Biol 591 Introduction to Bioinformatics (Fall 2002): Problem Set 6 (Part 2) Statistical Analysis of Microarray Data

PS6.2. What is your level of confidence with each of the following statements, and why?

A. Patients with ALL have a different set of genes than patients with AML.

B. Patients with ALL have different levels of gene expression than patients with AML.

C. Patients with ALL have a different set of proteins than patients with AML.

PS6.3. You add RNA from a patient known to have ALL to a filter spotted with a large number of human genes. In parallel, you add RNA from a patient known to have AML to a different filter spotted with the same human genes. You find that the signal for a certain gene is much higher in the first filter than in the second.

PS6.3a. Does this mean that the gene is expressed at a higher level in the ALL patient than in the AML patient? *Hint: No.*

PS6.3b. Think of some reasons why not.

PS6.4. In Fig. 2 from Golub et al, all four curves in each graph appear to converge to a single horizontal line.

PS6.4a. Why is that?

PS6.4b. The curve reprenting the actual results appears to converge more slowly. Why is that?

PS6.5. 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.

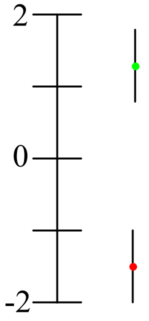
PS6.5a. By examining Fig. 3B, predict which two patients were clinically diagnosed differently than predicted by the test (you'll have to look at the figure in color).

PS6.5b. What might account for the discrepancy?

PS6.5c. Test one of the possible reasons you came up with in PS6.5b by examining the raw data of the training set (bringing it up in Excel will help here).

PS6.6. Let's get a visual picture as to how the correlation coefficient used by Golub et al works.

PS6.6a. Consider the genes with the highest correlation coefficient and the lowest correlation coefficient (you'll get these from your solution to problem PS6.1, or you can pull them off of Fig. 3B). Draw a diagram for each, with the Y-axis giving the value of the correlation coefficient. Show for each the mean value for ALL patients, bracketed on each side by the standard deviation, and the same for AML patients. (see below for a picture of what I mean by a mean bracketed by the standard deviation).



PS6.6b. Do the same for two arbitrary genes that didn't make the cut.

PS6.7. 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.

PS6.7a. Using this set of genes, you test another group of patients with liver problems and find that gene expression in the set does not predict those with liver cancer but predicts rather well those with cirrhosis of the liver. Explanation?

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.

PS6.7b. Why is that? (By the way, I don't mean "close"; I mean "identical")

PS6.7c. 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, 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?