Statistical Testing with Genes

Saurabh Sinha
CS 466
Hypergeometric test

- A population of $N$ genes
- Microarray data identified a subset of $n$ genes as being up-regulated in cancer
- Also know a set of $m$ genes involved in cell division
- Want to test for association between cancer and cell-division
Hypergeometric test

Is the intersection (size “k”) significant large, given N, m, n?
Hypergeometric Distribution

Given that \( m \) of \( N \) genes are labeled “cell division”. If we picked a random sample of \( n \) genes, how likely is an intersection equal to \( k \) ?

\[
f(k; N, m, n) = \frac{(m)_k (N-m)_{n-k}}{N_n}.
\]
Hypergeometric Distribution

Given that $m$ of $N$ genes are labeled “cell division”. If we picked a random sample of $n$ genes, how likely is an intersection equal to or greater than $k$?

$$P = \sum_{j \geq k} f(j; N, m, n)$$
“Enrichment” analysis

• If \( P \) calculated this way is below some threshold \( \alpha \), e.g., 0.05, we say that the association between the cancer set and the cell division set is statistically significant.

• In general, such “enrichment analysis”
  – has a set of genes with known function (or other property, e.g., “fast evolving”)
  – has a set of genes identified by the particular study (e.g., microarray study of cancer)
  – tests if the identified genes are enriched for the function or property
Gene Ontology

• Gene Ontology (GO) is a highly popular source of “known gene sets”, e.g., gene sets with common known function

Source: “Bionformatics” - Polanski and Kimmel
Motivation for GO

• Large numbers of genes in different organisms, with large numbers of protein products and functions
• Several databases (often species specific) store such information (e.g., FlyBase)
• Need a way to organize all this information from all these databases
Consistent and Structured organization

• Browsing through genome databases like FlyBase is powerful, but …
• … lack of consistency from one database to another. (The “interface” is not the same.)
• Annotations (e.g., functions of genes) could be more “structured”, for much easier and automated access
  – e.g., GoogleScholar/Pubmed versus Bibliography software/databases
Gene Ontology database

- In response to these needs.
- Aims at structured and consistent terms to describe gene function
- A “standardized” and “structured” vocabulary to describe genes and their products
Structure of GO

• Tree structure, with three main branches:
  – “molecular function”: activities of gene product at molecular level, e.g., “DNA binding”
  – “biological process”: process (series of events), e.g., “metabolism” or “oxygen transport”
  – “cellular component”: e.g., “cell nucleus”

• Every GO term is a descendent of one of these three branches

• Tree structure captures natural hierarchy of terms, e.g., “metabolism” is an ancestor node of “amino acid metabolism”
GO database

• GO is not only a hierarchy of descriptive terms, it is also the assignment of one or more of these terms to genes
• Large numbers of genes in different organisms have been manually annotated with GO terms
Accessing GO

- Can download entire hierarchy of terms, as well as assignment of these terms to genes in a species.
- Can access through specialized interfaces, e.g.,
  - GeneMerge for enrichment analysis of sets of genes
  - AmiGO: Search by gene name (all terms for that gene) or by term name (all genes for that term)
  
http://www.genedb.org/amigo/perl/go.cgi
Statistical Tests

• Enrichment analysis is a statistical test:
• We calculate the random chance of seeing something “as good” (e.g., intersection as large) as what is observed
• If this is very small, we “reject the null hypothesis” that the observation happened “by chance”
Statistical Testing with genes: another example

• Consider a single gene $g$. Take its expression values in experimental condition 1, and in experimental condition 2.

• Compare the two samples, e.g., calculate the difference of means.

• Null hypothesis: both samples come from populations with equal means.

• Is the difference of means significantly different from zero?

• A statistical test about gene $g$
Multiple genes

• Null hypothesis about equality of means is “gene-wise” null hypothesis

• How do we extend this framework to test a set of genes (for difference of expression between the two conditions)?

• One natural extension: Declare equality in both conditions for every gene. This is the complete null hypothesis.

Source: Stat. Methods in Bioinformatics; Ewens & Grant.
Complete Null Hypothesis

• But this may not be the hypothesis we are interested in accepting or rejecting: rejecting it doesn’t tell us which (subset of) genes are differentially expressed

• Rejecting the complete null hypothesis only tells us “some gene is not equally expressed in the two conditions”.
What we’d like

• A procedure that predicts which genes are differentially expressed
• Remember that “prediction” in this context is not a definitive prediction, there is some margin of error
• We’d like our procedure to provide us with an overall margin of error
To be more specific

• Let’s go back to gene-wise null hypothesis $H_0$
• $\Pr (X \geq k \mid H_0) < \alpha$
• If we observe random variable $X$ to have a value of $k$ or more, we reject $H_0$, i.e., we predict that $H_0$ is not true.
• However, this prediction has a margin of error: $\alpha$
• It is possible that $H_0$ is true, yet we rejected it; in fact the probability of this “false positive” error is $\alpha$!
• So we are able to control “false positive” rate in the single gene test.

• “Positive” because rejecting null hypothesis usually incriminates the gene as being “interesting” in some way. “False” because $H_0$ being true means that the prediction is false.
Controlling false positives

• In the multiple gene test procedure (not defined as yet), we would like to predict a set of genes as being interesting, i.e., as violating null hypothesis.

• But we would like to have some control over (i.e., some idea of) the false positive rate
False Discovery Rate

- Is one such procedure
- The final outcome will be a set of genes predicted to be differentially expressed
- We will have some control on the *proportion* of false positives among these predicted genes
- Say there are 10,000 genes and 100 are true differentially expressed. It be OK to predict some set of 100 genes, with the disclaimer that 50 of these may be false positives.
  - False Discovery Rate of 50% may be OK!
Some definitions

• Consider the genes for which null hypothesis is rejected. (Say $R$ in number.)
• Let $V$ be # genes (from these $R$) for which null hypothesis is true (i.e., falsely rejected, or false positive)
• Let $S$ be the # genes (from these $R$) for which null hypothesis is false (i.e., correctly rejected)
• Define $Q = 0$ if $R = 0$
• $Q = \frac{V}{R}$ if $R > 0$
Some definitions

• V, S, R, and Q are random variables, and even after all tests are done we don’t know Q (the false positive proportion)
• Define FDR = Expectation(Q)
• The FDR “procedure” will aim to control this expected value of Q
A difference between p-values and FDRs

• FDR is fundamentally different from a p-value.

• P-value assesses significance of data. If we publish some data that we claim to be significant, we should present a small p-value for the data (e.g., <= 0.05)

• FDR is generally used as a “culling tool”; the investigator wants to predict a set of genes to test experimentally, and an FDR of 0.5 may be acceptable to her (she’ll do twice as much experimental work, which may be fine)
The “Procedure”

• Proposed by Benjamini and Hochberg in 1995. Many other procedures since then, but we’ll only see this original one.

• Begin with a per-gene p-value, i.e., \( \Pr(X \geq k \mid H_0) \), for every one of the \( g \) genes being studied.
  
  – \( g \) gene-wise null hypotheses, with their corresponding p-values computed.
Benjamini-Hochberg procedure

1. Let the g gene-wise p-values be denoted by \( p_{(i)} \).
2. Consider these g p-values to be sorted in ascending order, i.e., \( p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(g)} \).
3. Let \( H_{(i)} \) be the hypothesis corresponding to \( p_{(i)} \).
4. Let \( q_i = \frac{i}{g} \alpha \) for \( i = 1, 2, 3 \ldots g \) and \( \alpha \) is the desired FDR.
5. Let k be the max i such that \( p_{(i)} \leq q_i \).
6. Procedure: Reject null hypothesis \( H_{(1)}, H_{(2)}, \ldots H_{(k)} \) and accept all others.
7. Theorem: \( E(Q) = \frac{g_0}{g} \alpha \leq \alpha \).