Jim Bogenpohl
bio 4342
5/7/04

Final Project
Annotation of Littoralis Fosmid XAAA121

Overview: For this paper, a 47 kb fosmid of Drosophila littoralis DNA was sequenced and annotated. Genscan was used to predict genes, and these genes were compared to homologous genes in Drosophila melanogaster. Synteny was noted. For fosmid XAA121, three genes were found that had syntenous homologs on melanogaster chromosome 3R: CG17267, Cdc2c, and Oamb. The littoralis fosmid was also examined for repitious DNA. Excluding low complexity repeats, only one LTR repeat was found by RepeatMasker in the fosmid. Including the low complexity repeats, the fosmid was only 3.38% repitious DNA. A map of the fosmid is shown below (Fig 1).

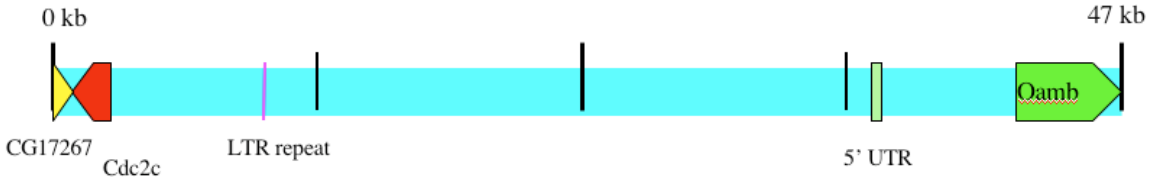

Figure 1. A map of fosmid XAAA121. The vertical black lines are merely hash marks to help measure distance.

Genes: The first step in the analysis of genes was to examine the output of the Genscan gene finder (Fig. 2). Genscan predicted six genes. For the rest of the paper, these features (brown lines) will be referred to by a number 1-6, numbered from left to right as seen in Figure 2 below. The predicted features were further examined by blast analysis.

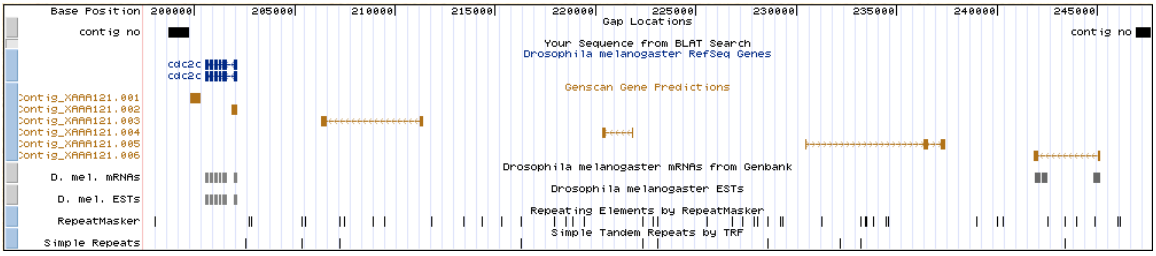

Figure 2. Genscan gene finder output for fosmid XAAA121. Predicted genes (brown lines) will be referred to as features 1-6, numbered left to right as seen in the figure above.

Feature 2 was perhaps the easiest feature to annotate. Although it did not predict even close to the right size or number of exons, Genscan aligned this small feature to the melanogaster gene Cdc2c (ID CG10498) on chromosome 3R. This gene, also known as Cell Division Control protein 2C is a kinase involved in control of the cell cycle. Blasting the fosmid against the melpro database found that all of Cdc2c hit to this region of the fosmid with an e-value of 2e-85. The protein sequence of Cdc2c was obtained

from Ensembl, and blast2 was used to perform a tblastn of each exon against my fosmid. This allowed me to pin down the exact location of the boundaries of the coding sequence of each exon except for one. This blast-based method for finding exon boundaries only works when the boundary falls between two codons. If the boundary fell in the middle of the codon, the genetic code had to be manually examined and the boundary was found by hand. The boundaries were confirmed by excising the regions of my contig, appending them into a single file, and translating them. The locations of the exons are reported below in Table 1.

Table 1. Locations of coding sequence of exons of Cdc2c.

Exon	Start	End
1	2394	2266
2	1875	1681
3	1610	1500
4	1447	1286
5	1229	1032
6	969	823

The next feature that was annotated was feature 6. Genscan aligned some melanogaster ESTs to this region. Blasting this region against the melpro database yielded a hit to a seven-exon melanogaster gene Oamb (ID CG3856) with an e-value of 8e-78. This gene, also known as Octopamine Receptor in Mushroom Bodies, is an adrenergic receptor molecule in the fly brain, and it is also found on melanogaster chromosome 3R. Again, the blast2 method was used to determine the exon boundaries in Table 2. My fosmid was found to contain the coding sequence of exons 4-6. Exon 7 was off the right end of my fosmid.

Exons 1-3 of Oamb are untranslated, and if conserved, they should lie somewhere in the area of feature 5. I performed a blast2 of the predicted mRNA of feature 5 against the 5' untranslated region of Oamb (Fig. 3). The two sequences matched in two small areas close to the translational start site with an e-value of 0.01. Then, a blast2 was performed of the 5' UTR of Oamb against my fosmid, in order to get the exact position of the hits (Fig. 3). Both small hits were located around base 37,000, which would correspond to exon 3, by synteny. This exon is poorly conserved, and I could not determine its boundaries.

Table 2. Locations of coding sequence of exons of Oamb.

Exon	Start	End
4	42189	42440
5	42504	42740
6	44892	45419

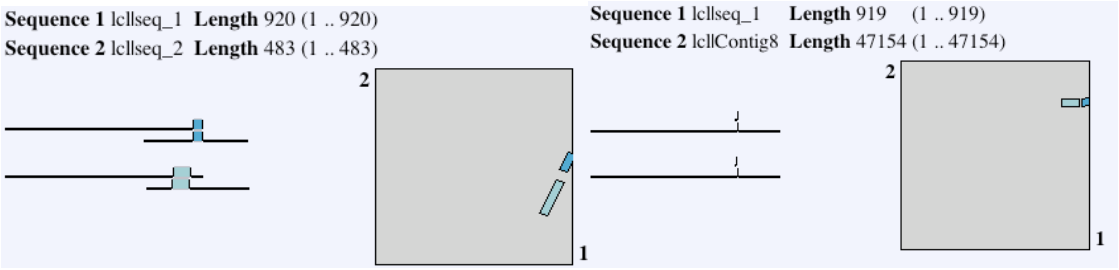

Figure 3. Left, blast2 output of the 5' UTR of Oamb (1) blasted against the predicted mRNA of feature 5 (2). Right, blast2 output of the 5' UTR of Oamb (1) blasted against my fosmid (2).

The next feature examined was feature 1. None of the databases we were given in class (melpro, nr, swissprot, etc.) found a hit for this feature. However, I had a hunch that something was there, by synteny with melanogaster. So, I extracted the first portion of my contig and blasted it against ncbi's Drosophila database, and I got a hit to CG17267, the syntenous gene, with an e-value of 2e-17. Further blast analysis showed that my fosmid contained the third exon of this small gene, and the first two exons were off the left end of my fosmid. Table 3 states the position of this exon.

Table 3. Location of coding sequence of exon of CG17267.

Exon	Start	End
3	169	585

Feature 3 was a strange feature. None of the blast searches I did on goose found anything, but when blasted against the ncbi Drosophila database, it returned a hit to CG6300, another gene on melanogaster chromosome 3R from a region ~1 megabase away from the syntenous genes described above. The hit had an e-value of 8e-11, and it matched to one small portion (~50 aa) of the gene. This feature is probably either a pseudogene or more likely just sequence that randomly is similar to CG6300 by chance.

Feature 4 showed no matches to any blast search, and is most likely a false prediction made by Genscan. Table 4 summarizes the features predicted by Genscan.

Table 4. Summary of features predicted by Genscan.

Feature	Homology
1	exon 3 of CG17267
2	Cdc2c
3	nothing
4	nothing
5	5' UTR of Oamb
6	exons 4-6 of Oamb

Synteny: All of the genes identified on fosmid XAAA121 had syntenous homologs on Drosophila melanogaster chromosome 3R. There is one hypothetical gene, CG31205, annotated on melanogaster 3R that was not found in the littoralis DNA. This may be

because the gene does not really exist, or it might be explained by millions of years of divergence between the two species. The probe used to pull out this fosmid was Best. There was no good hit to Best on this fosmid. The strongest hit to Best on the fosmid had and expect value of 171267! Three genes were found in the 47 kb fosmid, so the frequency of genes for is 0.064 genes/kb. This region is not highly repetitious in littoralis nor melanogaster. Figure 4 shows the synteny between XAAA121 and Drosophila melanogaster chromosome 3R.

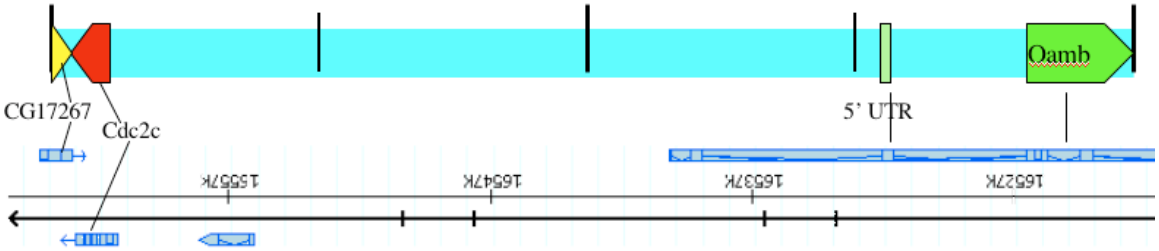

Figure 4. A map showing the synteny of XAAA121 to Drosophila melanogaster chromosome 3R. The third melanogaster gene from the left did not have a homolog in littoralis.

Clustal Analysis: Clustal, a multiple sequence alignment tool, was used to compare the Cdc2c gene of several species: D. melanogaster, D. littoralis, D. yakuba, Anopheles gambiae, Homo sapiens, Fugu rubripes, and Antirrhinum majus (snapdragon). The clustal output (Fig. 5) shows that Cdc2c is very well conserved across its entire length through several phyla. This is entirely surprising, given that the gene's function is involved in control of the cell cycle. Insects, vertebrates, and plants all undergo the cell cycle.

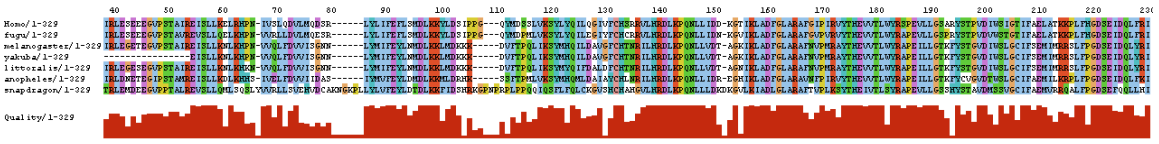

Figure 5. Clustal output for Ccd2c of D. melanogaster, D. littoralis, D. yakuba, Anopheles gambiae, Homo sapiens, Fugu rubripes, and Antirrhinum majus.

When done through the EBI webpage, Clustal also generates a cladogram showing estimated evolutionary distance between all of the molecules that you put in. Figure 6 shows the cladogram for Cdc2c of the species above.

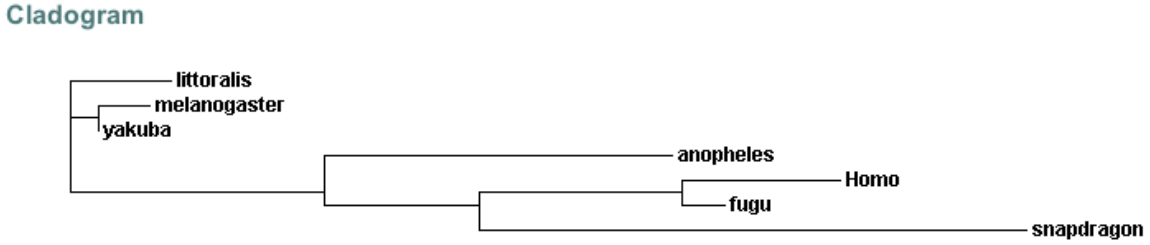

Figure 6. Cladogram of Cdc2c from D. melanogaster, D. littoralis, D. yakuba, Anopheles gambiae, Homo sapiens, Fugu rubripes, and Antirrhinum majus. Distance from left to right corresponds to estimated evolutionary distance.

Clustal can also be used to look for possible conserved regulatory elements. About 1500 base pairs upstream of each of the Cdc2c genes for D. melanogaster, D. littoralis, D. yakuba, D. pseudoobscura, and H. sapiens were extracted and run through Clustal (Fig. 7). No conserved promoter elements were seen.

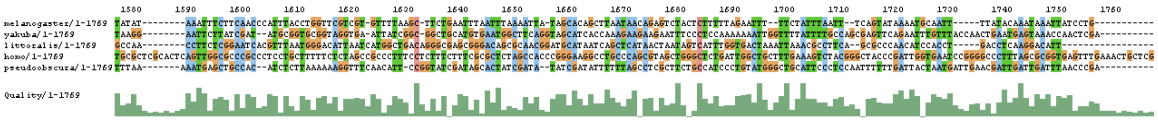

Figure 7. Clustal output looking for conserved promoter elements for Cdc2c in D. melanogaster, D. littoralis, D. yakuba, D. pseudoobscura, and H. sapiens. About 1500 bases upstream of each gene were examined.

Repeats: This portion of the genome is not highly repetitious in littoralis nor melanogaster. No repeats were found by blasting against the repeat libraries given to us. RepeatMasker found one LTR repeat with the low complexity sequence masked and 28 simple repeats and 16 low complexity repeats without low complexity masking turned on. The overall percentage of repetitious DNA in my fosmid was 3.38%. See attached spreadsheet for a full list of repeats and locations.

Appendix. All files requested for the appendix will be sent in digital form.