Introduction to Bioinformatics – AS 250.265
Assignment 5

It is often very difficult to understand how something is done until one does it oneself. This is true for many of the concepts of gene expression taught in chapters six and seven of your textbook. While there is no way for you to conveniently isolate cells and measure actual RNA expression levels in this course, this assignment will familiarize you with the concepts of analyzing the data once it has been collected.

As you complete this assignment, you will be treating the data as ideal, error-free representations of reality. It is never appropriate to do this in your own experiments: you must be constantly vigilant to look for places where systematic or experimental error could occur, even in the most trivial experiments. Indeed, underestimating the importance of performing the simplest experiments correctly can often lead to dire results when those "simple" experiments are interpreted.

This assignment is worth 100 points toward the homework portion of the class. The laboratory you do in class constitutes 50 of those points, and this handout makes up the remaining half. Both parts are due on March 16 at the start of class. Electronic portions of the assignment should be submitted through the course website with appropriate filenames (e.g. `lb02-microarray.xls`). The written portions of the assignment may be submitted by hand or uploaded as a Word document or PDF file.

Question 1: Laboratory Number 2 (50 points)

Submit your solutions the laboratory on microarrays that you completed in class.

Question 2: Northern Plots (25 points)

You are a cancer researcher, investigating the expression of a particular gene in two cell lines. One of the cell lines is from a tumor that you have been able to grow in culture (sample A). The other cell line are normal cells obtained from non-cancer patients (sample B). You have isolated total RNA from both cultures and want to determine whether your gene is overexpressed or underexpressed in the cancer cells.

a. [written] Describe how would you separate the mRNA from the rest of the cell's RNA? (3 points)

In a normal agarose gel (the material most often used for nucleotide gel analysis), nucleic acid strands migrate through the gel toward a positively charged electrode (since the phosphate backbone is negatively charged). After the gel has been "run" with the power on for a given time, the larger strands will have traveled a shorter distance than the shorter strands, because their bulk prohibits them from traveling as easily through the gel. In fact, it is shown that the distance a strand travels is roughly proportional to the logarithm of the number of base pairs.

b. [written] Your gel will have three lanes: one for markers (so you can identify absolute size), one for your RNA from sample A, and one for RNA from sample B. In your marker lane, you have the markers of the following size: 50 bases, 100 bases, 200 bases, 500 bases, 1000 bases. In

lanes A and B, you have the following distribution of sizes: 1 fragment is 90 bases, two are 130 bases, four are 150 bases (one of which is your gene), one is 225 bases, and two are 400 bases. Draw a picture of how the gel will look when you visualize it with UV light (which will show all the RNA on the cell). (Hint: setting the absolute difference between the shortest and longest marker sequences will determine your scale.) (10 points)

c. [written] Given that the size of the human genome is 3×10^9 bases, and assuming that these bases are distributed at relatively randomly, how long should your radioactive probe sequence be to ensure that it only binds your RNA? (Hint: there are 4^N unique sequences for an RNA transcript of length N. Also remember there's no such thing as 0.4 bases.) (3 points)

You synthesize your probe and label it radioactively. Then, you transfer the pattern of your original gel onto nitrocellulose film and attach it there with UV light.

d. [written] After hybridizing the probe to the RNA immobilized on your nitrocellulose film, draw how the gel will now look when visualized with a phosphoimager, a photographic device that only exposes where radioactivity is present. Use approximately the same scale that you used for part (b). (2 points)

Modern phosphoimaging systems all provide a way to count the amount of radioactivity in each region of the sample. The amount of radioactivity is directly proportional to the amount of sample present on the nitrocellulose film.

e. [written] The phosphoimager reads 10,000 arbitrary radioactivity units on your band in sample A and 17,850 units from sample B. You know that normal cells when prepared as you have done produce 15 ng of RNA for your gene. How much RNA is present from your preparation of cancer cells? (2 points)

d. [written] You would like to assert your expression findings in a scientific journal. As this point, why might this be difficult to do? What can you do to improve the credibility of your results? (Hint: think from a statistical perspective.) (5 points)

Question 3: Digital Differential Displays (25 points)

When large collections of EST's are brought together from a single organism, and when the assumption is justified that the amount of mRNA is proportional to the amount of protein product in the cell, one can use the differences in the number of ESTs to examine differential expression in the cells, effectively doing a microarray analysis without leaving one's desk.

In this problem, we will explore the NCBI tool "Digital Differential Display," or DDD. It is a part of the UniGene project and can be accessed through the main UniGene web page. As a warm up to this problem, visit the DDD page on the NCBI website and read the documentation given on the main page.

In this problem, we will use DDD to examine the human disorder pulmonary fibrosis. It is a disorder of the lungs that causes them to become tough and unable to transfer oxygen to the blood. To perform this analysis, we will need to select three different types of cell lines: the first should be

a lung cell exhibiting normal characteristics. The second should be a lung cell that exhibits pulmonary fibrosis. The final class should be something unrelated: often you can combine datasets to make something that should be representative of "generic" cells.

Construct a DDD for the following human libraries: normal lung tissue (10395), lung fibrosis (10419), and mixed (16264). The "mixed" tissue will be our control here, although we could have selected several different types of cells from different stages of development.

a. [written] If you were investigating pulmonary fibrosis, what would be your first candidate to investigate? (3 points)

b. [written] Why might this candidate not turn out to be important? In other words, what would you have to prove for this protein to be important in pulmonary fibrosis? (5 points)

When subtracting the number of ESTs for one gene from another, one must assess the statistical significance of the result. One of the techniques that can be used to determine the probability that a difference is significant (the p-value) is a method called "Fisher's Exact Test." In this test, a table is relating the number of observations in two different pools. The table that is constructed is below.

	Gene of Interest (# of ESTs)	**All Other Genes (# of ESTs)**	**Total**
Pool A	a	A	$N_A = a+A$
Pool B	b	B	$N_B = b+B$
Total	c = a + b	C = A + B	$N_A + N_B = c + C$

In our case, the first pool is normal lung tissue and the second pool is pulmonary fibrosis tissue. As we set this table up, we know a, b, N_A and N_B. Fisher showed that the probability of constructing a table like this by chance is:

$$p = \frac{N_A!N_B!c!C!}{(N_A + N_B)!a!b!A!B!}$$

As usual, probability values of 0.05 or less are taken to be significant for a single analysis (i.e. a single pair of proteins.) The NIH site corrects this to account for the fact that multiple analyses are performed (a Bonferroni correction), but we will ignore that detail for now. From your digital differential display data, you should have all the information you need to calculate a p-value for your distribution (N_A and N_B are simply the number of clustered ESTs from each pool). Unfortunately, calculating p isn't as easily done as might be hoped.

c. [written] Why is it not possible to calculate the p-value using the standard mathematical functions on a computer or calculator? (2 points)

While it is not possible to evaluate p using simple means, there are mathematical tricks to make this number attainable using conventional computing machines. It is possible, for example, to compute binomial coefficients using very large numbers that would not be computable using simple minded

factorial functions. Recall that a binomial coefficient for *n* and *m* is combinatorically equivalent to the number of ways of selecting *m* items from *n* objects when order is irrelevant. The mathematical form for this value is:

$$\binom{n}{m} = {}_nC_m = \frac{n!}{m!(n-m)!}$$

Looking carefully at this expression, it should be clear that the value *n*! does not have to be evaluated explicitly. Instead, if *m* is greater than *n-m*, the expression can be rewritten as follows:

$$\frac{n!}{m!(n-m)!} = \frac{n(n-1)(n-2)...(m+1)}{(n-m)!}$$

Mathematically, the above expression looks no simpler, but this "reduction trick" will get us what we need. Specifically, it will allow us to calculate binomial coefficients that are much larger than what would otherwise be possible using simple-minded factorial functions (although the denominator must still be calculated using the simple-minded way). Practically, with the trick we can now calculate the p-value for the protein we identified in part (a). Of course, in the expression above, if *m* is less than *n-m*, then we simply assign the value of *n-m* to *m*—the result will be equivalent.

d. [written] Express the Fisher p-value in terms of the binomial coefficient $\binom{n}{m}$. (Hint: your final expression will have two binomial coefficient terms in the numerator and one in the denominator.) (5 points)

e. [programming] Write a python program that can calculate the p-value for the difference you observed in part (a). You will have to use the reduction method discussed above. (10 points)

For this problem, you may find it easiest to write three functions: First, write a factorial function. This function will be used to calculate the denominator in the expression above. Then, write a function called `choose(n, m)` that calculates the value of ${}_nC_m$ using the reduction trick. Next, write a function to calculate the p-value in terms of your expression in part (d). Finally, use the function you just wrote to print out the p-value.

Your final solution for part (e) should be fairly short: the "official" solution is about 30 lines of code, including several empty lines included for readability. Spending ten minutes thinking hard about the solution will likely save you an hour of finding bugs and incorrect behavior if you simply dive in without thinking.