BIOL591: Introduction to Bioinformatics
Microarray Analysis Case Study: Golub et al (1999)

Outline:

I. Overview of article
II. Materials and Methods (first run through)
III. Results – Class prediction

I. Overview of article

As usual in bioinformatic problems, the Scenario presents us with a situation of too much information and not enough insight. (By the way, if you haven't yet looked at the story for this unit, now might be a good time to do so). The story leaves us with the admonition that we need to know something about statistical measures and how they may be applied to microarray data. Rather than address our ignorance more generally, let's instead consider how a specific problem was solved, as described in the paper by Todd Golub, et al, available from the Scenario web site or from the Calendar. Download the article and go for it.

First of all, take a 10-second tour of the article to see what you have before you. It is an unsegmented article (i.e. no explicit sections marked **Introduction**, **Materials and Methods**, and so forth). As a result, it may be difficult to find what ought to be the main point of any research article: what was found and *how* it was found.[1] Much essential information is to be found in figure legends and (surprisingly) at the back of the article in tiny notes. While you can safely ignore the references interspersed in the text of most articles, to do so with a *Science* research article risks making it incomprehensible. Moral: Check out those endnotes (given as italicized numbers in parentheses).

Confronting an article without sections, I find it useful to supply them myself. Skimming through the article, I see that most of the first page describes the problem in general terms. I'll define the **Introduction** section as extending to the end of the paragraph beginning "*We began with class prediction…"*. This paragraph sets forth the general aims of the article, a common strategy to close the **Introduction**.

One typically finds a **Materials and Methods** section next, and the last full paragraph on page 1 would fit well into that section. Unfortunately, much else that would also fit in are strewn throughout the article in figure legends and endnotes. There really isn't a **Materials and Methods** section, which will make our life difficult at times.

The **Results** Section begins at with the last line of the first page. However, this is an article that addresses multiple issues: how to predict whether a patient falls into one of two known classes of leukemia (class prediction); how to define classes by statistical means when such distinctions have not already been made (class discovery). The **Results** section has been divided along these lines, even if the subsections are not explicitly labeled. We end up with two articles connected by a common **Introduction** that are organized along similar lines (see outline on next page).

In these notes, I'll confine myself just to the results pertaining to class prediction. That's plenty for one day!

[1] The journals *Science* and *Nature* hold to the unsegmented style, and I find articles in those journals to be amongst the most difficult to understand in the way I want to understand them.

Outline of Golub et al (1999)

I. Introduction *(p.531, pars. 1-7)*
Begins: *The challenge of cancer treatment…*
Ends: *We began with class prediction:…whose appearance is highly similar.*

II. Materials and Methods *(p.531, par. 8)*
Begins and ends: *Our initial leukemia data set…scanned microarray image.*
Includes: Figure legends and many endnotes.

III. Results *(p.531, last line to p.535, par. 2)*

III.A. Class prediction *(p.531 last line to p.533, par.5)*

III.A.1. Are there genes whose expression correlate with AML vs ALL?
Begins: *The first issue was to explore whether…*
Ends: *For the 38 acute leukemia samples…based on expression data.*
Figures: *Fig. 1A* and *Fig. 2*

III.A.2. Construction of a class predictor
Begins and ends: *The second issue was how to use… threshold of 0.3.*
Figure: *Fig. 1B*

III.A.3. Validation of class predictors
Begins: *The third issue was how to test…*
Ends: *The choice to use 50… the AML-ALL distinction.*
Figure: *Fig. 3*

III.A.4. Discussion of genes found to be informative
Begins: *The list of informative genes…*
Ends: *We had expected that…cancer pathogenesis and pharmacology.*

III.A.5. Correlation between class prediction and response to chemotherapy
Begins and ends: *The methodology of class prediction…this hypothesis.*

III.B. Class discovery *(p.533, par.6 top.535, par.2)*

III.B.1. Introduction
Begins: *We next turned to the question…*
Ends: *To cluster tumors, we used… of the data points nearest to it.*

III.B.2. Application of self-organizing maps (SOMs) to known case (AML-ALL)
Begins: *We applied a two-cluster SOM…*
Ends: *We then tested the class predictor… previous biological knowledge.*
Figures: *Fig. 4A and Fig. 4B*

III.B.3. Extension of SOM to discern finer distinctions within AML and ALL
Begins: *We then sought to extend…*
Ends: *We again evaluated these classes… primarily of B-lineage ALL.*
Figures: *Fig. 4C and Fig. 4D*

IV. Discussion *(p.535, par. 3 to end of article*

IV.A. Discussion of class discovery
Begins: *The class discovery approach thus…*
Ends: *Class discovery methods… remaining genes.*

IV.B. Discussion of class prediction
Begins: *We also describe techniques for class prediction…*
Ends: *Most importantly, the technique… eventual outcome is known*

II. Materials and Methods (first run through)

I'll let you scan the Introduction on your own, then we'll proceed with the Results concerning class prediction. The authors say that they used a data set from 38 patients, probing expression from 6817 genes. You can see the data by downloading it from the Scenario web site (first data set of 38 patients). Upload the file into Excel and examine it.

SQ1. Can you identify the 38 patients? (*Warning*: no one said they were in order) Which are ALL patients (color them light yellow)? Which are AML patients (color them light blue)?

SQ2. Can you identify the 6817 genes? Why are there so many genes? Take a look at the names of the first few. They certainly don't sound like normal human genes. Why are they there?

SQ3. From the description in the article (and your own prior knowledge), what sense can you make of the format of the data in this data set? For example, what is the significance of negative numbers?

III. Results – Class prediction

Could it be, ask the authors, that somewhere in this mix of genes are some whose expression levels correlate well with the distinction between ALL and AML patients? To determine this, they sorted the genes "…by their degree of correlation (*16*)." What does that mean? Nowhere in the body of the article will you find a clue, unless you follow the indicated endnote. Visit endnote 16. There's a lot of jargon here, but if you can wade through it, you'll find a critical and readily understandable core, the definition of a measure of correlation, **P(g,c)**, defined as the difference of the means of expression of a gene (one class minus the other) divided by the sum of the standard deviations of gene expression over the two classes. This measure makes a lot of sense (Fig. 1).

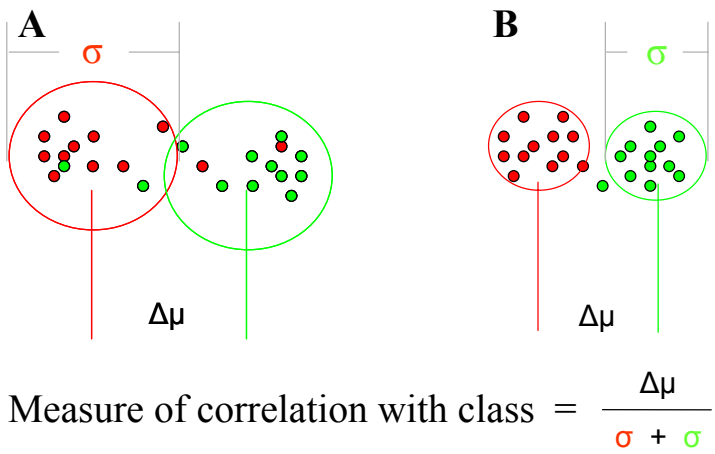

Fig. 1: Relationship between correlation and means and standard deviations. Means of each of the two classes (red and green) are represented by vertical lines. The corresponding standard deviations are represented by circles. (**A**) There is a relatively large difference in means between the two populations but the standard deviations are also large. (**B**) Both the difference in means and standard deviations are relatively small. You can see intuitively that these two populations are better separated than those in **A**, even though the difference in means is smaller.

Let's do it. On the Excel spread sheet, calculate in cell CA2 the average expression, μ_1, of the gene in line 2 (AFFX-BioB-5_at) over all ALL patients.[2] Now in cell CB2, calculate the average expression, μ_2, of the same gene but over all AML patients. In cells CC2 and CC3, calculate the standard deviations, σ_1 and σ_2, over all ALL and AML patients, respectively.[3] Finally, calculate **P(g,c)**, using the formula supplied in the endnote. Once you've done this, extend the five formulas from the top to the bottom of the page.[4] Now you can sort the page by correlation measure.

SQ4. What is the smallest measure of correlation? What is the largest?

SQ5. Examine the genes at the top and the bottom of the list. See any that sound interesting?

Of course in any large collection of numbers you expect some to be large some to be small. How can we tell, asked the authors, whether the extreme measures of correlation are larger or smaller than you would expect by chance in a collection of genes as large as ours? We've encountered a similar question before: How can we tell whether an alignment score is higher than you'd expect by chance in a database of the size examined? In that case, an **E** score was calculated, representing the number of times you'd expect to get a match that good or better by chance. In the current case, it isn't easy to see how you can ***calculate*** an expected number of genes with a given measure of correlation. The authors evidently didn't see a way either, so they used the approach that would occur first to a bioinformatician: run a simulation.

By chance… as always, that's not easy to accomplish. Should we just make up micrarray values? How? The authors solution is both elegant and generally applicable. Again, you'll find nothing in the body of the article that would offer you understanding of their method. Instead, you'll need to go to the endnote and the rather obscure legend to Figure 2 in the paper. The key phrase is *"…by permuting the coordinates of **c**."* That ***c*** was defined in the previous endnote and in Figure 1 as… well, never mind. What they did was simple: take the top row of the table of data and shuffle the values while leaving the data alone. The effect is that the labels indicating ALL patient (1 through 28) are randomly assigned and so sometimes land on an ALL column and sometimes not. For each permutation, the calculations of μ, σ, and **P** are repeated.

With the relationship between the columns and patient identity scrambled, you'd expect that any apparent correlation between the ability of a gene to predict a mythical class distinction would be just part of the variation that would arise by chance. We are talking about a lot of calculations now. How can it be represented in a way that the overall message might be grasped by the reader? Figure 2 is their attempt to do this.

To understand Figure 2, let's first see how it represents the actual measures of correlation. The X-axis is the measure of correlation that you calculated yourself, but the Y-axis is somewhat more mysterious. The legend is of little help. The Y-value of each point is meant to represent the

[2] To do this, type in =SUM(C2:BD2), summing expression from the first to the last ALL patient. You might be concerned that you're summing letters as well as numbers. Don't worry. Excel ignores the letters.
[3] To do this, type in =STDEV(…) putting in the range of columns over which to calculate.
[4] Select the five boxes and copy them, then select the cell CA3 and scroll down to the bottom of the page and (with the Shift key depressed) select cell CE7130. Paste, and the results of the formula should appear in all the selected cells.

number of genes with correlation coefficients better than that of the X-value for the point. My explanation is probably not much better, so let's make the chart ourselves.

Go back to your Excel file (which you left sorted by measure of correlation) and create a new column, CF, numbered from 1 to 7030.[5] Now you can plot column CE (number of genes) against CF (measure of correlation).[6]

SQ6. Compare your graph with that of Fig. 2. What part of Fig. 2 did you graph? What discrepancies do you observe? Do you see why?

SQ7. Precisely how many genes have correlation coefficients better than –1.0?

It isn't easy to do the permutations in Excel (that's why we have general computer programming languages like Perl), but you can get the idea what will happen by swapping five of the AML patients with five of the ALL patients (***be sure to save the original chart!***). If you leave all the formulas intact, the graph should be transformed into a blob, because the rows are no longer sorted by measures of correlation. If you resort the table, the graph should reappear, but shifted to the right.

That's just one permutation. The authors did 400 of them. How can you represent 400 curves on a single graph? I still can't figure it out from what's in the paper, but here's my best guess. The curve labeled "1%" indicates that in 1% of the permutations (i.e. 4 times) there were as many or more than the indicated number of genes that had an indicated measure of correlation. For example, in only 1% of the permutations did one gene or more have a measure of correlation equal to or greater than about 0.85 (where the measure was calculated with respect to ALL). Figure 2 of the paper shows curves representing the top 1%, 5%, and 50% (which one is that?) of the permutations.

SQ8. Complete the following sentence: *In only 5% of the permuted data sets were there X genes or more with a measure of correlation equal or greater than 0.5 (considering the measure of correlation from the perspective of ALL).*

SQ9. In the actual data, 397 genes had measures of correlation greater than 0.5. How did I know this? About how many randomly permuted data sets had at least 397 genes with measures of correlation greater than 0.5? About how many genes had measures of correlation greater than 0.5 in the top 5% of the randomly permutated data sets?

SQ10. Suppose you want to identify all genes that can reasonably be expected to correlate with the ALL-AML distinction by reasons other than chance. I'll define "*reasonably be expected*" by saying that only 5% of random permutations will have even 10% the number of genes in my collection, and so 90% of my collection are probably valid. What cutoff value should I chose for the measure of correlation?

[5] Type the value 1 in cell CF2 and the formula =CF2+1 in cell CF3. Now copy CF3 to the bottom of the column.
[6] Click the cell CE2 and (with the Shift key depressed) CF7130. Then click the chart wizard icon in the tool bar (it looks like a multicolored bar graph). Select XY (Scatter), unconnected points as the subtype, and click Finish. Finally, click on the Y-axis and change it to log scale.

I've focused on Figure 2 of the paper, as it seems to me the core of the authors' method of finding its class of predictors. Figure 1A is more difficult for me to understand. However, I think that it can be derived from Figure 2 as shown in Fig. 2 to the right. The concentric circles represent a slice through the curves from shuffled data, taken at a certain value for the measure of correlation. If the data is random, then a small number of genes should have that measure or better. If in fact there are considerably more than number, then it is reasonable to suppose that the correlation occurred for reasons besides chance.

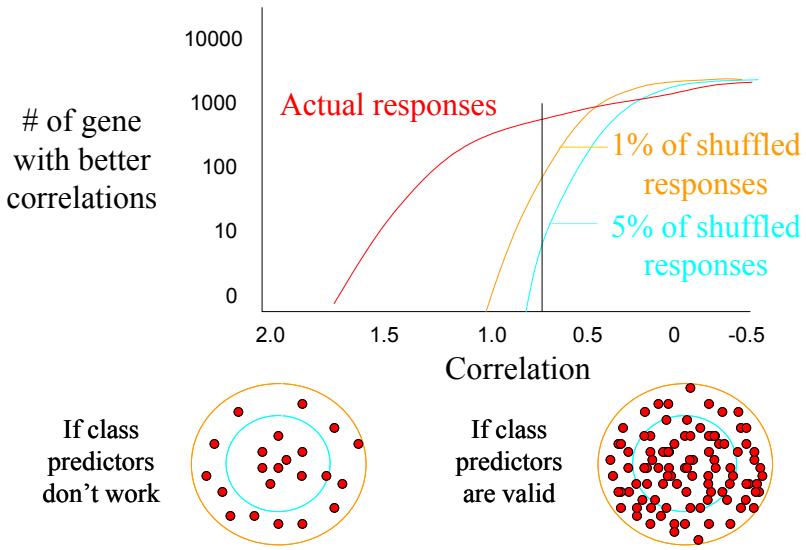

Fig. 2. Relationship between Figure 1 and Figure 2 of Golub et al (1999).

That ends the collaborative portion of our program, corresponding to a consideration of III.A.1 in the outline on page 2. Now you try to go through the next section, III.A.2, where a set of class predictor genes are used to decide whether a patient belongs in class ALL or class AML. Each of the genes in the set vote for either ALL or AML and the sum of the votes determines the assigned class. The main challenge is figuring out how votes are determined and counted. The article is considerably clearer here (in my opinion) than in the previous section, so you stand a good chance figuring out a voting scheme from the article itself. Figure 1B (and its legend) should help a good deal, and one of the endnotes should fill in what gaps remain.

SQ11. What is the formula for determining the vote of a gene in the class predictor set?

SQ12. What does each element of the formula mean?