GS01 0163
Analysis of Microarray Data

Keith Baggerly and Kevin Coombes
Section of Bioinformatics
Department of Biostatistics and Applied Mathematics
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
kcoombes@mdanderson.org

13 October 2005

Lecture 12: Differential Expression and Borrowing

- Modelling Redux
- Borrowing Strength Across Genes
- Combining Borrowing with Ranks
- Borrowing Tail Ranks
- Conditioning on Biology

The Modelling Punchline

Incorporating external information can help sharpen our inferences.

Incorporating such information often goes by the name of modelling, but it can also be viewed as “conditioning on relevant subsets of information”.

The crux of the problem is defining precisely what constitutes a “relevant subset”, which includes what we mean by “relevant”.

Types of Conditioning

One of the more common types of conditioning is to assume that some other quantity being measured shares some distributional characteristics with measurements of the quantity of interest.

In shorter words, we can use other data to give us better estimates of standard deviations, or the shape of the distribution, or so on. We saw this last lecture with the use of a third group of microarray measurments to sharpen inferences about differences between the first two.

Are Other Genes Relevant?

Are there similar characteristics to microarray measurements of different genes?

If there are, how can we use them?

Most frequently, the answer to the first question is assumed to be yes based on empirical observations. Occasionally, a modelling of the underlying physical processes can further suggest the nature of the similarity.

An Example

Our first example: normalization.

This can assume either that “most genes don’t change” (single scaling factor normalization) or, more stringently, that “the quantiles of the intensity distributions should be about the same” (loess normalization).

Are the Assumptions Valid Here?

In general, yes. In checking normalization methods, people have produced some nice-looking smooth curves, but the latter in particular are working under the assumption that if we start with genes of the same rough level of expression, the distributions of values when nothing is going on will be about the same.

Other Extensions of Borrowing

borrowing strength on the p-value scale.

BUM

Empirical Bayes

Extending this idea to diff. e.

What can we do here?

Say that we have our standard question of trying to compare the levels of a given gene in two different groups, A and B.

How can we change the t statistic?

As before, our best guess about the central value of the gene in each of the groups is driven by the observed values for that gene:

$\bar{x}_A - \bar{x}_B$ is unchanged.

Pooling variance estimates

What can use to improve our estimate of the variance?

How about the variance of all of the genes?

This is likely to be too much.

What if we just use the genes that are close by in terms of overall (average) intensity?

This type of procedure makes some of the same underlying assumptions as the loess normalization, which also works with “locally similar” data.

What does this produce?

a stabilized variance and a “smooth” t-test.

This idea has been independently reintroduced in several forms.

Baldi and Long (2001) use a Bayesian approach to trade off between the sample variance for the gene of interest and the pooled variance estimate. This is known as a “shrinkage” estimate.

Newton et al (2001) use a Gamma-Poisson model which achieves the same effect.

Some More Papers

The “fudge factor” in the denominator of SAM is of this variety.

Baggerly et al (2001) use a Beta-binomial model based on the use of variance derived from replicate spottings to derive the the locally pooled variance estimate; there is no weighting tradeoff with the actual variance observed.

Are We Using It?

This last paper is the basis for some of the “standard analyses” done at MD Anderson. All of the above tests were developed in the context of cDNA microarrays.

We’ve also used it to analyze data from nylon membrane arrays (Coombes, 2001).

Why might the assumption be valid here?

There are plausible reasons why the variance of microarray readings should change in a smooth fashion as the overall intensity increases.

These have to do with lognormal expression values, background subtraction, and thresholding.

But we’re implicitly assuming that “most genes aren’t too correlated” so a variance estimate derived from several genes will be close.

Implications of Independence

We note that this assumption of independence means that in terms of trying to define the overall variance distribution, it is not a good idea to choose a bunch of genes known to be biologically related as our relevant subset. It is interesting to explore these connections, but here we are seeking reinforcement of a story by looking for groups of genes having similar expression patterns.

Does this help a great deal?

In our earlier discussions, we noted that better characterization of the variability did improve things, but maybe not so much.

However, that assessment was predicated on our having a good idea of the underlying distribution to begin with. If the data are skewed or subject to frequent outliers, things can get worse.

Skewness, Outliers, and Bears?

This, of course, is why we often shift to rank tests which don't depend on the particular shape of the distribution. But, as we saw last time, the discrete nature of the ranks may preclude a rank test from yielding a small p-value even when something extreme is going on.

Linking Borrowing and Ranks

Small p-values, however, can be obtained if we have more “effective samples” with which to characterize the underlying distribution, leading us to combine the idea of borrowing strength across genes with the idea of using ranks to remain less sensitive to the particular shape of the underlying distribution.

The Relative Rank Test

Oddly enough, we haven't seen that much written about borrowing with ranks, but here goes.

Assume that we are interested in deciding if the levels of gene g are different between two groups A and B, and that g is for the most part contained within a set of genes G having similar null distributions.

The standard procedure (Wilcoxon) is to rank the $n_A + n_B$ values of g and sum the ranks of those in one of the groups (say A). The p-values are then computed by permutation and counting arguments.

The Relative Rank Test

For Wilcoxon, we note that we could just as easily have worked with the difference in average ranks for A and B, respectively, as the total must stay fixed.

Here, we rank all $G * (n_A + n_B)$ expression values within the "relevant set" G, and focus on the difference between the average ranks for gene g in groups A and B. Here, the choice of just one sum (say the A ranks) or the difference does matter because there are intervening values present.

What does this potentially buy us?

- The ability to get small p-values
- The ability to get large p-values
- The ability to differentiate between “extreme cases”
- Some robustness against outliers (we lose some of this relative to the Wilcoxon test, however) or different distributions

What does this potentially cost us?

accuracy, if the distributions are starkly different (eg, including high real variability genes with low real variability genes).

The traditional borrowing of strength focuses on a single number (such as the variance) and presumes that will be stable. Rank sharing makes stronger distributional assumptions.

Some Math

What can we say about the distribution of the difference $\bar{r}_A - \bar{r}_B$?

Well, if G is large, then we can effectively ignore the discrete nature of the rank distribution. To make things easier (on me), let's divide the ranks by $G * (n_A + n_B)$ so that we have values ranging from 0 to 1.

Some Math

When nothing is going on, the expected difference in average ranks is 0. The variance of a single draw from a uniform distribution is 1/12, so the variance of the difference is

$$\frac{1}{12}\left(\frac{1}{n_A}+\frac{1}{n_B}\right).$$

Approximate normality kicks in fairly quickly, and for finite samples the shape involves the repeated convolution of uniforms (giving B-splines).

Some Outcomes

So, what do the results of using this test look like?

Looking at the prostate cancer data, the values returned by the relative rank test look intermediate between those of Wilcoxon and t-tests. By using multiple genes to more finely partition the ranks, we recapture some of the parametric sensitivity of the t-test. Here, the data were approximately normal to begin with.

Relative Tails?

Can the relative rank approach be used to help with the tail rank test for biomarkers?

Well, in the description of the tail rank test given earlier, it was stated that we needed to specify two things before using the test:

ψ, the desired specificity of the biomarker, and

γ, the bound on the FWER.

The way that the relative rank approach can help is hidden in the way the value of ψ is used.

Defining Quantiles

Specifically, in order to use the tail rank statistic we need to estimate, for each gene g, a threshold value τ_g such that $P(X_g < \tau_g) = \psi$

The difficulty is that τ_g represents a tail quantile of a distribution, because we want ψ to be close to 1. Tail quantiles are hard to estimate well unless (a) you have lots of samples (which we typically won't) or (b) you have some knowledge of the parametric form of the distribution of X_g.

Tradeoffs

The question then becomes one of which assumption is more plausible:

that we know a parametric form well enough to characterize tails,

or

that the distribution of expression values in a given intensity range when nothing is going on may be similar enough across genes for them to be productively combined.

The Upside

If we collect the ranks for G genes as above, and focus on the results in the control samples, then our “effective sample size” will increase, typically to the point where we can get point estimates of some extreme quantiles (such as 99%).

Further Extensions

What other ways can we use the relative rank approach?

Kruskal-Wallis can be revisited.

Is there some way to build sensitivity into the tail rank procedure? Probably not, since we're assuming that the behavior of the biomarker is "atypical" for the subset that it flags as interesting.

Sensitivity and Biomarkers

There is an asymmetry here, which reflects the asymmetry in the question we're asking.

For good biomarkers, we want the specificity to be high, but we're willing to live with low sensitivity.

The rationale for this is that the heterogeneity of the disease suggests that if markers are to be productively used, this should be as part of a panel.

We don't yet know how to assemble a good and parsimonious panel.

We may be able to assemble a good panel.

Conditioning on Biology

Above, we’ve tended to group probes as “similar” based on their observed expression values, giving intensity-dependent variance estimates, normalization, and so on.

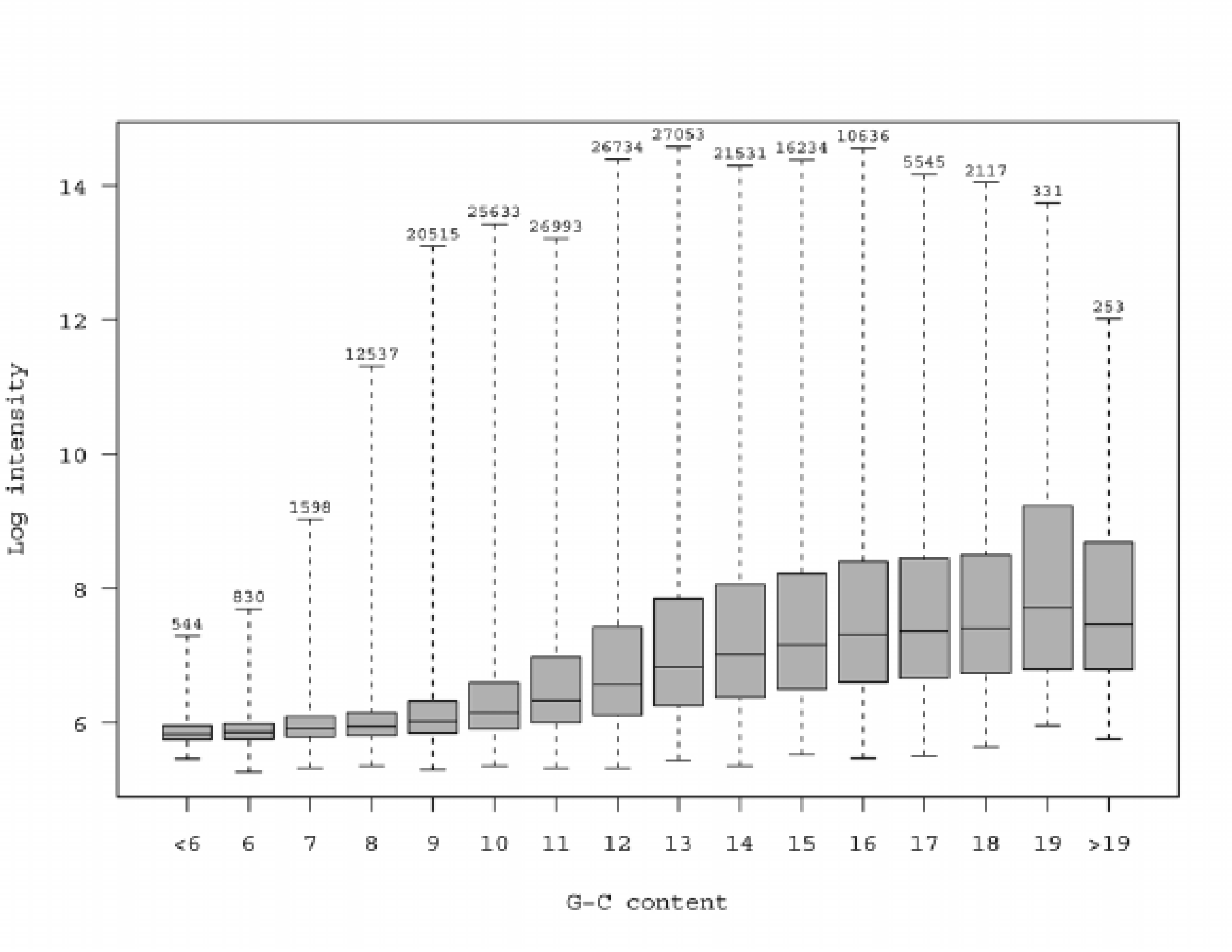
Are there other ways of gathering probes or probesets into groups that we might expect to have similar behavior with respect to baseline brightness and variability?

Exploiting Sequence: PDNN and GCRMA

Some of the more effective methods work by grouping probes in part based on their sequence.

What properties of a probe sequence might make it better at binding, or make the bonds it does produce stronger?

Intensity by GC Count



G-C has 3 hydrogen bonds, A-T has 2

Defining Affinity

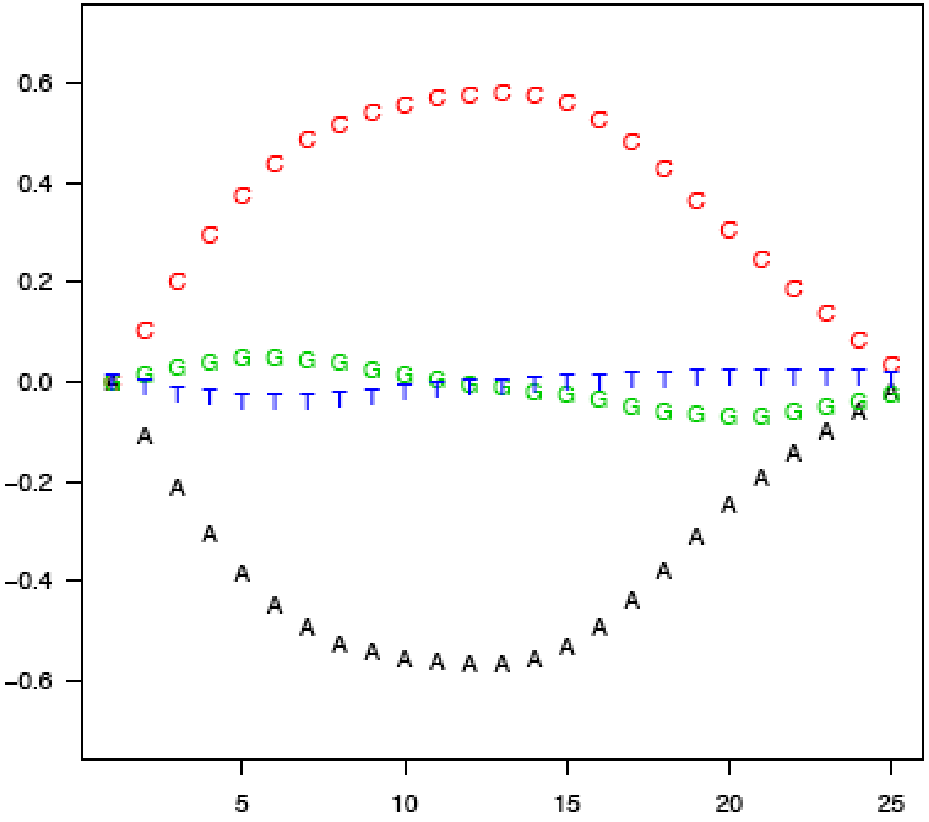
A model for intensity

$$log(PM_{ij}) = \Theta_i + \alpha_j + \epsilon_{ij}$$

A formula for affinity (Naef and Magnasco, 03, Phys Rev E v68)

$$\alpha = \sum_{k=1}^{25} \sum_{j \in \{A,T,G,C\}} \mu_{jk} 1_{b_k = j}, \quad \mu_{jk} = \sum_{m=0}^{3} \beta_{jm} x^m$$

Probe Affinity by Position



Why the asymmetry?

Invoking GCRMA in R

uses some information from the MM values! Follow outline in BioC...

```
library(gcrma)
library(hgu95av2probe)
library(hgu95av2cdf)
library(affydata)
data(Dilution)
ai <- compute.affinities(cdfName(Dilution))
Dil.expr <- gcrma(Dilution,affinity.info=ai)
```