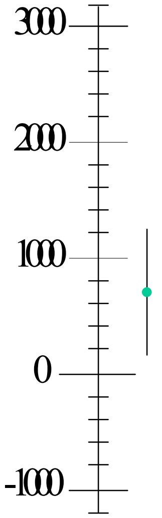
Biol 591 Introduction to Bioinformatics (Fall 2003)
Problem Set 6: Statistical Analysis of Microarray Data

P6.1. Let's get a visual picture as to how the correlation coefficient used by Golub et al works. For each specified gene, draw next to grid to the right the mean$_{ALL}$ and mean$_{AML}$ (both as circles) bracketed on each side by one standard deviation. An example is shown to the right: mean = 700, stdev = 500.



1a. Represent in this way the expression of thrombospondin-p50 (line 2759 of the training set) in ALL and AML patients.

1b. Represent the expression of activation-induced C-type lectin (line 4855 of the training set) in ALL and AML patients.

1c. In which gene is the difference in means greater? In which gene is the measure of correlation more extreme (i.e. distant from zero)?

1d. Examine the actual expression values for these genes to understand why this is the case.

P6.2. Compare **P(g,c)**, the measure of correlation used by Golub et al to **r**, the Pearson correlation coefficient, and **t**, the T-test statistic. To which is P(g,c) more similar? The most important element in judging similarity should be the ***types*** of quantities in the formulas.

$$r = \frac{\sum_{i=1}^{N} \left(\frac{x_{ALL,i} - \mu_{ALL}}{\sigma_{ALL}}\right)\left(\frac{x_{AML,i} - \mu_{AML}}{\sigma_{AML}}\right)}{N - 1}$$

$$P(g,c) = \frac{\mu_{ALL} - \mu_{AML}}{\sigma_{ALL} - \sigma_{AML}}$$

$$t = \frac{\mu_{ALL} - \mu_{AML}}{\left(\frac{\sigma^2_{ALL}}{N_{ALL}} + \frac{\sigma^2_{AML}}{N_{AML}}\right)}$$

P6.3. The 50 gene set used by Golub et al for predicting the ALL/AML class distinction was tested by seeing whether it correctly assigned each of the 38 patients in the training set to the correct class. On p. 532 (bottom right), you'll see that the test failed in 2 of the 38 patients.

3a. By examining Fig. 3B, predict which two patients were clinically diagnosed differently than predicted by the test (you'll have to look at the figure in color).

3b. What might account for the discrepancy?

3c. Test one of the possible reasons you came up with in 3b by examining the raw data of the training set (bringing it up in Excel will help here).

P6.4. Reproduce the analysis that led to the two curves of observed data in Fig. 2 of Golub et al (1999):

4a. Calculate in Excel the measures of correlation for all genes then plot them vs the number of genes whose measures of correlation exceed a specific value.

4b. Calculate in Perl the measures of correlation for all genes and output them to a file. Read them in Excel and produce the identical charts as in **1a.**

P6.5. Modify the Excel calculation and curve you generated in **P6.4** so that the T-test[1] is used in place of P(g,c) to score the ability of expression of a gene to distinguish ALL from AML. How do the two methods compare? What's a good way to compare the two sets of results?

P6.6. Run *Permute_training_set.pl* (available at the unit web site), a program designed to calculate the curves for permuted data shown in Fig. 2 of Golub et al (1999). The program takes a couple of minutes to run (to save time, I have it run only 20 permutations rather than the 400 specified in the paper).

6a. Upload the output file into Excel and generate a graph of the data, in a format similar to the curves of Fig. 2. Does the curve resemble the appropriate curve?

6b. Modify the program so that it prints out each permutation it considers and prints out the measure of correlation for each permutation that is ranked #10 (i.e., there are 10 genes with more negative measures of correlations). Compare this set of numbers to the value finally generated for y=10 in the file you uploaded into Excel in **6a**.

6c. The program goes through only 20 permutations. If you chose to go through systematically every possible permutation, how many would there be?

P6.7. Consider Fig. 2 from Golub et al (1999).

7a. Pick a point from the leftmost curve of the lefthand graph (Observed, High in ALL) and state carefully what that point represents.

7b. Pick a point from the rightmost curve of the righthand graph (median, High in AML) and state carefully what that point represents.

7c. All four curves in each graph appear to converge to a single horizontal line. Why is that?

7d. The curve representing the observed results appears to converge more slowly. Why is that?

P6.8. Modify *Class_prediction.pl* so that saves to disk data on just the 50 most informative genes (as judged by extreme P(g,c) scores). Then use *Vote_for_ALL.pl* (available at the unit web page) to assess each patient in the independent data set (also available at the same place), classifying each as either ALL, AML, or no decision, according to the method of Golub et al.

P6.9. You are trying to devise a tool to help in the early diagnosis of certain kind of hepatic cancer. To do this, you collect liver tissue from a number of patients diagnosed with the cancer as well as liver tissue from a number of people who died of unrelated causes. RNA extracted from each person is used to probe a human gene chip, and after appropriate normalization, and you look through the results for those genes that show the highest or lowest correlations with the distinction between the two classes.

9a. Using this set of genes, you test another group of patients with liver problems and find that gene expression in the set is not good in predicting those with liver cancer but predicts quite well those with cirrhosis of the liver. Explanation?

[1] To do a t-test in Excel, type in a cell TTEST(*range of ALL data, range of AML data,* 2, 3). "2" in the third position indicates that a two-tailed distribution is to be used (this means that you're considering how likely it is that a difference between the means of the observed ***magnitude*** might have arrived by chance (regardless of the sign of the difference. "3" in the fourth position indicates that the two populations do not necessarily have the same variance. This influences the calculation of the degrees of freedom.

OK. Start over. This time you find a set of genes that seems to work very well in predicting those with liver cancer. You found it using RNA from 8 patients with liver cancer and 12 patients with normal livers. To test the validity of this gene set, you shuffled the identities of the patients and recalculated the correlations found between gene expression in the set of genes and the distinction between 8 randomly chosen patients and the remaining 12 patients. The results are similar to those shown in Fig. 2 of Golub et al: the true distribution of correlation values was more extreme than even the most extreme 1% of the random permutations. If you look at the most extreme 0.1%, the curve is closer to the actual curve, and if you look at the most extreme 0.001%, the two curves are identical.

9b. Why is that?

9c. Your test enters the routine battery given to all adults over age 50 as part of their annual checkup, and you become a celebrity, riding the talk show circuit to talk about the need for prevention and the power of molecular medicine. One day you are surprised to find a complaint from a practitioner in Aberdeen, North Dakota. He tells you that the test gave a positive result for a nominally healthy male, age 55, but from liver biopsy and other procedures, it became clear that the patient had no liver tumor. The practitioner was surprised because of the low false positive rate of the exam (1 false positive per 1000 tests), but he was astonished when the very next patient to come into his office also tested positive for liver cancer and also turned out to be tumor-free by independent means. He wants to know whether others have complained about faulty test kits. Are you surprised? What explanation can you provide?

10. Discover Perl! In several programs this semester statements of the following forms have appeared:[2]

```
return sort(@results);
@current_correlations = sort { $a <=> $b } @current_correlations;
@training_set = sort { $$a[0] <=> $$b[0] } @training_set;
@hit_info = reverse sort { $$a[0] <=> $$b[0] } @hit_info;
```

Figure out what at least the first two forms do by making up and testing a program like:

```
my @array = ( put some values here );
put a statement here that uses sort on @array
print the array
```

Warning! $a and $b look like common variables, but they are not. Rather, they're more like place holders. Interchange them, but don't change their names.

[2] From FindMatches.pl, Permute_training_set.pl, Class_predictor.pl, and FindMotif.pl, respectively.