

Artemis Exercise: *Using Artemis to annotate a section of a genome.* Due Nov. 12

From <http://www.sanger.ac.uk/Software/Artemis/>: "Artemis is a free genome viewer and annotation tool that allows visualization of sequence features and the results of analyses within the context of the sequence, and its six-frame translation." You can read more at the Sanger Centre site.

First, start Artemis. If you are not at one of the course computers, you can do this remotely using the following steps:

```
cd
cp -r /genome/ICEbin/artemis .
module load java
artemis/art
```

1. In Artemis, OPEN the fasta file called MysteryIGERTSeq.txt (available as a txt file on the IGERT course webpage). How long is the sequence? Why are there 6 tracks corresponding to amino acid sequences, and what are the black lines in these tracks in the upper display?
2. You want to find ORFs longer than 100 codons. First Select All Bases, then choose Create/ Mark Open Reading Frames, selecting 100 as the minimum length when queried. How many ORFs of the minimum length are displayed?
3. Based on observations so far, this fragment is likely to have come from what group of organisms? Why is this likely?
4. Which ORFs seem likely to be real genes and why?
5. Select the longest ORF by clicking on it, then use View to see the fasta for the polypeptide. How long is the polypeptide? Does the start position seem to be correct?
6. Use this polypeptide as the query in a blastp search against the nr protein database at NCBI. What is the organism and protein of the closest hit? Is the match significant? What is the 4 letter gene name?
7. Which amino acid position and type appears to be the initial amino acid in the polypeptide, based on information from the blast alignments? Explain.
8. This ORF overlaps with some shorter ORFs on the opposite strand. Select the longest of these and use as query in another blastp search. What do results tell you about the likelihood that this protein represents a real gene? Explain.
9. Under Graph in Artemis, you can observe the %GC across the fragment. What is the average %GC? How does it vary with respect to ORF position?
10. Go to www.ecocyc.org, which is a comprehensive database for the functional genomics and metabolism of *E. coli*, the best studied cellular system. Under

Search, you can blastp your ORF (the long one) against the *E. coli* set of proteins. Why might this be helpful for annotating your fragment (in addition to blastp-ing against the larger NCBI nr protein database)? Does this search support the earlier indicators of the translation start site?

11. From your blastp results page in ecocyc, lick on the 4 letter gene name to go to the “gene” page. What is the function of this gene (briefly)? What are examples of GO terms for the Biological Process and Molecular Function categories? Approximately how many papers have been written about this gene, and what was the year of the first one?
12. Under Sequence Features on this gene page, two functional motifs are indicated. Are these conserved at the amino acid level between *E. coli* and your mystery organism?