Statistics 5570/6570 Spring 2014 Homework 5

Expected Completion of Content: 4 Apr
Assignment Due (by 11:59 P.M.): 18 Apr

Directions: (Same as with previous homeworks, with 5 points for neatness and format, including code in appendix; there are 88 points possible, plus 10 available bonus points.) The main purpose of this assignment is to help students gain familiarity with sequence alignment and next-generation sequence data analysis tools in Bioconductor, and to discuss and experience BLAST online. See partial R code on the last page. This homework will require the following packages be loaded in R: Biostrings, cluster, RColorBrewer, DESeq. (The bonus points may require another package you've used before.)

1. The following two lines represent the sequences of fragments of two particular proteins found in humans:

   `GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL`

   `GNPKVKAHGKKVLGAFSDGTFATLSELHCDKL`

   (a) (3 points) Use the PAM30 scoring matrix to perform a simple pairwise global alignment of these two sequences, with an affine gap penalty of -4 to open and -1 to extend a gap. Report the score and the alignment (copying and pasting the alignment from R is fine; note that you will need to use a "space-preserving" font such as Courier New to make the characters align properly).

   (b) (3 points) Same as part (a), but this time use a gap opening penalty of -20 and a gap extension penalty of -5. Report the resulting alignment and score.

   (c) (5 points) Comment briefly on how (and why) these two alignments differ.

2. The `hw5.fasta` file available on the course website contains the sequence information for five protein sequences - two from mouse and one each from human, chicken, and rat. The sequences in this file are in FASTA format.

   (a) (3 points) Load this sequence information into R and obtain a pairwise distance matrix where distance here is in terms of the "normalized score" from a local pairwise alignment with the BLOSUM62 scoring matrix, a gap opening penalty of -4, and a gap extension penalty of -1 (see the `get.phylo.dist` function defined and used in Notes 6.1). Do not report the matrix, but instead give a phylogenetic tree based on these pairwise distances, using average linkage.

   (b) (2 points) Give a symmetric heat map representing these pairwise distances.

   (c) (5 points) Comment briefly on what the phylogenetic tree and heat map indicate about these five proteins.

3. Using the interface on the NCBI website, perform a BLAST search for the following protein sub-sequence:

   `GSAQGHGKVALTNAVAHVDDMWPNALSALHAHKL`

   You may use the default algorithm parameters.

   (a) (5 points) Report the top-scoring alignment (in a format similar to that in Exercise 1a above.)

   (b) (2 points) Report the E-value of the top-scoring alignment.

   (c) (5 points) What is the correct interpretation of this E-value?

   (d) (7 points) Convert this E-value to a P-value.

   (e) (5 points) What is the null hypothesis corresponding to this P-value?

4. Referring to the next-generation sequencing (NGS) technology as presented in class:

   (a) (5 points) Explain briefly the similarities and differences of this technology compared to the Affymetrix GeneChip we've used previously.

   (b) (5 points) What do the differences in part (a) affect in terms of the statistical analysis of data from NGS technology? (That is, what do we need to do differently?)

5. The CEPH Family Pedigree provides genomic information for several multigenerational Caucasian families from Utah. A 2010 PLoS Biology paper by Cheung et al. reported RNA-Seq data for 41 unrelated grandparents (17 female, 24 male) in this study. The sample R code below will read in and organize these data in a format similar to the in-class example of Notes 6.4 ("genes" [named by Ensembl ID] in rows, samples [named by study ID] in columns; the first 17 columns are female, and the remaining 24 columns are male).

   (a) (2 points) How many "genes" are reported in these data?

   (b) Use the `DEseq` tools as in class to perform a test of differential expression between female and male participants.

      i. (10 points) Report a plot of the dispersion estimate vs. the standardized mean estimate for each gene, with a superimposed line to visualize the fitted dispersion vs. mean relationship. Comment briefly on what this plot suggests. (BONUS 10 points: Revise the `plotDispEsts` function to use a larger plotting character and to show point density.)

      ii. (3 points) Report a histogram of the "raw" p-values (without any multiple testing adjustment) from this test.

      iii. (5 points) Comment briefly on what causes the behavior of this histogram near p-values of 1.

      iv. (5 points) Comment briefly on what the behavior of this histogram near p-values of 0 suggests.

      v. (3 points) When controlling the FDR at 0.05 (i.e., now using the "adjusted" p-values), how many genes are called differentially expressed?

Partial R Code:

```
## (5)
eset <- read.csv("http://www.stat.usu.edu/jrstevens/stat5570/cheung.csv")
frame <- read.csv("http://www.stat.usu.edu/jrstevens/stat5570/cheungGender.csv")
library(DESeq)
countsTable <- eset[,-1]
rownames(countsTable) <- eset$X
conds <- frame$gender
```