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. 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? In your answer use the codon usage
   database at: http://www.kazusa.or.jp/codon/ look for the codon usage table for
   Strain K12 of *Escherichia coli*. Explain your approach.

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. Note that ORFs can overlap! Write a program to look
   for long open reading frames in DNA sequences. Analyze the DNA sequences of the
   following organisms: the sars virus (genbank ID NC_004718), *Haemophilus influenzae*
   (NC_000907), and *E. coli* (NC_000913). Once you find the genbank entry for the
   sequences, download them in a format called *fasta*, which is simpler than the genbank
   format. 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. Compare your findings to the number of genes
   that organism has.

Submit your program by email to the instructor.

**Submission:** In class on the due date, except for program printout.