Problem Set 4
BENG 183
Fall 2011

Due Friday November 4th 2011, 11:59 PM

Please submit by e-mail to q1ma@ucsd.edu.

Question 1: Minisatellites demonstrate intraspecies polymorphism, and this has been used for DNA fingerprinting of organisms. When radioactive probes containing a minisatellite sequence are annealed with DNA blots containing restriction endonuclease digests of DNA, multiple bands hybridize. This pattern of bands varies from one individual to another. For this reason, minisatellites are often used in paternity testing.

Magda claims that Joseph (Male 1) is the father of her child. Joseph says that the real father of the child is Larry the cable guy (Male 2). The Southern blot of each individual was probed with 35S tagged $(CAG)_5$, which recognizes many microsatellite sequences. Examine this blot below (Figure 1) and calculate the probability that Joseph is the father of the child, and, separately, the probability that Larry is the father of the child. Which is the father?

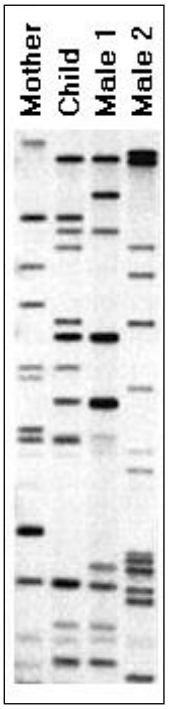

Figure 1. Paternity test.

Question 2: A very important advancement in genome assembly was the mathematical analysis done by Lander and Waterman (PMID: 3294162). The

Lander-Waterman Model can be used to plan shotgun sequencing projects.

Let's define a few variables:
G = haploid genome length in bp
L = sequence read length in bp
N = number reads sequenced

Lander and Waterman suggested that the number of times a base is sequenced follows a Poisson distribution:

$P(Y=y) = (\lambda^y * e^{-\lambda}) / y!$

where λ is the average number of times any base is sequenced (e.g. coverage), $\lambda = LN/G$. y is the number of times a particular base has actually been sequenced. For example, if y = 3, then the Poisson distribution will give you the probability that every bp has been sequenced exactly 3 times (not less, not more).

a) Determine the probability that every bp will have been sequenced if one does 3x, 5x, 10x coverage.

b) If we are sequencing the human genome, with 500 bp reads, how many reads must we do in order to sequence 90% of the genome at least twice.

Question 3: Review of terms

a). What's the definition of a 'Housekeeping gene'? How about a 'Luxury gene'?

b). There have been many different approaches to analyze the expression of RNA and proteins. Please list:

I) The classical approaches used to analyze RNA expression

II) The classical approaches to analyze protein expression

III) What classical approach could be used to analyze both RNA and protein expression level or profile?

c). The development of microarray technology greatly assisted research on mRNA expression. Describe the basic design of a microarray and the associated procedure to quantitatively analyze mRNA expression levels. (You can use a flow chart). What's the purpose of depositing DNA probe on the array in excess of the hybridized sample?

Question 4: Primary single color microarray data analysis

The microarray dataset GDS3217 (downloadable from the GEO database) describes the effect of estradiol (E2) on a breast cancer cell line expressing estrogen receptor (ER). Specifically, the analysis is of ER-positive MCF7 breast cancer cells from 0 to 48 hours following E2 treatment. ERs facilitate the transcriptional effects of hormones.

The expression results suggest potential correlations between ER binding and gene regulation. In this study, we highly recommend use of MeV (MultiExperiment Viewer) for data analysis.

a) Which are the most significant differentially regulated genes by comparison of Estradiol treated sample vs control sample? What are the top 10 up-regulated genes from the dataset (First three columns are 3 replicates for estradiol treated samples, and last three columns are 3 replicates for control samples)?

b) The data normalization step is very important for microarray analysis. It typically consists of global (mean or median) normalization followed by local (lowess) normalization. List the technical / mechanistic reasons for why this normalization would be required? Another way of asking the same question is this: "What would the experimentalist and experimental platform have to satisfy in order for such data normalization to be unnecessary?"

c) Use MeV or Excel or an equivalent approach to perform this normalization on your data. An "MA" plot is then used to assess the overall quality of a data set and whether it has been normalized properly (http://en.wikipedia.org/wiki/MA_plot). This plot graphs log (treated) – log (control) versus log (treated) + log (control). Generate and show two MA plots: one for normalized data and one for un-normalized data.

d) Compare the MA plots of normalized and un-normalized data. What is the difference between the plots? What is the improvement gained by normalization? How many significant genes are detected in either case? What is the volcano plot looks like before/after normalization?

Question 5: Genome Ethics

Several companies (23andMe, Navigenics, etc.) are now offering affordable opportunities for genetic tests that will allow individuals to identify genetic variants on a genome scale. Many individuals in the public are interested in obtaining these tests, while others are not. What are potential reasons that someone might choose to have such tests conducted? What are potential reasons that others might choose to not have the tests conducted?