

# Inferring Function From Known Genes

Naomi Altman

Nov. 06

# Objective

There are 3 major objectives for microarray studies:

- 1) Understand the function of genes.
- 2) Understand a biological process.
- 3) Classify samples.

3 is "easy" since it is purely observational.

1&2 start from genetic methods (i.e. 1 gene at a time)

# Genetic Methods

Tools:

- "knock-out" genotypes in which the gene is mutated to be non-functional
- "tagged" genotypes in which a fluorescent tag is activated when the gene is activated
- "in situ" hybridization in which the gene product is labeled in thin sections which can be viewed under a microscope
- "genetically modified" strains in which foreign genes have been added
- "genetically modified" strains in which selected genes are over- or under-expressed

# Using Known Genes

There are several ways in which known genes can be used to infer the function of unknown genes in a microarray experiment.

## 1) Seeded clustering

Assume that the nearest neighbors (according to some expression metric) of the known gene have similar function.

## 2) Unsupervised clustering

majority rule - in a cluster consisting of both known and unknown genes, assume that the cluster has the majority function

# Using Known Genes

There are several ways in which known genes can be used to infer the function of unknown genes in a microarray experiment.

## 3) Pathway analysis

If the genes are sufficiently well understood, they may be assembled into networks showing which genes regulate other genes.

Unknown genes that have expression patterns similar to those in the network can be placed in the network.

BioPixie (for yeast) will be demonstrated by 2 project groups.

PathAssist ?

# Understanding the Biological Process

We also use the known genes to infer the biological processes underlying our experimental treatments.

The primary tools are gene classification methods based on function and/or sequence.

The most used tool is probably the Gene Ontology Project.

# Gene Ontology Project

- [http://en.wikipedia.org/wiki/Gene\\_Ontology](http://en.wikipedia.org/wiki/Gene_Ontology)

- <http://www.geneontology.org/>

(try entering "DNA repair"

"ribosome"

"protease"

or the term of your choice

under gene or protein: try "ap2"

# How does GO "work"

The terms are in a tree-like structure but some nodes have multiple "parents".

The annotations are assigned by a team of collaborators and may also be submitted by other biologists (i.e. somewhat like Wikipedia, but with a more formalized central group)

Each annotation also has an "evidence" annotation, which can be used to assess reliability.

# Using GO with Microarray Data

- 1) Compile a list of differentially expressed genes.
- 2) Obtain the GO annotations of all genes on the array.
- 3) Extract the GO annotations of the DE genes.

Determine which GO annotations are over- or under-represented among the DE genes.

# Descriptive Use of GO

410

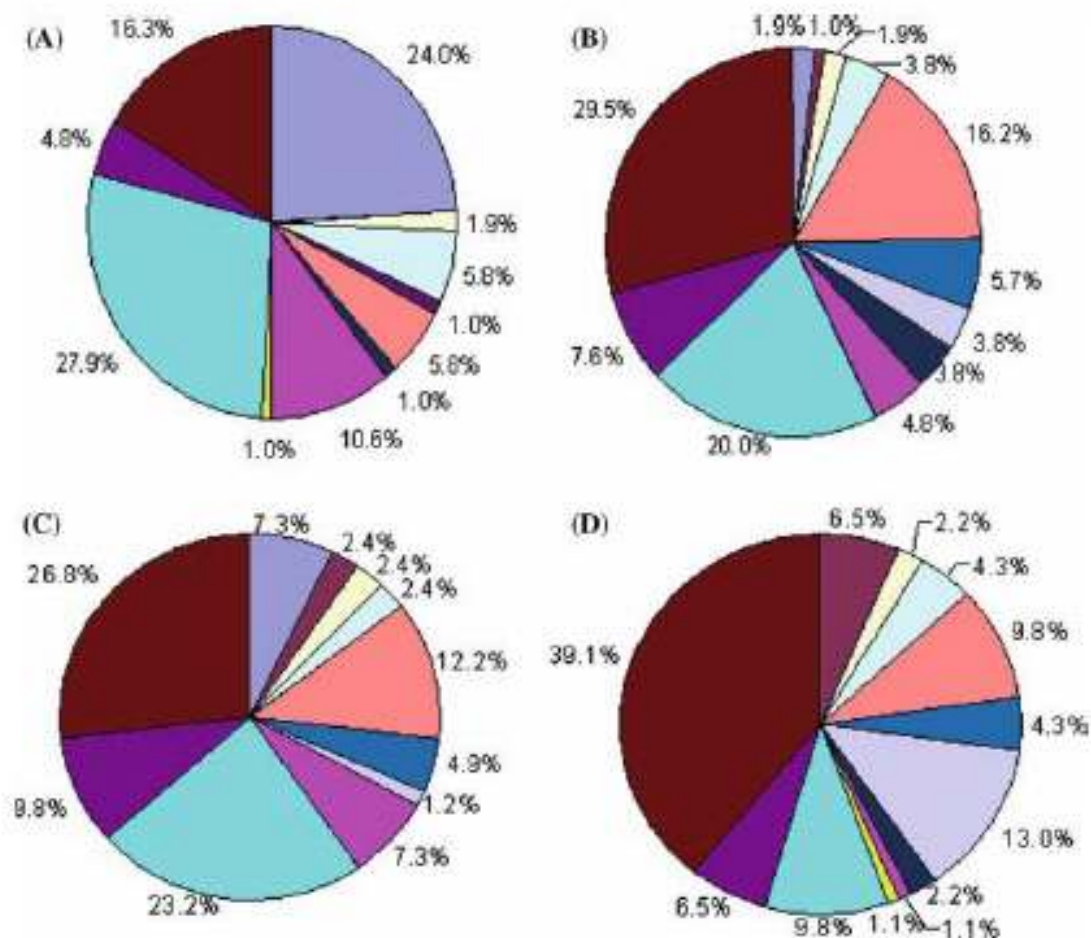


Figure 5. Distribution of genes in various biological process categories. (A) Genes preferentially expressed in young inflorescence. (B) Genes preferentially expressed in reproductive tissues. (C) Genes preferentially expressed in young inflorescence and stage-12 flower. (D) Genes preferentially expressed in stage-12 flower and silique. The categories were assigned based on Go Slim, and shown in clockwise order starting from 12 o'clock. ■ Cell growth/division; ■ cell structure; ■ Disease/defense; ■ Energy; ■ Intracellular traffic; ■ Metabolism; ■ other function; ■ Protein destination and storage; ■ Protein synthesis; ■ Signal transduction; ■ Stress/defense; ■ Transcription; ■ Transporter; ■ Unclassified.

# Testing For Enrichment

	In	Out	
DE	$N_{11}$	$N_{12}$	$N_{.1}$
not DE	$N_{21}$	$N_{22}$	$N_{.2}$
	$N_{1.}$	$N_{2.}$	$N_{..}$

$H_0$ : The percentage of DE genes In the GO category is proportional to the number on the array in the category

percentage on array:  $N_{1.}/N_{..}$

expected:  $N_{.1}N_{.1}/N_{..}$

observed:  $N_{11}$

test  $(O-E)^2/E$  (Chi-squared test)

# Testing For Enrichment

	In	Out	
DE	$N_{11}$	$N_{12}$	$N_{.1}$
not DE	$N_{21}$	$N_{22}$	$N_{.2}$
	$N_{1.}$	$N_{2.}$	$N_{..}$

## Problems:

- 1) Multiple testing: e.g. the Ontology for humans contains over 1000 terms
- 2) The ontology categories are nested.
- 3) The test statistic assumes that the genes are selected independently (but they are generally dependent).
- 4) Large  $N_{..}$  can lead to spuriously small p-values.

# GO in R

The GOstats package in R will:

take a genelist with Entrez IDs

take a reference genelist with Entrez IDs

take an annotation package (you can make your own)

Find all the "significantly enriched" or "significantly depleted" GO categories.

(The documentation was not very readable, but it was simple once I found an example.)

# Using GOstats with the Human vs Chimp Brain Data

The limma output was saved in `efit.contrast`. I selected all genes with highly significant differential expression among the treatments.

```
select=efit.contrast$F.p.value<.0000001  
genes=efit.contrast$genes[select,] #affy probe id  
library(hgu95av2)  
w1<-as.list(hgu95av2ENTREZID)  
#Entrez id for every probeset on the array
```

## Using GOstats with the Human vs Chimp Brain Data

```
entrez=w1[genes]
entrezID=unlist(entrez)
params <- new("GOHyperGParams", geneIds =
  entrezID, universeGeneIds = unlist(w1),
  annotation = "hgu95av2", ontology = "BP",
  pvalueCutoff = .001, conditional = FALSE,
  testDirection = "over")
HGOver=hyperGTest(params) #tests all nodes
htmlReport(HGOver,"h.html")#report on sig nodes
summary(HGOver) #report in R
```

# Results

Gene to GO BP test for over-representation

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0008065	0.00	-29.22	0	2	1.00	establishment of blood-nerve barrier
GO:0007399	0.00	2.69	21	48	325.00	nervous system development
GO:0048731	0.00	2.65	21	48	329.00	system development
GO:0019226	0.00	3.15	12	31	181.00	transmission of nerve impulse
GO:0007417	0.00	4.36	6	20	89.00	central nervous system development
GO:0007420	0.00	8.60	2	11	30.00	brain development
GO:0007268	0.00	2.97	11	29	177.00	synaptic transmission
GO:0050877	0.00	2.22	25	48	381.00	neurophysiological process
GO:0007154	0.00	1.49	133	171	2066.00	cell communication
GO:0030072	0.00	58.68	0	4	5.00	peptide hormone secretion
GO:0030073	0.00	58.68	0	4	5.00	insulin secretion
GO:0007215	0.00	11.04	1	6	14.00	glutamate signaling pathway
GO:0007267	0.00	1.80	27	45	425.00	cell-cell signaling
GO:0006813	0.00	3.16	5	13	74.00	potassium ion transport
GO:0009187	0.00	5.42	2	7	26.00	cyclic nucleotide metabolism

$N_{11}$

$N_{21}$