in a phase 2 start to the second exon. Thus, the best acceptor is the 3253-4 "AG". Note that the other acceptor is not only out of phase, but, if used, would cause one of the conserved amino acids (the L) to be omitted from the protein.

Itemize what other evidence you could consider if you have two or more possible donor/acceptor pairs.

When two or more intron donor/acceptor pairs are found, the following should be considered:
1. The pair that maximizes the inclusion of conserved amino acids would be strongly favored.
2. Pairs that have more bioinformatic support would be favored. This includes co-incidence with gene predictors (the more the better) and higher scoring predicted splice sites (see red arrow in previous screen shot).
3. For any combinations that are indistinguishable by the above criterion, by convention, the pair that creates the longest protein should be picked.

Finally, record the base coordinates for the first exon and the beginning of the second exon as deduced from your complete analysis.

The final result has coordinates of the first exon as 3035 – 3191 (with phase 1). The start of the second exon is phase 2 and mapped to 3255 (the first base in the exon, e.g. the first base after the ‘AG’ that ends the intron).

**Question 7**

*Use the results of the alignment of the second and third exons in question 5 to locate the 3’ end of the second exon and the beginning (5’ end) of the third exon.*

Following the general procedure above, the end of the second exon (CDS #21_...) is found at 3449 and is phase 1, and the beginning of the third exon (CDS #20_...) is found at 3971 with phase 2.