Using Categories to Model Genomic Data

R. Gentleman

May 18, 2005

Introduction

The analysis of genomic data is an important challenge in bioinformatics. The particular aspect of that problem that we will address is the analysis of gene expression data in conjunction with category data. Category data, are data that map from specific entities (genes in this case) to categories, or classes. In some cases the mapping will be a partition, so that each gene maps to one and only one category, but in most cases genes will map to multiple categories. One important aspect of category data is that the categories and the mappings from entities to categories are known *a priori* and are not determined from the experimental data.

There are very many biological examples of category data, and to some extent they may be approached using similar statistical methods. Examples include the mapping from genes to chromosomes (which is nearly a partition), the mappings from genes to pathways, the mappings from genes to GO (**?**) classifications, or the mappings from genes to protein complexes. The categories themselves may have complex relationships, as is the case with GO and protein complexes, but for now we concentrate on the mappings between genes and categories.

To make some of the concepts concrete and to provide extensive, but related, examples we will make use of a microarray data set (**?**). The data come from a clinical trial in acute lymphoblastic leukemia (ALL). We will focus our attention on the patients with B-cell derived ALL, and in particular on comparing the group with BCR/ABL (9;22 translocation) to those with no observed cytogenetic abnormalities.

Category analysis is similar to the approach taken in **?**. However, category analysis can be viewed as an extension of that methodology that makes its application both simpler and richer. Among the important concepts is the notion that the search for sets of differentially expressed genes is not always the right approach for analyzing gene expression data. Given that a microarray typically measures the levels of coordinated gene expression averaged over a few thousands of cells it seems that a more holistic approach is sometimes warranted. We are often more interested in categories where the constituent genes show coordinated changes in expression over the experimental conditions than in sets of differentially expressed genes. The methods presented in this paper are one approach to taking a more global view.

Let's consider the comparison of two different groups, or phenotypes and we assume that there are some number of DNA microarrays that have been obtained for each group. The actual method of comparing the expression levels in the two groups is, in some sense, irrelevant to the subsequent discussion and readers can easily substitute their own favorite methods. We refer readers to **?** or **?** for a general discussion of some of the issues involved in gene filtering. Here we will use a t-test.

First we subse the ALL data to the two phenotypes that we would like to compare, those with BCR/ABL and those with no cytogenetic abnormalities, labeled NEG.

```
> data(ALL)
> esetA <- ALL[, intersect(grep("^B", as.character(ALL$BT)), which(as.character(ALL$mol) %in%
+     c("NEG", "BCR/ABL")))]
> esetA@annotation = "hgu95av2"
> esetA$mol.biol = factor(esetA$mol.biol)
> lowQ = rowQ(esetA, floor(0.25 * 79))
> upQ = rowQ(esetA, ceiling(0.75 * 79))
> selected <- (upQ - lowQ) > 0.5
> sum(selected)

[1] 4507

> esetASub <- esetA[selected, ]
> BCRcols = ifelse(esetASub$mol == "BCR/ABL", "goldenrod", "skyblue")
> library("RColorBrewer")
> cols = brewer.pal(10, "RdBu")
```

0.1 Category Analysis

We have a set of data where there have been G measurements on each of n samples. We use $\mathbf{E}$ to denote the G by n data matrix. We consider the case where a univariate test statistic can be computed for each entity (gene in our case) and denote the resulting G vector by $\mathbf{x}$.

There is a given fixed set of categories, $\mathcal{C}$, and a set of entities (genes) $\mathcal{G}$ from which we can compute the incidence matrix $\mathbf{A}$, where $a[i, j] = 1$ if entity j is in category i. The question of interest is the identification of the elements of $\mathbf{z} = \mathbf{Ax}$ that are unusually large or small.

0.2 Implementation

Consider the two-sample problem. Assume that there are n microarrays available, and that they have collected data on n samples under two conditions. Suppose that we have chosen to use a test statistics, T. There are several different methods of generating values of $\mathbf{x}_T$ under the null hypothesis that there is no difference between the two conditions. These include permuting the sample labels, carrying out a bootstrap simulation, or using any one of a number of other methods for generating a reference distribution. Once this reference distribution, $\{\mathbf{x}^b\}_{b=1}^B$ has been computed, it induces a distribution on $\mathbf{z}$, where $\mathbf{z}^b = \mathbf{Ax}^b$. Hence, for each z_i we can compute marginal tests of whether that particular z_i is extreme relative to the joint distribution.

0.2.1 Parametric Assumptions

Suppose that $\mathbf{X}$ is multivariate $N(\mu, \Sigma)$. The statistics are computed as $\mathbf{Z} = \mathbf{AX}$, and hence $\mathbf{Z}$ also follows a multivariate Normal distribution with mean $\mathbf{A}\mu$ and variance $\mathbf{A}\Sigma\mathbf{A}'$. But if Σ is unknown then it is not possible to carry out inference on $\mathbf{Z}$. For our situation Σ is too large to easily be estimated.

We note that if $\mathbf{X}$ is made up of two sample t-statistics with n reasonably large then the elements of $\mathbf{X}$ are approximately $N(0,1)$ random variables. If the genes were independent $\mathbf{Z}$ divided by the square root of the row sums of $\mathbf{A}$ will itself be approximately multivariate Normal with mean zero and Σ will be the m by m identity matrix. We use this approximation below.

To carry this out we make use of the `rowttests` function in the annotate package.

```
> ttests = rowttests(esetASub, "mol.biol")
> fL = findLargest(geneNames(esetASub), abs(ttests$statistic),
+     "hgu95av2")
> fL2 = probes2Path(fL, "hgu95av2")
> length(fL2)
```

```
[1] 1015
```

```
> inBoth = fL %in% names(fL2)
> fL = fL[inBoth]
```

In the computation to get `fL2` we first found all duplicate mappings to LocusLink identifiers and then selected the one with the largest t–statistic. Note that we passed the absolute value of the observed t–statistic in and so obtain the most extreme value. Then we remove all LocusLink identifiers that do not map to any known pathway and find that we have 1015 genes left.

Next we create the adjacency matrix that maps genes/probes to pathways. We will make the decision that we are not interested in pathways that have fewer than 5 members that are in the experimentally observed genes.

```
> Amat = t(PWAmat("hgu95av2"))
> AmER = Amat[, names(fL)]
> rs = rowSums(AmER)
> AmER2 = AmER[rs > 5, ]
> rs2 = rs[rs > 5]
> nCats = length(rs2)
```

There are 94 pathways (categories) that will be used for the analysis.

```
> tA = AmER2 %*% tobs$statistic
> tA = tA/sqrt(rs2)
> names(tA) = row.names(AmER2)
```

And now we can examine the resultant qq-plot, which is shown in Figure 1.

In Figure 1 we see that there is one pathway for which the aggregate statistic seems to be unusually large, and negative. While there are a number of pathways that seem to have elevated levels of activity among those with BCR/ABL none of the stand out the way that the ribosome pathway does. We can find that pathway and examine the data a bit more. Then in the next section we make use of the permutation approach to assess significance.

```
> byTT = names(tA)[tA < -5]
```

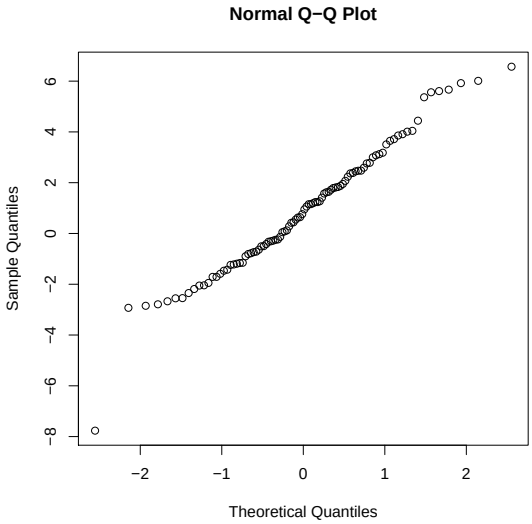

Figure 1: The per category qq-plot.

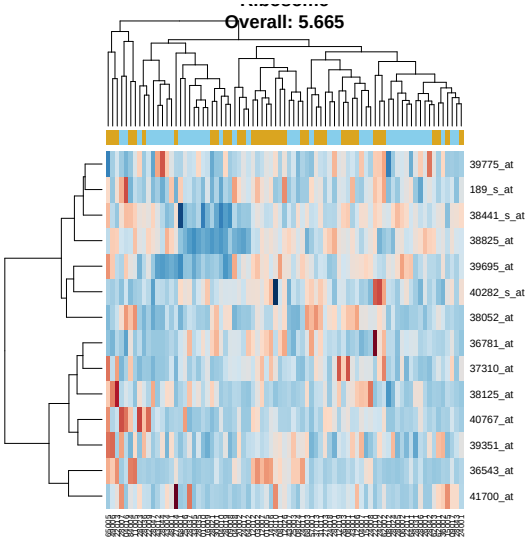

Figure 2: The mean plot for the category with the smallest p-value.

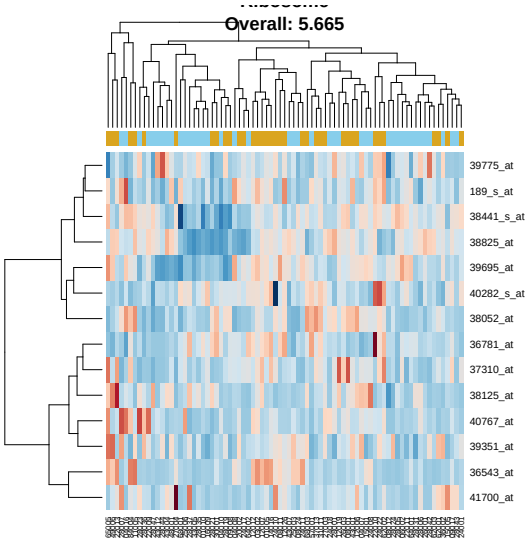

Figure 3: The mean plot for the category with the smallest p-value.

Figure 7 suggests that the level of activity among genes involved in the Ribosome pathway seems larger in those labeled NEG than the samples from patients with BCR/ABL. Perhaps indicating that ribosomal activity is suppressed in those with BCR/ABL.

We can further investigate the relationship between gene expression and the Ribosome pathway by examining heatmaps for this pathway. First, we examine the heatmap for those genes that were selected by our filtering approach, Figure **??**, and then we examine the heatmap for all genes that are on the chip and annotated at the ribosome pathway, Figure **??**.

Here the patterns of expression do not seem to corroborate the finding. Yes there are differences in the patient samples with respect to the expression of mRNA for ribosomal mRNAs, but the differences are not so striking. So what is going on? Well, one of the genes gives us a hint - if you examine the heatmap carefully you will notice that one probe seems to be expressed in one set of samples and not in another - and the probe is `41214_at`, which corresponds to RPS4Y1, which is a sex-linked ribosomal protein. If we then redraw the heatmap but color the sidebar according to sex, (green for males and slate for females) we see that these probes almost exactly separate the males and females.

So we have found something that appears to be real, at least biological, but is also uninteresting. Gender is slightly confounded with the relationship between BCR/ABL and NEG cytogenetics and hence we will sometimes find effects that may be more correctly attributed to sex.

0.3 A permutation distribution

It is important to assess the significance of the observed test statistics with respect to a reference distribution. To that end, we consider permuting the sample labels (that is which of the two groups, `BCR/ABL` or `NEG`, a patient belongs to). In the code below we consider 500 permutations.

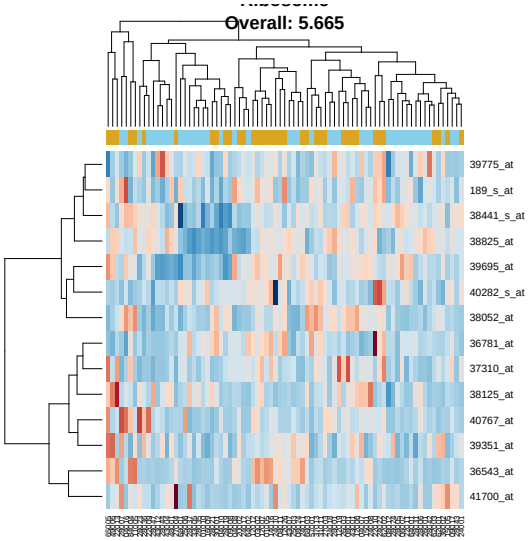

Figure 4: The mean plot for the category with the smallest p-value.

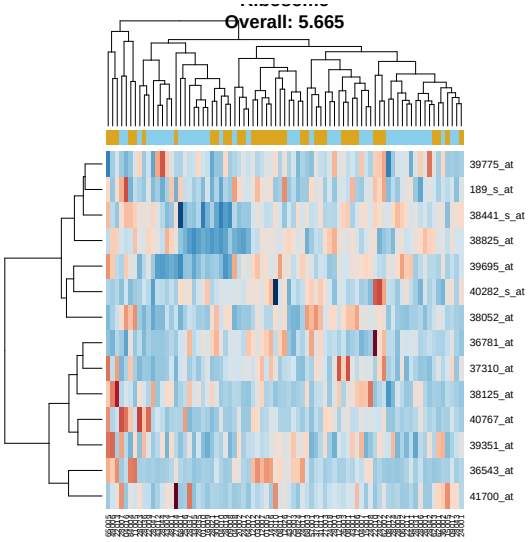

Figure 5: A heatmap of the selected probes for mRNA in the Ribosome pathway. The color bar is green for males and grey for females.

```
> v1 = ttperm(exprs(eS), eS$mol.biol, B = NPERM)
> permDm = do.call("cbind", lapply(v1$perms, function(x) x$statistic))
> permD = AmER2 %*% permDm
> permD2 = sweep(permD, 1, sqrt(rs2), "/")
> pvals = matrix(NA, nr = nCats, ncol = 2)
> dimnames(pvals) = list(row.names(AmER2), c("Lower", "Upper"))
> for (i in 1:nCats) {
+     pvals[i, 1] = sum(permD2[i, ] < tA[i])/NPERM
+     pvals[i, 2] = sum(permD2[i, ] > tA[i])/NPERM
+ }
> ord1 = order(pvals[, 1])
> lowC = (row.names(pvals)[ord1])[pvals[ord1, 1] < 0.05]
> highC = row.names(pvals)[pvals[, 2] < 0.05]
> getPathNames(lowC)

$"03010"
[1] "Ribosome"

$"00100"
[1] "Biosynthesis of steroids"

$"00220"
[1] "Urea cycle and metabolism of amino groups"

$"03022"
[1] "Basal transcription factors"

$"00062"
[1] "Fatty acid biosynthesis (path 2)"

> getPathNames(highC)

$"04210"
[1] "Apoptosis"

$"00561"
[1] "Glycerolipid metabolism"

$"00903"
[1] "Limonene and pinene degradation"

$"00563"
[1] "Glycosylphosphatidylinositol(GPI)-anchor biosynthesis"

$"04630"
[1] "Jak-STAT signaling pathway"
```

```
$"00410"
[1] "beta-Alanine metabolism"

$"00910"
[1] "Nitrogen metabolism"

$"04810"
[1] "Regulation of actin cytoskeleton"

$"04060"
[1] "Cytokine-cytokine receptor interaction"

$"05040"
[1] "Huntington's disease"

$"00340"
[1] "Histidine metabolism"

$"00513"
[1] "High-mannose type N-glycan biosynthesis"

$"04070"
[1] "Phosphatidylinositol signaling system"

$"05050"
[1] "Dentatorubropallidoluysian atrophy (DRPLA)"

$"00590"
[1] "Prostaglandin and leukotriene metabolism"

$"00760"
[1] "Nicotinate and nicotinamide metabolism"

$"00604"
[1] "Ganglioside biosynthesis"

$"04080"
[1] "Neuroactive ligand-receptor interaction"

$"04010"
[1] "MAPK signaling pathway"

$"04510"
[1] "Focal adhesion"
```

```
$"00770"
[1] "Pantothenate and CoA biosynthesis"

$"04512"
[1] "ECM-receptor interaction"

$"00531"
[1] "Glycosaminoglycan degradation"

$"04020"
[1] "Calcium signaling pathway"

$"04520"
[1] "Adherens junction"

$"04350"
[1] "TGF-beta signaling pathway"

$"04610"
[1] "Complement and coagulation cascades"

$"04530"
[1] "Tight junction"

$"04620"
[1] "Toll-like receptor signaling pathway"
```

```
> lnhC = length(highC)
```

Notice that we have used quite a large p-value, although our adjustment should be for the 94 categories that we are testing, and so is will not be too dramatic.

We can visualize the differences in group means, just as we did before. These are shown in Figures 6 through 9.

Unfortunately, as we can see from the visualizations none of these category plots are especially compelling. If we return to the category Ribosome then we see that the permutation p-value is 0.994.

1 Using other functions

In the examples above we used, perhaps the simplest form of per category statistic, the summation. Extending the model to deal with virtually any other per-category function is quite simple and the cade to do this is available as the `applyByCategory` function.

```
> byCmeds = applyByCategory(tobs$statistic, AmER2, FUN = median)
> byCrankTest = applyByCategory(tobs$statistic, AmER2, FUN = wilcox.test)
```

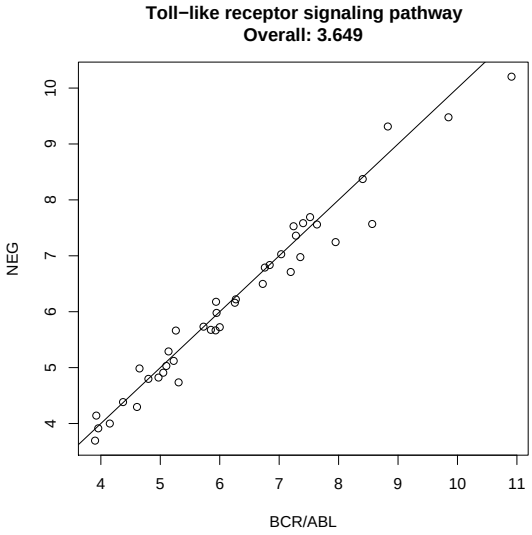

Figure 6: The per category mnplot for pathways that are deemed to be differentially expressed using a permutation approach.

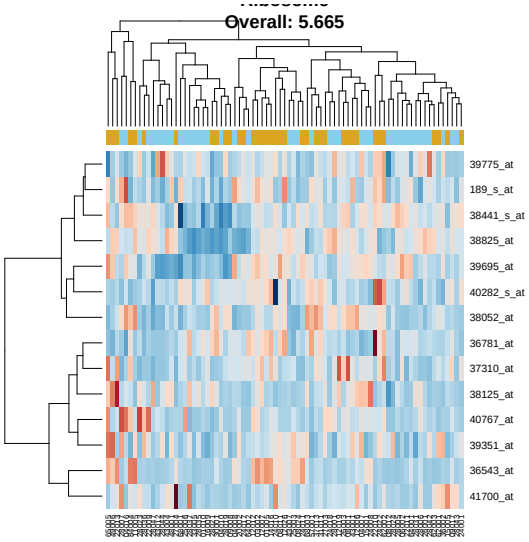

Figure 7: The mean plot for the category with the smallest p-value.

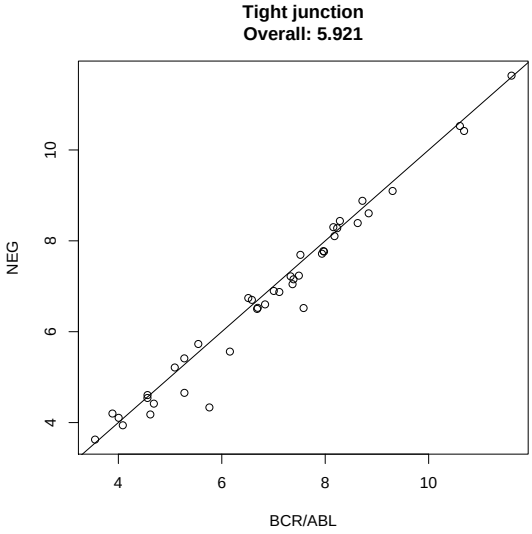

Figure 8: The per category mnplot for pathways that are deemed to be differentially expressed using a permutation approach.

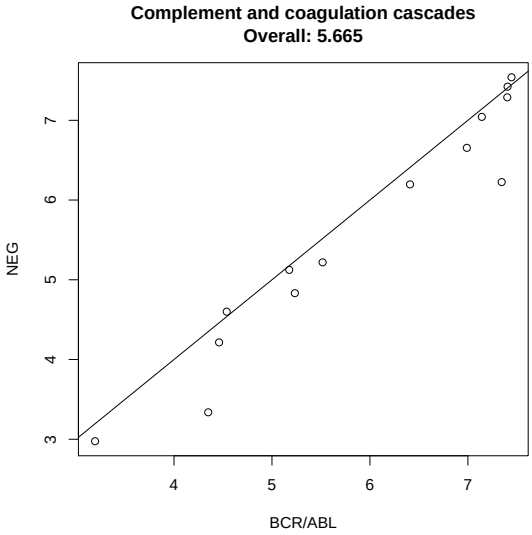

Figure 9: The per category mnplot pathway that are deemed to be differentially expressed using a permutation approach.

```
> ranS = sapply(byCrankTest, function(x) x$statistic)
> ranpv = sapply(byCrankTest, function(x) x$p.value)
```