Annotating molecular profiles using Gene Ontology and Functional Enrichment

Greg Tucker-Kellogg

October 28, 2015

1 Background

A common problem in analysis of genomic studies is to *annotate* molecular profiling data. We would like to infer the functions (or processes) that are involved in a phenotype, either from an experiment or an observational study.

Example experiments include response of cells to stimulus, response of patients to treatment, differentiation of cells in development, etc.

Example observational studies include non-diseased versus diseased individuals, differences between strains of organisms, etc.

Molecular profiling data might be measurements of transcript levels, protein levels, metabolite levels, or other phenotypic data at a molecular level. Sequencing technologies, microarrays and other tools allow simultaneous measurement of thousands of transcripts, or proteins, or metabolites. For simplicity, let's consider gene products, such as transcripts or proteins, and let's refer to them as "genes" for short.

2 How to annotate genes from a study

Suppose we have performed an experiment using molecular profiling of a large number m of genes and also performed a statistical analysis of the data for each gene G_i, where i is in $\{1, 2, .., m\}$. As a result we have a set $\boldsymbol{A}$ of test results, where A_i is the result for our assessment of G_i. The values of A_i might be p values, t statistics, F statistics, or some combination of effect size and significance. For our purposes in this module, it doesn't matter; we can assume the test results are provided to us by a statistician. For the rest of this discussion let's assume we have used a cutoff to select n genes of interest, where $n < m$.

Our task as bioinformaticists may be to determine if any interesting *function* or *process* is associated with the n genes that are statistically interesting. That's what enrichment analysis does.

3 The Gene Ontology gives us sets of genes

The Gene Ontology, as we have seen, organises biological function into a collection of hierarchies. When genes are associated with categories of function, as they are in the Gene Ontology, we can use them as *sets* of genes. There are many ways to get such sets. All of the pathway databases can be used to generate sets, so the Gene Ontology is not the only way to annotate genes with function. However, the Gene Ontology is a well organised, collaborative effort to annotate gene function in a consistent manner, which why it's so widely used. The Gene Ontology gives us a shared vocabulary of function we can use.

4 The wrong way to assess function

There a lot of wrong ways to assess whether a biological function is associated with a molecular profiling experiment. An intuitive — but wrong — way to do it is to *count* the number of significant genes n_t associated with a function t (e.g., a Gene Ontology category) and then conclude that the list with the highest number, or the highest percentage, is the most important.

Why is that wrong? Because random chance may give us a high number, or a high percentage. If the Gene Ontology category is very broad, then many genes may belong to it by chance. Let's call that total number of our m genes annotated with function $t = m_t$. If we do an experiment using a microarray, and 25% of the genes on the microarray are annotated as "transport" (i.e. $m_t/m = 0.25$), then seeing 25% of your statistically significant genes annotated as "transport" is meaningless, even if that's the highest percentage.

5 The Fisher's exact test method

The correct way is to look for whether a category of function is over-represented among the significant genes in the profile. We use a *null model* of random chance: what is the chance at random of seeing $n_t = k$ genes annotated with function t from among the n genes that pass our cutoff? This comes from the hypergeometric distribution, and is exactly:

$$\Pr(n_t = k) = \frac{\binom{m_t}{k}\binom{m-m_t}{n-k}}{\binom{m}{n}} \tag{1}$$

But Eq. 1 is not quite right, since we want to know the chance of seeing *at least* than many genes in the set annotated with the function. So we sum over all the possibilities, which is the one-sided Fisher's Exact Test.

$$\Pr(n_t \geq k) = \sum_{k=n_t}^{\min(m_t,n)} \frac{\binom{m_t}{k}\binom{m-m_t}{n-k}}{\binom{m}{n}} \tag{2}$$

As you might expect, this is not easy to calculate by hand.

6 The GSEA approach

Another approach, called "Gene Set Enrichment Analysis", or GSEA, is also common. I mention it because you are likely to see it mentioned if you read the genomics litarature. Unlike the Fisher's Exact Test approach, the GSEA does not first identify a subset of statistically interesting genes. Instead, it takes *all* the profile data, and keeps it in rank order (from most statistically interesting to least). The null model is that genes annotated with a function are randomly distributed across the rank order sorted data set. GSEA looks for deviations from that null model, typically enrichment of genes with high rank.

7 What is a 'set'?

In both cases, the "set" that is being enriched is the set annotated with a function, not the set of genes from the experiment. In the case of the Fisher's Exact Test, it is a test of whether that (functional) set is enriched in the subset of genes that we have decided are of interest. In the case of GSEA, it is whether that set is enriched among genes ranked highly by our test.