Lab Weeks 10 & 11: Advanced Annotation Instruction Sheet

(adopted from the Genomics Education Partnership - http://gep.wustl.edu/)

Special situations in gene annotation

There are a few situations that deserve special comments. You may not run across these situations when creating your own gene model, but the comments below may be helpful if you do. Do note, however, that some of the algorithms, techniques and programs discussed below have not been introduced or discussed in the introductory GEP exercises and there is no GEP training material available for them. In your role as annotators, you are encouraged to use these techniques if you have time and motivation but are certainly not required to use them.

Conservation on one end of an intron

Sometimes when searching for similarity between *D. melanogaster* exons and your genomic region, you will find one exon with a very good match to the intron/exon junction at one end of the intron and no match at the end of the adjacent exon at the other end of the intron. It is still sometimes possible to find the unaligned site by using a string search instead of a BLAST similarity search. This is probably best explained by example. Consider these two alignments for the first and second exon of some hypothetical gene. In *D. melanogaster* the first exon is 68 aa long and the second exon is 28 aa long:

Exon 1 sequence:

```
 1    MDINNEIENIISDIDINIKAQEKLKEQELKAQQYQQNQNK
41    YNPASGPITETQTTTTVVVTKKDSEEET 68
```

And alignment to our hypothetical genomic region (*D. melanogaster* is Query sequence below)::

```
Query  5736  MDINNEIENIISDIDINIKAQEKLKEQELKAQQYQQNQ  5773
             MD NN+I NIISDIDINIKAQEKLK+ E ++ +   +Q
Sbjct  1     MDFNNQILNIISDIDINIKAQEKLKQNECQSGELDLHQ  38
```

Exon 2 sequence:

```
 1      ESANVSKTVDLRKIFTPATDAAEILPKN  28
```

And alignment (*D. melanogaster* is Subject sequence below):

```
Query  6259  ESSNLSRTVDLRKIFTPATDAPEILPKN  6342
             ES+N+S+TVDLRKIFTPATDA EILPKN
Sbjct  1     ESANVSKTVDLRKIFTPATDAAEILPKN  28
```

Notice the strong alignment at the start of exon 1. This gives good evidence where exon 1 starts but given its length in *D. melanogaster* it may be difficult to find the donor site at the end of exon 1. We would certainly follow the exon length conservation rule above and look downstream about 90 bases (i.e. we are missing 30 aa or 90 bases of the end of the first alignment), but this just gives the general area where we might expect the end of the exon. Interestingly, exon 2 starts with a very strong alignment. Thus, when considering the amino acid sequence of the *D. melanogaster* protein with the amino acid sequence of the potential new gene, we have identified two conserved domains separated

by a region without conservation. It would be somewhat unlikely that the 5' end of the downstream conserved domain coincides exactly with the 5' end of exon 2. If these two 5' ends do not coincide we would expect exon 1 should end with at least a few conserved amino acids. Since BLAST did not detect these amino acids (i.e. only the 5' end of exon 1 shows an alignment) it is likely that the number of conserved amino acids at the 3' end of exon 1 is very small (one or two). In these cases a search for a short DNA sequence that would code for one or two conserved amino acids next to an in frame splice junction may be fruitful. In the example above then exon 1 might end with the same 1 or 2 aa as the exon in *D. melanogaster*.

To start the search, we must first find the phase of the acceptor site at the beginning of exon 2. Since there is a very strong alignment that ends at the first amino acid we expect an acceptor site to be 0, 1 or 2 bases upstream of base 6259. In this example, we will assume that the acceptor site at the start of exon 2 immediately precedes the codon for the glutamic acid (E). As such we would look for a DNA sequence to end exon 1 that codes for the amino acid E (glutamic acid) and then T (threonine) and then a donor site. If the acceptor site at the beginning of exon 2 had one base and then the codon for the glutamic acid we would look for a sequence to end exon 1 which was a codon for E, a codon for T, any single base and then the donor site. Since the codon table is degenerate we will need to look for a number of different sequence that could code for ET(donor). Checking the codon table we see that E has two codons, GAG and GAA, and T has four codons, ACT, ACC, ACA, ACG. If these two amino acids were conserved and we want a phase of 0 to match the acceptor of exon 2, then any of 8 different sequences could code for ET(donor):

GAGACTGT	GAAACTGT
GAGACCGT	GAAACCGT
GAGACAGT	GAAACAGT
GAGACGGT	GAAACGGT

If the potential region where these amino acids can be is small, it is possible to simply search by eye. It is also possible to use BLAST to search your sequence if you change some of the parameters to specifically allow for these very short alignments. First, set the word size to a number less or equal to the length of the sequencing you are searching with. Also, be sure to turn OFF the filter and set a very large expect threshold (different implementations of BLAST calculate E-scores differently when comparing two sequences, it is best to experiment with the version of BLAST you are using to empirically determine the best threshold). Since there are 8 different ways to code for "ET(donor)" you would need to do 8 BLAST searches. (To ensure that the BLAST tool you are using can detect these very small alignments you may wish to do a positive control search with a sequence you know does exist within your subject to verify that it can be found).

If one of these sequences is found in the correct location (i.e. downstream of the exon 1 alignment but before exon 2 and on the correct strand) and it is in the correct frame (i.e. the same frame as the early exon 1 alignment) and is in the proper phase (i.e. links with the correct frame in exon 2) you have found pretty strong evidence for the end of exon 1 and this site should be picked.

Searching for conservation of sequences when the default BLAST search fails

While BLASTx (or tBLASTn) is your primary tool for searching for conservation, just like all other programs, they have their limits and can fail. This will happen with increasing frequency when you search with smaller exons or when you are attempting to annotate more rapidly evolving proteins. If you do a search and get no significant similarity found, your first step should be to increase the expect threshold. Remember you are looking for "the most similar" sequences not necessarily "a statistically significant" match. The technique then is to keep stepping up the expect threshold until you get alignments, no matter how large the absolute value of the E score. Those alignments in the proper location (position and strand) should then be checked to see if they can be used to generate an exon.

If BLAST fails (and it will in at least some cases), the next best search technique is to use CLUSTALW to do a DNA-to-DNA search. This is a very different kind of computational algorithm that can sometimes succeed when BLAST fails. The best technique is to extract the DNA sequence of only that region of the contig that you wish to search (you must use the proper strand in a CLUSTALW alignment) and compare it to the DNA sequence of the exon you are trying to place. The result will always be a single best alignment as determined by the CLUSTALW algorithm. The position of this alignment should then be checked to see if there is sufficient evidence for an exon (i.e. viable donor/acceptor sites).

If CLUSTALW fails then DNA-to-DNA BLASTn should be attempted. To avoid large numbers of irrelevant and misplaced alignments, be sure to either extract the region to be searched or use the "Subject subrange" feature at NCBI BLAST. Here again the expect threshold should be increased in increments until alignments are found. The alignments generated are again good starting points for further investigation.

Evidence for conservation can also be found in the multi-species multiz tracks (if available). These tracks look for high levels of sequence conservation by comparing more than two species.

Very small exons

It can be quite difficult to find very small exons. Be sure to increase the expect threshold until you start to see hits no matter how poor the alignment looks. To avoid sorting through lots of false alignments, you can restrict the region searched. For example, if you have the upstream and downstream exons already mapped, do not search the entire region; rather use the "from:" and "to:" boxes in BLAST 2 sequences to restrict your search to the region between the two mapped exons. You can also try to make the search more sensitive by changing the word size from 3 to 2 but be aware that this may cause you to miss high quality blast hits if the sequence you are searching is large so be sure to restrict the area searched using the above technique if you reduce the word size. Remember very weak similarity, if in the right place and with usable donor and acceptor sites, is probably identifying the correct exon.

The initial exon is often very small. These exons can be searched for using the short degenerate search trick described above, but will often fail. In this case, you should look for any candidate methionine codon upstream of your second exon which is also very close to an in frame medium or high quality donor site. If more than one is found in the proper region, pick the one closest to the next exon.

You may also try using the GEP's on-line 'Small Exons Finder' (SEF) at http://gep.wustl.edu -> Projects -> Annotation Resources -> SEF. See the GEP's Small Exon Finder User Guide at http://gep.wustl.edu/help/documentations/ for help with this search feature.