14

Hypergeometric Testing Used for Gene Set Enrichment Analysis

S. Falcon and R. Gentleman

Abstract

After the set of interesting genes has been determined, say those that are differentially expressed, a next step in the analysis is to attempt to find functional relationships among those genes that might help better elucidate the underlying biology. These methods typically rely on existing or predefined sets of genes. In this chapter we show how to carry out Hypergeometric tests to identify potentially interesting gene sets.

14.1 Introduction

The **Category** and **GOstats** packages provide extensive facilities for the testing of over- and underrepresentation of gene sets in a specified list of interesting genes. In this chapter we focus most of our examples on the gene set collection induced by the Gene Ontology GO; (The Gene Ontology Consortium, 2000). However, the techniques demonstrated can be easily translated for use with other gene set collections supported by the **Category** package including KEGG, PFAM, and chromosome band annotation and these are covered in the exercises.

In this chapter we describe the preprocessing required to construct inputs for the main testing function, `hyperGTest`, the algorithms used, and the structure of the return value. We use a microarray data set (Chiaretti et al., 2004, 2005) from a clinical trial in acute lymphoblastic leukemia (ALL) to work an example analysis. In the ALL data, we focus on the patients with B-cell derived ALL, and in particular on comparing the group with BCR/ABL translocations to those with no observed cytogenetic abnormalities (NEG).

F. Hahne et al., *Bioconductor Case Studies*, DOI: 10.1007/978-0-387-77240-0_14,

To get started, load the packages needed for this analysis:

```
> library("Biobase")
> library("ALL")
> library("hgu95av2.db")
> library("annotate")
> library("genefilter")
> library("GOstats")
> library("RColorBrewer")
> library("xtable")
> library("Rgraphviz")
```

14.2 The basic problem

The reasoning behind this approach is that there is some universe of objects (in our case genes) that are of potential interest, and that these objects can be divided into two groups (those that are interesting and those that are not). In addition, there are other characteristics of the objects that are also binary, such as belonging to a particular GO category, or having a particular biological property. And hence one would like to ask whether there is an association between being interesting and having the particular property. This question is easily answered using basic probability, and the resulting test is also widely known as Fisher's exact test.

The probability calculation can be carried out in two different ways, but the resulting test statistics and p-values are identical. Consider an urn containing one ball for each gene in the universe and imagine that those that are interesting are colored black, and those that are not interesting are colored white. Then, under the null hypothesis that there is no relationship between being interesting and being in a given GO category containing K genes, we can model the number of interesting genes using a Hypergeometric distribution. If there are j interesting genes in the GO category, we simply compute the probability of seeing j or more black balls in K draws, without replacement, from the urn. This probability is symmetric, in the sense that we could also have described the problem with the balls colored according to the GO category, and select, without replacement, one ball for each gene in our gene list. Another way of thinking about this is to draw the balls from the urn that represent genes annotated at the given term and fill out a two-way table as shown in Table 14.1.

In principle there is no reason why either grouping needs to be binary. You could have three types of genes (really interesting, sort of interesting, and not interesting) and a category that has three levels. If so, the multivariate generalizations of the Hypergeometric distribution will be needed.