CountBio

Mathematical tools for natural sciences

Biostatistics with R

Home Tutorial index Next Previous

Gene Enrichment Analysis

Generally, a set of genes expression under certain conditions are short listed from high throughput experiments like microarrays, RNAseq etc.

Most of the cellular processes involve many genes acting together.

How these two things, ie., individual gene selection based on expression and the collective processes compared?. How many of these genes individually expressed participate contribute to the collective process?

Gene Enrichment Analysis aims to derive potentially interesting biological themes from individual gene-based measurements.

Published first in a classif paper,

"*Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*"
Aravind Subramanian, Tablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander and Jill P. Mesirov a,k
PNAS, October 25, 2005, vol. 102, no. 43, 15545–15550

The limitations of studying individual genes from a control-treatment type of expression analysis are as follows:

1. After correcting for multiple testing hypothesis, not many individual genes might cross the threshold set for statistical significance, especially for the biological phenomena that involve low levels of gene expression.

2. On the other extreme, we may end up with a very large list of differentially expressed genes, making the interpretation of the disease or the phnomena difficult.

3. Study of single genes will miss out the collective action of pathway. For example, an increase of 30 expression levels in particular 8 genes in a pathway may be more effective than factor of 6 increase in just two of them!. Expression analysis will detect these two genes alone, leaving out the others with low expression level.

4. It has been observed in expression analysis that two experiments studying same disease report different sets of differentially expressed of gene list!

In order to overcome these issues, "Gene Set Enrichment Analysis" was developed by the above group. This became the basis for the innumerable number of tools like DAVID, and the method is improved over last 15 years.

The Gene Set Enrichment Analysis take two sets of genes at the input:

1. The set of genes (L) which are ranked from top to bottom according to the level of differential expression. For example, take all the gene expression values of a microarray experiment and rank in descending order of differential expression level. (No threshold filtering).

2. Get another set S of genes based on the prior knwoledge of the same phenomena or disease from previous experiments, or gene products in a *biochemical pathway (eg. metabolic pathway)* or *coexpression networks* or *sharing same GO terms* etc.

Note that the gene sets L is from expression studies and S is from independent information.

What the Gene Set Enrichment Analysis does?. To use the author's words in the paper,

1. "The goal of GSEA is to determine whether members of a gene set S tend to occur toward the top (or bottom) of the list L, in which case the gene set is correlated with the phenotypic class distinction."

In other words,

2. "Given an a priori defined set of genes S (e.g., genes encoding products in a metabolic pathway, located in the same cytogenetic band, or sharing the same GO category), **the goal of GSEA is to determine whether the members of S are randomly distributed throughout L or primarily found at the top or bottom. We expect that sets related to the phenotypic distinction will tend to show the latter distribution."**

Before going into details of various steps involved in this analysis, we will require to understand a funcdamental idea behind this. We can start with Fisher's exact test for such calculations. We will understand other methods from this.

Fisher's exact test for contingency tables

In an experiment, suppose we have measured differentially expressed genes by treating certain cells with a drug.

We also have a GO annotation term which we want to study in this condition.

Is there any association between these differentially expressed genes and annotations for the GO term considered?

Put it in other words, "Is the GO term enriched with the set of genes from the expression analysis?"

Suppose we take the particular GO term biological process. We prepare a contingency table by counting the genes in experiment that are there or not there in the GO term list:

	Differential Expression	No differential Expression	Sum
Present in GO term	a=12	b=3	a+b=15
No present in GO term	c=3	d=12	c+d=15
Sum	a+c=15	b+d=15	N=a+b+c+d=30

From above table, we see that we have observed 12 differentially observed genes from experiment out of 15 in the GO term set. What is the statistical significance of this?.

If we repeat the experiment, we may end up in 10 genes instead of 12. Is this also significant?

We do Fisher's exact test to get a significant level for this table. With this, we get the hypergeometric distribution given by,

$$P_h(x, n, N_1, N) = \frac{{}_{N_1}C_x \times {}_{N_2}C_{n-x}}{{}_{N_1+N_2}C_n} = \frac{{}_{N_1}C_x \times {}_{N-N_1}C_{n-x}}{{}_{N}C_n}$$

For the contingency table, we have,

With this, we get the probability density of hypergeometric distribution given by,

$$P_h(x = a, n = a + c, N1 = a + b, N = a + b + c + d) = \frac{{}_{a+b}C_a \times {}_{c+d}C_c}{{}_{N}C_{a+c}} = \frac{{}_{a+b}C_b \times {}_{c+d}C_d}{{}_{N}C_{b+d}}$$

The above expression give a **p-value** for the observed contingency table under the **null hypothesis that the presentce or absence of genes in GO terms are independent of their expression levels** .

Now, for a given row and column sums, there are many possible tables. We can compute the probability for each one of the possible tables. From this, we get the probability of getting a value more than or equal to the observed 'a'.

Suppose, we take a contingency table with a+b=15 and c+d=15

Pssoble value of a,b,c,d that can satify this sums and corrreponding probabilities are:

$a = 0 \quad b = 15$

$c = 15 \quad d = 0 \qquad p = 6.45 \times 10^{-9}$

$a = 1 \quad b = 14$

$c = 14 \quad d = 1 \qquad p = 1.4 \times 10^{-6}$

$a = 2 \quad b = 13$

$c = 13 \quad d = 2 \qquad p = 7.1 \times 10^{-5}$

$a = 3 \quad b = 12$

$c = 12 \quad d = 3 \qquad p = 1.3 \times 10^{-3}$

$a = 4 \quad b = 11$

$c = 11 \quad d = 4 \qquad p = 1.2 \times 10^{-2}$

$a = 5 \quad b = 10$

$c = 10 \quad d = 5 \qquad p = 5.8 \times 10^{-2}$

$a = 6 \quad b = 9$

$c = 9 \quad d = 6 \qquad p = 1.6 \times 10^{-2}$

$a = 7 \quad b = 8$

$$c = 8 \quad d = 7 \qquad p = 2.6 \times 10^{-2}$$

$$a = 8 \quad b = 7$$

$$c = 7 \quad d = 8 \qquad p = 2.6 \times 10^{-2}$$

$$a = 9 \quad b = 6$$

$$c = 6 \quad d = 9 \qquad p = 1.6 \times 10^{-2}$$

$$a = 10 \quad b = 5$$

$$c = 5 \quad d = 10 \qquad p = 5.8 \times 10^{-2}$$

$$a = 11 \quad b = 4$$

$$c = 4 \quad d = 11 \qquad p = 1.2 \times 10^{-2}$$

$$a = 12 \quad b = 3$$

$$c = 3 \quad d = 12 \qquad p = 1.3 \times 10^{-3}$$

$$a = 13 \quad b = 2$$

$$c = 2 \quad d = 13 \qquad p = 7.1 \times 10^{-5}$$

$$a = 14 \quad b = 1$$

$$c = 1 \quad d = 14 \qquad p = 1.4 \times 10^{-6}$$

$$a = 15 \quad b = 0$$

$$c = 0 \quad d = 15 \qquad p = 6.4 \times 10^{-9}$$

From above results we compute the probability of getting 12 or more differentially expressed genes that are also in the gene set with GO terms as,

$$P \; = \; \sum_i P_i \; = \; 1.3 \times 10^{-3} + 7.1 \times 10^{-5} + 1.4 \times 10^{-6} + 6.4 \times 10^{-9} \; = \; 0.0014$$

From this result, we claim that the probability of our observed data or that more extreme under the assumption that there is no association between expression and gene set membership is 0.0014