

Statistical analysis of microarray data

Karen Kafadar
Department of Mathematics
University of Colorado-Denver

- Introduction
- cDNA slides and Oligonucleotide chips
- Data quality: cDNA slides
 - manufacturing
 - measurement error
 - background signals
- Inference: Affymetrix algorithms
- Inference: cDNA analysis
 - transformations
 - background
- Inference: Multiplicity
 - FDR
 - Effect of correlated tests
 - Clustering by correlations
- Final comments

Introduction: Two types of gene chips

cDNA gene experiments (courtesy of Kim Kafadar)

- DNA (genetic code) in cell nucleus (A,G,C,T)
- Genes = organized strings of nucleotides
- Proteins (sets of triplets) made in cell cytoplasm
- Need to get the genes into the cytoplasm
- Cell doesn't pass all of DNA into cytoplasm -- first makes a copy; then splices out introns (mRNA)
- Ribosomes then code to make proteins
- Gene expression = measure of mRNA concentration, which *may* correlate with protein production

Example: Calcium signaling in yeast cells:

Procedure:

Grow 2 batches of cells: w/Ca, w/o Ca

Wait 30 minutes; harvest cells; collect mRNA

Result:

mRNA is reverse transcribed back into cDNA

cDNA comes from

“no CA” cells: labeled with green fluorophore

“w/ CA” cells: labeled with red fluorophore

Both hybridized to gene chip.

cDNA Gene chip

(tailored to specific organism):

- Place copy of each gene in defined location on chip
- cDNAs find their matching partners
- Laser scanner records fluorescence at each location

Data (Visual):

Red if gene is expressed more strongly in Ca batch

Green if gene is expressed more strongly in no-Ca batch

Yellow if not much difference

Data values: For each channel:

Foreground (spot) pixel intensities (also diameter)

Background pixel intensities

#pixels, mean, **median**, standard deviation

Manufacturing problems:

- *Gene library:*
 - Isolate sample cells
 - place in solution; unravel DNA
 - Restriction enzymes → gene probes
 - Splice probes into "vector" (bacterial DNA)
 - Transform back to host cell
 - Place on agar plate to grow and multiply
 - Reverse: splice out genes from vector

"Normal" cells?
separate genes, or gene fragments?
- *Consistency of slide preparation:*
 - inkjet-like technology
 - 4x4 or 8x4 print tip templates
 - non-uniform spots across slide
 - broken/worn-out print tips
 - registration errors
 - probe concentration/homogeneity
- *Sample preparation: 'labeling efficiency'*
 - equal volumes of experimental & control cells?
 - equal amounts of Cy3 / Cy5?
- *Instrumentation errors (Laser scanner):*
 - range of fluorescence levels in both channels
 - scanning accuracy/precision in both channels
 - effect of sample degradation over time between scan at 532nm (green) and 635nm (red)

- *background photon counts from scanner*
- *spatial effects on chip/slide*
- *comparisons of values across and within gene slides*

Analysis questions:

- Cut-off level for meaningful expression:
At what level of differential expression is gene declared Calcium-induced? ("2-fold change")
- Gene chip designs to maximize "signal-to-noise"
(Certain genes known experimentally to be induced by calcium but don't make 2-fold cut-off)
- Comparing results from different experiments
- mRNA bias? (amount of mRNA \neq protein produced)
- Other sources of variability and their effects?
(e.g., laboratory errors)
- **Background adjustment**
- **Data transformations**

Aside: More problems with humans:

- About 30,000 genes code for 100,000+ proteins.
** How to identify *groups* of genes that react together (“expression”) under similar stimuli? **
- Censored information:
We don’t know the location of every gene, so we don’t have all gene templates.

Terry Speed: series of Welch t statistics using

$$r = \log_2 R \ ; \ g = \log_2 G$$

Adjust for pin tip location/scale effects:

$$r - g \rightarrow r - g - h_i(r + g) , i = 1, \dots, 16 \text{ pins}$$

$$r - g \approx N(0, a_i^2 \sigma^2) , \sum \log a_i^2 = 0$$

$$M_{ij} = \log (R_{ij} / G_{ij}) , i = 1, \dots, 16; j = 1, \dots, n_i$$

$$\log \hat{a}_i^2 = \log \left(\sum_{j=1}^{n_i} M_{ij}^2 \right) - (\text{mean})$$

(actually used MADs)

Affymetrix chips

"Probes" = manufactured strings of 25 nucleotides

"oligos" ~ 8 amino acids + 1 base

16 – 20 oligos per gene ("replication")

"PM" = "Perfect Match": actual sequence of 25 bases in specific gene of interest

"MM" = "Mis-Match": Same as "PM", *except* at base 13 (as different as possible/random/...)

imperfect control

Manufacturing of gene array: "Photolithography"

- impurities in nucleotide chemical manufacturing
- screen deformities
- registration issues
- "edge" effects
- many others

Hybridize with cells of interest

if gene is expressed, $PM \gg MM$ (!); else $PM \approx MM$

Data: measures of light intensity (fluorescence)

Affymetrix algorithms

"have been developed as a result of mathematical modeling of data generated from experiments with GeneChip probe arrays hybridized with known amounts of target transcripts. The algorithms were optimized using extensive empirical testing."

--- User's Manual 4.0, Appendices 2-4 (2000)

"Call": Present, Absent, or Marginal

(1) *Measure fluorescence of a PM or MM probe cell*

Small: $24 \mu\text{m} \times 24 \mu\text{m}$

Large: $50 \mu\text{m} \times 50 \mu\text{m}$

Chip size: Approximately 12 mm square

Resolution: $3 \mu\text{m} \Rightarrow 8 \times 8$ pixels per probe cell

Ignore 28 border pixels \Rightarrow 36 pixel intensities

Define expression level as 75th percentile ($p_{(27)}$)
(63rd percentile gives mean of exponential?)

(2) *"Background signal intensity"*:

Size	#cells	sectors(1/16)	Lowest 2%
$(24\mu m)^2$	$512^2 = 262,144$	16,384	328
$(50\mu m)^2$	$256^2 = 65,536$	4,096	82

Average of 328 or 82 cell intensities within sector;
 subtract average from each probe cell in sector
 (→ negative values)

Divide chip into 16 sectors; calculate background
 intensity within each sector
 (crude assessment of spatial variation?)

Better: (robust) average log intensities

(3) *"Noise calculation"*:

For each background cell ($N = 16 \times 328$ or 16×82):
 Calculate std dev of pixels ($n_i = 36?$) = s_i

$$Q = (1 / N) \cdot \left[\sum_{i=1}^B s_i / \sqrt{n_i} \right] \cdot SF \cdot NF$$

SF = scale factor

NF = normalization factor

(4) *Probe pair "call": Comparing PM and MM:*

"Positive" if $PM - MM \geq 2Q$ and $PM/MM \geq 1.5$

"Negative" if $MM - PM \geq 2Q$ and $MM/PM \geq 1.5$

(2 and 1.5 are user-changeable)

16-20 probe pairs per gene (say, $n = 20$):

Let $X =$ number positive calls $\sim \text{Bi}(n, \frac{1}{2})$ under H_0

(a) Positive fraction X / n

(b) Positive/Negative ratio $X / (n - X)$

(c) Log avg ratio = $10 \sum_{i=1}^{n^*} \log(PM_i / MM_i) / n^*$

if $n \leq 8$, then $n^* = n$

else let $y_{(i)} = PM_i - MM_i$, sorted

drop min and max; calculate mean and SD = s_t

$n^* = \#\{i: |PM_i - MM_i| \leq 3 \cdot s_t\}$

Ranges for calls:

Statistic	Absent	Marginal	Present
Pos fraction	≤ 0.33 ($0 \leq X \leq 6$)	$0.33 - 0.43$ ($X = 7, 8$)	> 0.43 ($X \geq 9$)
Pos/Neg ratio	≤ 3.0 ($X = 15$)	$3.0 - 4.0$ ($X = 15, 16$)	> 4.0 $X > 16$
Log avg ratio	≤ 0.90	$0.90 - 1.30$	≥ 1.30

$\ln(X / (n - X))$: when $X \in \{1, \dots, 19\}$:

mean=0, sd=0.4735 (vs 1.12, 1.26 without ln)

Comparison algorithm: Baseline and Experimental

Each chip analyzed as before; scale and normalize

Baseline average:

$$A_B = 2\% \text{ trimmed mean of all } (PM_i, MM_i) \text{ cells}$$

Experimental average:

$$A_E = \text{same, for experimental chip}$$

Target average:

$$A_* = \text{derived from somewhere}$$

$$NF = A_B / A_E$$

$$SF = A_* / A_E$$

Multiply *all* values in experimental array by $SF \cdot NF$
(so both chips have same average)

Fold Change Increase: Two criteria

$$(PM - MM)_{\text{exp}} - (PM - MM)_{\text{base}} \geq CT$$

$$\frac{(PM - MM)_{\text{exp}} - (PM - MM)_{\text{base}}}{\max \{Q/2, \min(|PM - MM|_{\text{exp}}, |PM - MM|_{\text{base}})\}} \geq PCT$$

CT = "calculated by software using SDT (Statistical Difference Threshold) of both experimental or baseline data" (or user-defined)

PCT = Percent change threshold (default = 80)

Fold Change Decrease: Interchange "exp" and "base"

A very straightforward analysis:

Efron, Tibshirani, Storey, Tusner: *JASA* Dec 2001

"Probe reduction": summarize gene activity via average log of differences:

$$\sum_{i=1}^{20} [\log PM_i - c \cdot \log MM_i] / 20$$

$$c = \frac{1}{2} : \sum_{i=1}^{20} \log(PM_i / \sqrt{MM_i}) / 20$$

Reasons for $c = \frac{1}{2}$:

- Biological: MM_i may not be so different from PM_i (differs in only 1 base)
- $P\{\text{change} \mid \text{data}\}$ is "maximized" at $c = \frac{1}{2}$: "mild advantage ... but other comparisons confirmed its superiority ... little value in using trimmed means"

Compares summary gene average across different experiments using FDR

Summary: Much work on analysis of these gene chips remains to be done

Back to cDNA slides: An illustrated analysis

Comparing 2 mouse cell lines:

UC0Z-22 (immature) --- mast cell precursor, requires SCF (stem cell factor) *and* IL-3 (interleukin-3) for growth and proliferation

MC9 (mature) --- mast cells respond to *either* SCF *or* IL-3

Differentiation from immature (precursor) mast cell to mature stage is poorly understood

Goal:

Understand mast cell maturation process:
which genes are differentially expressed?

Layout:

Pin template 8×4 : 32 blocks (23x23)

1	2	3	4
5	6	7	8
...			
29	30	31	32

Each block has $529 = 23 \text{ rows} \times 23 \text{ columns}$ of spots

Two issues in the analysis of data:

Transformations

Background estimation

Transformations

Wide range of foreground (spot intensity) medians:

Green channel: 18 to 27667

Red channel: 289 to 43869

Variance as a function of the mean:

Propagation of error formulas

("Delta method" for random variables):

$$f(X) \approx f(\mu_x) + (X - \mu_x)f'(\mu_x) + \dots$$

$$\text{Var}(f(X)) \approx [f'(\mu_x)]^2 \cdot \sigma_x^2$$

Curtiss (1943):

if σ_x is a function of μ_x , and

ideally $\text{Var}(f(X))$ is constant, then solve for f

$$\sigma_x^2 = \alpha + \beta(\mu_x - \gamma)^2$$

$$f(x) = \log[y + (\alpha/\beta + y^2)^{\frac{1}{2}}], \quad y = (x - \gamma)$$

Notes:

argument of log is always positive

x near γ : linear function of x

x far from γ : $\log(x)$

Alternative: Tukey's g -family of distributions

$$(X - a) / b = (e^{gZ} - 1) / g$$

Z is standard normal $N(0,1)$

g is skewness parameter ($g = 0 \rightarrow Z$)

a is location parameter

b is scale parameter

$$Z = z(X) = g^{-1} \cdot \log[g \cdot (X - a)/b + 1]$$

Quantiles transform consistently:

define z_p, x_p :

$$P\{ X \leq x_p \} = P\{ Z \leq z_p \} = p$$

$$\Rightarrow z_p = g^{-1} \cdot \log[g \cdot (x_p - a)/b + 1]$$

Similar transformations:

- $f(\cdot)$: assume σ_x is really quadratic;
find f so variance is exactly the fitted quadratic
- $z(\cdot)$: assume X is really lognormal
find z so $z(X)$ is as normal as possible
(variance will be approximately quadratic)

Fitting a , b , g : Hoaglin (1985, EDTTS Ch 11):

$$x_p = a + b(e^{gz_p} - 1) / g$$

$$x_{1-p} = a + b(e^{-gz_p} - 1) / g$$

$$x_{\frac{1}{2}} = a \equiv x_M \quad (\text{median})$$

Then

$$g_p = -(1 / z_p) \cdot \log[(x_{1-p} - x_M) / (x_M - x_p)]$$

Plot g_p vs $-\log(p)$; find common g

Plot x_p vs $(e^{gz_p} - 1) / g$; *slope* $\approx b$

Repeat for all 32 blocks

Background estimation

Many background algorithms

Counts appear even in absence of spots
smearing, artifacts on slide, environment, ...

Adjustment: *foreground* – *background* (may be negative)

Many background algorithms

We will fit a simple model to each block's
background medians:

$$b_{ij} = m + row_i + col_j + res_{ij}$$

Plot row_i vs i ($i = 1, \dots, 23$ rows)

Plot col_j vs j ($j = 1, \dots, 23$ columns)

residuals structure → fit term for non-additivity:

$$b_{ij} = m + row_i + col_j + T \cdot row_i \cdot col_j + res_{ij}$$

" m " for each block: 8 "layers" and 4 "stripes"

→ fit two-way model to block terms:

$$m_{ls} = M + layer_l + stripe_s + res_{ls}$$

Subtract *fitted* background from foreground

Transform adjusted foreground values → (Z_R, Z_G)
approximately Gaussian distributed

Analysis steps:

1. Median polish applied separately to background counts in each block (possibly with smoothing of the fitted row and column effects, and possibly with the extra term for non-additivity), yielding 32 sets of fitted background counts
2. Adjust foreground counts in each channel by subtracting fitted background counts in Step 1 from reported foreground counts.
3. Estimate for each block the parameters in the g transformation to the adjusted foreground counts obtained in Step 2.
4. Transform the adjusted foreground counts using g , a , b , to obtain \approx Gaussian distributed quantities

$$Z_R = g_{kr}^{-1} \log[g_{kr}(R_{ijk} * -a_{kr})/b_{kr} + 1]$$

$$Z_G = g_{kg}^{-1} \log[g_{kg}(G_{ijk} * -a_{kg})/b_{kg} + 1]$$

5. Estimate the correlation $\hat{\rho}_k$ between Z_R and Z_G in each block.
6. Calculate an approximate standard error for the difference $Z_R - Z_G$ as $[2(1 - \hat{\rho}_k)]^{1/2}$
7. Denote as “significant” those differences that either exceed a set number of standard deviations, or achieve significance via FDR.

Results:

About 178 genes were “significant”

Additional issues:

- *foreground – background < 0:*

Red foreground/background near row 7, column 22

	Foreground				Background			
	20	21	22	23	20	21	22	23
5	677	684	600	676	273	267	263	250
6	658	1302	1149	869	276	267	258	250
7	557	630	255	786	278	272	260	240
8	660	774	1840	676	283	269	264	250
9	653	761	626	575	273	267	263	260

Red foreground/background near row 23, column 23

	Foreground				Background			
	20	21	22	23	20	21	22	23
20	694	705	696	669	345	344	328	301
21	547	828	844	848	347	344	327	297
22	695	784	609	767	361	344	319	293
23	872	858	665	277	357	339	311	283

- *Fitting common g in both channels:*
bivariate version of lognormal distribution
- *Robust correlation estimates*
- *Automation*
- *Applicable to oligonucleotide arrays*

Final comments

1. Sensible estimates of background and appropriate transformations of gene expression data are essential for the analysis of microarray data
2. Better algorithms are needed for oligonucleotide arrays [Wing Wong, D-Chip (www.dchip.org)]; methods here may be applicable (Red/Green -> PM/MM)
3. Adjustment for multiplicity is necessary
4. Usual FDR procedure safe to apply (Benjamini + Yekutieli) unless correlations are negative
5. Robust correlations of any form are probably sufficient for assessing rough direction (\pm)

FDR = False Discovery Rate

Controls # of false significances relative to number of significances claimed, rather than to number of tests made

Validity established assuming *independent* test statistics

Benjamini + Yekutieli (2001): "The Control of the False Discovery Rate in Multiple Testing under Dependency"

Primary messages:

- *Positive* correlation among test statistics presents no problem (Positive regression dependency)
- *Negative* correlation, general correlation structures, require changing the q (FDR) to $q' = q \cdot \sum_{i=1}^N (1 / i)$
 $q = 0.05$, $N = 6000$ tests on genes $\Rightarrow q' = 0.0054$, many fewer significances

Researchers at Health Sciences Center have much data from which they can assess correlation among test statistics on genes

Two questions:

1. How well does "cob" (Mosteller+Tukey 1977, p.211) perform as a robust estimate of correlation?
2. How well does FDR identify non-zero correlations among test statistics?

cob = robust correlation coefficient

(Mosteller + Tukey 1977, p.211)

$$cob(x, y) = \text{sgn}(b) \cdot [1 + s_{bi}^2(y - bx) / (b^2 s_{bi}^2(x))]^{-\frac{1}{2}}$$

where

$$s_{bi}^2(y) = [n \sum (y_i - T_{bi})^2 (1 - u_i^2)^4] / [(\sum \psi_i) \cdot (-1 + \sum \psi_i)]$$

$$u_i = (y_i - T_{bi}) / (9 \cdot MAD);$$

$$\psi_i = (1 - u_i^2)(1 - 5u_i^2)I_{[-1,1]}(u_i)$$

b = slope estimate from rreg

How well does this perform?

Brief simulation exercise:

1. 300 simulations
2. X, Y vectors have length n , $n = 20$ or $n = 50$
3. Correlations: $\rho = -0.9, -0.6, -0.3, 0, 0.3, 0.6, 0.9$
4. Uncontaminated: $X \sim N(0,1), Y = \rho X + (1 - \rho^2)^{\frac{1}{2}} Z$
5. Contaminated:
 $W \sim 0.90 N(0,1) + 0.10 N(0,100)$
 $X \sim 0.90 N(0,1) + 0.10 N(0,100)$
 $Y = \rho X + (1 - \rho^2)^{\frac{1}{2}} W$

6. Calculate $rreg(x, y)$, s_{bi}^2 on X and $rreg$ residuals

Simulation results:

1000r	(U) n=20		(U) n=50		(C) n=50	
	Mean	Median	Mean	Median	Mean	Median
-900	-809	-804	-806	-818	-810	-814
-600	-540	-534	-556	-543	-559	-580
-300	-275	-296	-288	-284	-320	-302
0	1	-4	-9	6	3	2
300	292	282	271	324	304	290
600	545	546	560	574	558	582
900	809	808	800	814	816	809

Satisfactory for small values of ρ ($|\rho| < 0.5$;
otherwise badly biased.

Might be useful for assessing direction only,
"correlation cliques" (Dennis Cox)

FDR on Correlation Matrices

Use FDR to identify high correlations in a correlation matrix (variable combinations, factor analysis, ...)

Small simulation experiment:

- (1) Given known correlation matrix R among k variables.
- (2) Generate 1000 random $N_k(0, R)$ vectors.
- (3) Compute \hat{R} .
- (4) Use FDR on the $k(k - 1) / 2$ estimated correlations.
- (5) Which were captured/missed?

A somewhat more realistic R than Toeplitz matrix, tri-diagonal, or other known correlation matrices that probably never occur in practice:

3053 U.S. counties, cancer rates [$\log_{10}(1 + rate)$]

X_1 = Prostate cancer mortality rate, white males

X_2 = Prostate cancer mortality rate, nonwhite males

X_3 = Melanoma mortality rate, white females

X_4 = Melanoma mortality rate, white males

X_5 = measure of county “urbanicity”

X_6 = Log_{10} (Total county population)

X_7 = Average January temperature

X_8 = Average July temperature

X_9 = Average # days sunlight in January

X_{10} = Average # days sunlight in July

X_{11} = Cube root of elevation

Standardized all variables by $1.5 \cdot \text{MAD}$ (*mad* in S-Plus)

Replaced all correlations within 0.08 of 0 to 0

(S-Plus choked on chol(R) otherwise)

==> 32 non-zero correlations (some quite small)

$k=11$, $n=1000$ observations ==> \hat{R}_j , $j = 1, \dots, N=1000$

How many “significant” correlations were usually found?

#signif	30	31	32	33	34	
#freq	134	306	331	128	33	
#signif	28	29	35	36	37	38
#freq	7	30	22	6	2	1

93.2% were within 2 of the correct number (30--34).

Which ones were most often missed?

-- 92.3% of them corresponded to three correlations
0.09, 0.09, -0.09 (of questionable utility)

Which ones were most often inserted as "active"?

-- none in particular, ranging in frequency from
3.4% to 8.4%

Conclusion: A successful use of FDR for this application

“The type and severity of multiplicity being used ought to depend upon the field of use, the purposes of the analysis, and the state of admitted knowledge. No one multiple-comparisons procedure can serve all purposes.”

J.W. Tukey, “Seventeen points relevant to multiplicity in clinical trials,” Merck-Temple Workshop, 13 Nov 1992

“[The two Y’s] have been working on the idea of controlling the fraction of the positive statements that you make that are wrong...It does seem to have some good properties.”

-- J.W. Tukey, in a conversation at Princeton following a Symposium in honor of his 80th birthday, 20 June 1995.