**Statistics 446/646, Spring 2019**
**Homework 4**

**Both Local & DL Students:** Submit assignment through Ecampus. Please save your submission as a PDF, with format "lastname_firstname.pdf". Include all R code.

1. Use seqinr to obtain the human Genbank DNA sequence for the PSEN1 gene.
   (a) How many sequences are returned, what are they named, and what are their lengths?
   (b) Provide the FASTA file entry for the first sequence (AB159776).
   (c) What percentage of the bases in the third sequence (BC011729.PSEN1) are either G's or C's?

2. Consider the global alignment of these two DNA sequences: GAG and GTAG. Let the score for a match be +1, the score for a mismatch be -1, and the score for a gap be -2 (so, opening a gap costs -2 points, extending that gap by one space costs an additional -2 points, and so on).
   (a) Write down the score matrix for this alignment, and fill in all of the cells. Your matrix should resemble that on slide 16 of the "alignment.pdf" slides under Notes on eCampus.
   (b) What is the optimal alignment, and what is its corresponding score?

3. Consider these two amino acid sequences: ASEDLTI and AEEDFGI. For the purposes of alignment, use the PAM30 error matrix and a gap penalty of -2 (gap open =0 and gap extend = -2).
   (a) What are the optimal alignments, using both the global and local algorithms?
   (b) Verify the overall scores for these alignments.
   (c) Use the randomization-based significance test we used in class to assess the local alignment of the two amino acid sequences:
      i. Report a density plot of the randomization based alignment scores; use the density function. Shade in the area under the density curve that corresponds to the randomization based scores that are $\geq$ our observed score of 19.
      ii. What is the resulting p-value?

4. The H1F0 gene is present in multiple species. The versions of the gene in human, mouse, and cow are said to be homologous.
   (a) What does homologous mean?
   (b) Consider the global alignment of the sequences of this gene in human (homo sapiens) and mouse (mus musculus). Use seqinr to download the DNA sequences. You will get multiple sequences for each species. Just use the first of each. The lengths of the human and mouse sequences will be the same. [ Use gap open = 0 and gap extend = -2]
      i. Comparing the human and mouse versions of H1F0, what proportion of bases differ?
      ii. Report a dot plot showing their alignment.

iii. Use the randomization-based significance test to assess the global alignment of H1F0 in human and mouse. What is the p-value?[ Use gap open = 0 and gap extend = -2]

(c) Use the muscle package to carry out multiple alignment of H1F0 in human, mouse, and cow (bos taurus). Again, just use the first sequence returned by seqinr for each species. Report the first 50 aligned bases.

5. Download the "QC spike" data from eCampus. These are mass spectrometry proteomics data. There are 12 samples. In each sample, there is mostly Salmonella proteins, but a small number of "quality control" (QC) proteins are spiked in as well. The first four samples are replicates for which the QC proteins were spiked at a high concentration, the next four samples have QC proteins at a medium concentration, and the last four samples have QC proteins at a low concentration. The numeric values in the data are quantifications of the prevalences of 4,591 peptides in each sample (larger numbers mean more copies of the corresponding peptide). The first column is a unique identifier for each peptide. The next few columns provide additional information about the peptides and the proteins they correspond to. The column named "is_qc" is an indicator for whether each peptide came from a QC protein (1 if yes, 0 if no). The remaining 12 columns contain the peptide intensities themselves.

(a) Filter out all peptides for which at least one comparison group has fewer than 2 observations (non-NA values). After completing this, you should be left only with peptides for which each comparison group has at least 2 observations. How many peptides are you left with? How many of these are QC peptides, and how many are Salmonella peptides? Use these filtered data in what follows.

(b) Report side-by-side boxplots comparing the 12 samples (so, 12 boxplots side-by-side), for the QC peptides.

(c) Report side-by-side boxplots comparing the 12 samples, now for the Salmonella peptides.

(d) For each of the Salmonella peptides, compute a two-sample t-test, assuming equal variances, comparing the first comparison group (highest concentration of QC proteins) to the third comparison group (lowest concentration of QC proteins).

i. Report a histogram of the resulting p-values.

ii. What proportion of the Salmonella peptides had p-values $< 0.05$? What proportion of the Salmonella peptides did you expect to have p-values $< 0.05$?