CSC 314, Exam II Review.

Note: This review is not comprehensive and focuses on the earlier material (GenBank, probability) and sequence alignments. To prepare for the exam you should understand the material in this review as well as all labs completed since the first exam.

1. ***Sequence Database questions***
   (Start at GenBank: https://www.ncbi.nlm.nih.gov/genbank/)

   a. How many RefSeq molecules are there associated with the keyword insulin?

   b. How many of these molecules are from humans (*Homo sapiens*)?

   c. How many RefSeq entries are for the insulin gene INS in humans?

   d. The top hit from the above search is the gene with accession number NM_001185098. How many exons does this gene have?

   e. What are the first nine nucleotides in the coding sequence (CDS), and the corresponding amino acids?

2. ***GEO Questions*** (http://www.ncbi.nlm.nih.gov/geo/). The following questions are about the GEO series GSE48060, which analyzes expression of genes in patients with and without recurring incidences of myocardial infarctions (heart disease).

   a. How many samples were analyzed in this study?

   b. What microarray platform were these individuals profiled on?

   c. For sample GSM1167073, what is the expression value for the probe 1053_at, and what gene does this probe correspond to? (Note: you need to look at the platform data to find the gene).

   d. Use the Analyze With GEO2R feature to identify the top 250 genes differentially expressed genes between patients with and without recurrent events

      i. What is the most differentially expressed probe and gene?

ii. The p-value for this gene corresponds to the following probability: if there really was no difference between patients with and without recurrent heart disease, the probability of observing a log2 fold change of at least _____ is equal to _____?

iii. Is this gene *upregulated* or *downregulated* in patients with recurrent heart disease?

3. Probability Questions

a. A standard deck of cards contains 26 black and 26 red cards. If a card is randomly selected, what is the probability it is black?

b. If two cards are randomly selected, what is the probability they are both black?

c. Assume that cards are sampled *with replacement* (i.e., put back into the deck each time).

   i. If two cards are drawn, what is the probability that they are both black?

   ii. If five cards are drawn, what is the probability that at least one card is red?

4. Sequence Alignments

For (a) and (b), use a linear gap penalty of 4 points, a match score of +5 points, and a mismatch score of -1 point. You must show your dynamic programming matrix to receive credit.

a. Find the score and alignment for the optimal global alignment between *handy* and *say*.

b. What is the score and optimal global alignment between *ha* and *say*? (Note: you should use your dynamic programming matrix from part (a) to answer this question)

c. What is the score and optimal *local* alignment between *snow* and *knot*?

5. Using the BLOSUM-62 matrix, a gap opening penalty of 5, and a gap extension penalty of 1, find the score of the *semiglobal* alignment given below (Note that semiglobal alignments do not penalize gaps at the beginning or end of the alignment).

```
FRIDA--Y
--P-ARTY
```