
CS/ST 480

HOMEWORK 1 (DUE JANUARY 30TH)

1. **Single nucleotide substitutions [15 pts].**

A single nucleotide substitution at which position in a codon would most likely have the greatest impact on the function of the encoded protein: the first, the second, or the third? Why?

2. **Point mutations [15 pts].**

Which of the following of point mutations would most likely have the greatest impact on the function of the encoded protein: a single nucleotide substitution (e.g. A mutates to G), or a single nucleotide deletion (e.g. an A is deleted from the sequence)? Why?

3. **Codon usage [20 pts].**

Although the genetic code is universal (with some exceptions), each organism has its own preference for codon usage. The web page http://www.molbiol.ox.ac.uk/~cocallag/refdata_html/codonusagetable.shtml gives statistics on the codon usage of *Escherichia coli*. Your colleague has the following sequence fragment from *E. coli*: AAGUCAUUAUUUUCG.

Assuming this is the coding strand, can you help her to identify the most likely translation frame?

4. **Gene prediction [50 pts].**

In class we defined an Open Reading Frame (ORF) as a stretch of sequence that starts with a start codon, ends with a stop codon, with no stop codon in the middle, and whose length is a multiple of 3. Write a program to look for long open reading frames in DNA sequences. Analyze the DNA sequences posted on the homework page, which includes the sequences of the sars virus, two prokaryotes, chromosome 1 of the yeast genome, and chromosome 22 of the human genome. Describe your experiences with this simple gene recognition technique. What is the size distribution among the long (say > 100 codon) ORFs you find, and how does that compare to random sequences.

Include a printout of your program.