Introduction to Gene Set Testing and the Global Test

Utah State University – Spring 2014
STAT 5570: Statistical Bioinformatics
Notes 4.3

References

- Goeman et al. Bioinformatics 20(1):93-99 (2004)
- Goeman et al. JRSS-B 68(3) 477-493 (2006)
- Goeman & Buhlmann. Bioinformatics 23(8):980-987 (2007)
- Chapters 19-22 of Bioconductor Monograph (course text)

Motivation for “over-representation”

- Suppose we perform a test of DE and find a list of 132 significant genes (out of 2,000)
- Consider a specific GO term, like apoptosis

	apop.	not apop.	
significant	100	32	132
not signif.	753	1,115	1,868
	853	1,147	2,000

Questions in Gene Set Tests

- Is “significance” independent of “apoptosis”?
- Is the gene set “apoptosis” over-represented among “significant” genes?
- Is the gene set “apoptosis” differentially expressed?

Traditional test of "independence"

- EV = (row total) x (column total) / (table total)
- Test statistic:

$$\chi^2 = \sum \frac{(\mathrm{OV}-\mathrm{EV})^2}{\mathrm{EV}}$$

- If truth is "independence" and sample size is large, then $\chi^2 \sim \chi^2_{df}$ $\quad df = (\#\,\mathrm{cols.}-1)(\#\,\mathrm{rows}-1)$
- Or – obtain P-value by permutations (exact test, based on hypergeometric dist'n)

Hypergeometric distn. & Fisher's exact test

- Suppose an urn contains *n* balls:
  r black and *n-r* white
- Draw *m* balls without replacement
- Let *X* = # black balls drawn;
  then X ~ hypergeometric(*r*,*n*,*m*):

$$P(X=k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}$$

- Use this to generate probabilities of all possible 2x2 tables with same row and column totals; look at probability of first cell count
- P-value = proportion of tables more extreme than the original table (usually based on χ^2 values)

(implemented in hyperGTest function of package GOstats)

But wait – possible null hypotheses

- Competitive null

  compare DE of gene set (G) to a standard defined by the set's complement (G^c)

$$H_0^{comp} = \text{The genes in G are at most as often DE as the genes in } \mathrm{G}^c$$

- Self-contained null

  compare gene set to a fixed standard (does not compare to set's complement)

$$H_0^{self} = \text{No genes in G are DE}$$

Comparing null hypotheses

- Competitive Null:
  - relinquishes power in order to make a stronger statement
  - singleton gene sets treated differently from single gene tests
  - cannot be used to test set of all genes
- Self-contained Null:
  - more statistical power due to restrictive nature (will tend to reject null more often)
  - will reject null when gene set is a singleton
  - can be used to test set of all genes – useful as precheck

Possible P-value calculation strategies

- Gene sampling

    genes are sampling units;
    permute gene labels – (col. & row sums constant)
        (gene set: yes/no; DE: sig/no)

- Subject sampling

    subjects (arrays) are sampling units;
    permute array labels (trt/ctrl)

Comparing resampling strategies

- Gene sampling
    - urn model (like hypergeometric)
        - have two measurements describing each gene: yes/no, sig/no
        - look at all possible re-arrangements of table values, keeping fixed marginal totals
    - reverses roles of samples and genes for testing
        - traditional: sample of subjects with fixed measurements
        - here: sample of measurements from fixed samples
    - assumes observations for genes are iid
    - sample size = number of genes

- Subject sampling
    - sample size = number of samples

Resampling strategies and P-values

- If we were to replicate the experiment many times (and null is true), P-value is expected % of replicated experiments yielding more extreme results than the actual biological experiment

- Gene resampling strategy
    - a replicate experiment would measure new genes on the same subjects
    - does not mimic actual biological experiment
    - strongly discouraged
- Subject resampling strategy
    - a replicate experiment would measure new subjects on the same genes

---

Hypergeometric test: what does "significance" mean?

- competitive null, gene sampling
- Null assumes genes in gene set are not unusually differentially expressed, and genes in the gene set are independent
- A "significant" P-value → reject Null
    - usually – treat this as concluding "genes in gene set are unusually DE"
    - but – could be due to dependence of genes in gene set (which is to be expected among functionally-related genes)

Other gene set testing methods

- GSEA (Gene Set Enrichment Analysis) [Excellent 6570 project!]
  - Mootha et al. Nature Genetics 34(3):267-273 (2003)
  - Subramanian et al. PNAS 102(43):15545-15550 (2005)
- Global Ancova
  - Mansmann and Meister. Methods of Information in Medicine 44(3):449-453 (2005)
- SAFE (Signicance Analysis of Function and Expression)
  - Barry et al. Bioinformatics 21(9):1943-1949 (2005)
- Global testing
  - Goeman et al. Bioinformatics 20(1):93-99 (2004)
- ADGO (Analysis of Differentially expressed gene sets using composite GO annotation)
  - Nam et al. Bioinformatics 22(18):2249-2253 (2006)
- GXNA (Gene eXpression Network Analysis)
  - Nacu et al. Bioinformatics 23(7):850-858 (2007)
- more … a very active area right now

Global test for groups of genes

- Do subjects with similar gene expression profiles have similar class labels?

- Look within single groups of genes
  - GO term
  - KEGG pathway
  - cluster from cluster analysis

Notation

- n samples (arrays), p genes
- m = subgroup size (# of genes in gene set)
- X = (x_{ij})
  = matrix of "normalized" expression values
  (n rows, p columns)
- Y = vector of "clinical outcome" (usually 0/1)

Motivation for global test

- Suppose a gene set can be used to predict the clinical outcome

- Then within the gene set, changes in gene expression patterns should match changes in clinical outcome

- Y depends on X – but how?

Statistical justification for global test

- Generalized linear model:

$$h\left(E[Y_i \mid \beta]\right) = \alpha + \sum_{j=1}^{m} x_{ij}\beta_j = \alpha + r_i$$

(could be logit function)

- Does gene set predict clinical outcome?

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0$$

- But – it could be that m >> n
  - then classical tests fail

Revising model as random effects

- Assume: $\beta_1, \beta_2, \ldots, \beta_m \sim F$ (generic dist'n)

$$E[\beta] = 0,\ \mathrm{Var}[\beta] = \tau^2$$

- Then: $H_0 : \tau^2 = 0$
- Equivalently: $r = \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix}$, $E[r] = 0,\ Cov[r] = \tau^2 XX^T$
- Interpretation: $\tau^2 > 0:$ similar $X_{i,\bullet}$ and $X_{\ell,\bullet}$

$$\Rightarrow \text{ similar } E[Y_i] \text{ and } E[Y_\ell]$$

Score test – from random effects

- Let $R = \frac{1}{m} XX^T$

$$\mu = h^{-1}(\alpha) = E\left[Y \mid \tau^2 = 0\right]$$

$$\mu_2 = E\left[(Y-\mu)^2 \mid \tau^2 = 0\right];\ \mu_4 = E\left[(Y-\mu)^4 \mid \tau^2 = 0\right]$$

$$Q = \frac{(Y-\mu)^T R (Y-\mu)}{\mu_2}$$

$$E[Q] = trace(R);\quad Var[Q] = 2trace(R^2) + \left(\tfrac{\mu_4}{\mu_2^2} - 3\right)\sum_i R_{ii}^2$$

$$c = Var[Q]/(2E[Q]);\quad \upsilon = 2(E[Q])^2/Var[Q]$$

- Under $H_0 : \tau^2 = 0,\ \ Q/c \mathrel{\dot\sim} \chi^2_\upsilon$

Global test – interpretation of significance

- Testing a set of genes

  Null: none of the genes in the gene set are correlated with clinical outcome

- “For a significant result it is not necessary that the genes in the [gene set] have similar expression patterns, only that many of them are correlated with the [clinical] outcome.”

```
library(affy); library(ALL); data(ALL); library(genefilter)

eset <- exprs(ALL)
T.cell <- c(rep(0,95),rep(1,33))

Eset <- new("ExpressionSet", exprs=eset)
pData(Eset) <- data.frame(trt=as.character(T.cell))
annotation(Eset) <- "hgu95av2"

library(globaltest)
gt.all <- gt(trt,Eset)
 # this tests whether B-cell and T-cell
 # patients have the same overall gene expression profiles
gt.all
```

```
    p-value Statistic Expected Std.dev   #Cov
1 2.1e-34       9.98    0.787   0.226 12625
```

Interpretation: Strong evidence that at least some genes are correlated with T-cell status

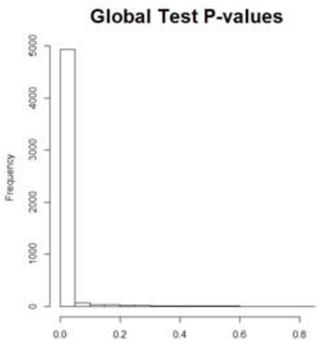
NOTE: In general, use non-filtered data for gene set tests

```
print(date())
gt.GO <- gtGO(trt, Eset, ontology="BP",
  minsize=10, maxsize=2000)
print(date()) # about 3.5 minutes

result <- data.frame(GO.ID=names(gt.GO),
     alias=gt.GO@extra[,2],
     pvalue=gt.GO@result[,1])
head(result)

hist(result$pvalue, xlab=NA,
   main='Global Test P-values',
   cex.main=2)

dim(result)  # 5133  3
```



```
     GO.ID                                  alias       pvalue
GO:0033077 T cell differentiation in thymus 3.710824e-73
GO:0045061            thymic T cell selection 1.367810e-72
GO:0001775                    cell activation 1.403571e-72
GO:0043368          positive T cell selection 1.825101e-71
GO:0045058                   T cell selection 2.686312e-71
GO:0046649              lymphocyte activation 1.215176e-70
```

NOTE: ontology, minsize, and maxsize options in gtGO function – why?

Global Test – more

- We focus here on logistic regression, but can also do linear regression, multinomial logistic regression, Poisson regression, and Cox proportional hazards model – depending on nature of phenotype variable
- Generalized linear models: Goeman et al. (2011) Biometrika 98(2):381-390 – possible 6570 project
- Multiple hypothesis testing – still a problem; but can adjust (even accounting for structure among GO terms – see package vignette)

Summary

- Testing Gene Sets
  - GO classes
  - KEGG pathways
  - (also Broad Institute pathway databases)
- Some methods based on 2-way table
  - best if based on subject sampling scheme (and self-contained null) – like global testing
- Coming up next: visualization of gene set testing results
- Later: meta-analysis approaches to gene set testing (allow more meaningful alternative)