Annotate Your Chimp Sequences *Spring 2006*

In the last few labs, you've seen the basics of how to annotate genes and pseudogenes. You now have a chance to apply this knowledge to annotate some DNA sequences from chimpanzee.

What to Produce

The endpoint of the annotation process will be a map of all the features found within the sequences, along with a brief record of how the feature was annotated. Here's what we would like to see:

1) At the beginning of the report create a list of all gene-like features, including their endpoints in the sequence and a 1-2 line summary of what you think the feature is. The list should include genes and pseudogenes (though you may not be able to tell these apart with certainty) as well as anonymous EST matches that were detected but don't seem to correspond to any other annotated feature. If your annotation refers to an existing GenBank or Swissprot record, be sure to include the accession number for this record.

Note that, even if you have multiple BLAST matches to a particular feature, you should give it just a single line. Your annotation should integrate the evidence from the various matches.

2) Follow the list, with a concise but suitably detailed summary (paragraph-length) indicating what you know or believe about the feature. Give any opinion you have as to whether it is a gene, a pseudogene, or something else, as well as what kind of evidence supports your hypothesis. If you can't make up your mind what the feature is, say so (but give evidence in favor of your two or three favorite hypotheses). We're looking for something on the order of what you wrote at the end of Lab 1, plus any comments related to EST evidence.

You should also annotate the longer transposable element relics, such as LINEs and retrovirus-like elements, found by RepeatMasker. These features provide good evidence that a particular part of the sequence has a known (if now inactivated)
function. Don't bother drawing in all the Alus.

As discussed, work on these individually, but if you get stuck feel free to discuss with others regarding the proper interpretation of the data. Remember you will be graded mostly on the amount and type of evidence you find for any feature and how you interpret it.

Recommendations on How to Proceed (command-line version)

I would start with Genscan output as well as any significant BLAST hit to **Swissprot** and the human EST library. You could also blast against protein **nr**. As you find pieces worth investigating more closely, you can extract these pieces with the **extractregion** script if necessary and blast them against the nucleotide database. Follow up with an examination of the region in the human and/or chimp browsers.

For each feature, you should think about at least the following questions:

A) What family of genes or other sequences does the feature match? For regions exhibiting only anonymous EST matches, you might try to check them with **blastx** against the protein **nr** database, if you haven't already done so for your entire contig.

B) Within a gene family, is there evidence as to which member gave rise to the feature? Remember to use evidence from mRNAs as well as protein to back up your answers.

C) Does the feature appear to match its human ortholog? Does it match at the right genomic location in human? Does it match human ESTs or mRNAs as well as would be expected for an orthologous chimp sequence?

D) Is the match full-length at the protein level? At the DNA level? Are certain parts of the feature better conserved than others? Do you see evidence that is inconsistent with the feature producing a working version of its annotated protein (if any)?

E) What is your best guess as to the feature's boundaries? If you're looking at an mRNA, is the EST evidence consistent with the annotated mRNA boundaries?

F) When a single feature matches two or more discontiguous segments of the contig, are there repetitive elements in between them? (This can be good evidence of either an intron or a pseudogene, depending on other evidence.)

G) How certain are you about your calls based on the available evidence?

Remember that you have at your disposal BLAST itself, the Herne output viewer, the HTML BLAST output, all its links to GenBank, Genscan, the genome browsers and the Swissprot / Expasy database and other databases. And, of course, there's always Google! If you're stumped as to what a particular feature might be, be creative in trying to gather evidence.

Recommendations on How to Proceed (web-based version)

I would start with Genscan output as well as any significant BLAST hit to **Swissprot** and the human EST library. You could also blast against protein **nr**. As you find pieces worth investigating more closely, you can use the "set subsequence" option on the NCBI BLAST web server to search these specific regions against the nucleotide database. Follow up with an examination of the region in the human and/or chimp browsers.

For each feature, you should think about the questions posed above (A-G).

Remember that you have at your disposal the Genscan predictions, the HTML BLAST output and all its links to GenBank, the evidence tracks on the UCSC genome browser and the Swissprot / Expasy database. And, of course, there's always Google! If you're stumped as to what a particular feature might be, be creative in trying to gather evidence.