Statistical analysis of replicated microarray time course data

Terry Speed and Yu Chuan Tai
WEHI & UC Berkeley
Outline

Background and illustrative example

Multivariate approach to the problem

Small simulation, and sample results on the example.
Some characteristic features of replicated microarray time course data

• Typically short series: $k = 4$-10 time points for shorter, and 11-20 time points for longer series; often irregularly spaced; and few replications: $n \leq 5$.

• Gene expression values at different time points may be correlated, especially if common reference design is used, or a common pool of cells is sampled.

• Can be periodic, as in the cell cycle: Cho et al., 1998; Spellman et al., 1998, Storch et al., 2002.

• May have no particular pattern, as in developmental time courses: Chu et al. 1998; Wen et al. 1998; Tamayo et al. 1999.
Our illustrative example:
A plant’s response to a pathogen.

Healthy *Arabidopsis thaliana* (mustard weed) plant
**A. thaliana response to infection by E. orontii**

- Two lines of plants: *Columbia, Col-0* = wildtype (wt), and an enhanced disease susceptibility line *eds16* (mutant).

- **Objective**: to identify genes whose temporal expression patterns following infection differ between wt (*Col-0*) and mutant (*eds16*).
The experiments, I

BOX (12 Plants)

FLAT (6 Boxes)

- Plants are evenly positioned
- Wild type and mutant intermixed
- Same numbered leaves sampled each time point (leaves 5, 6, 7)
- “Random” sampling of plants from two flats for each time point
- New plants used for each time point (can’t resample)
- Sampling occurs at same time of day (with exception of 6hr)
- Each RNA sample contains leaves from 2 plants
The experiments, II

- Plants are grown in environmental chambers: T, RH, Light Intensity
- 4 week old plants were infected with heavy inoculum of powdery mildew (spores from 3 heavily infected leaves per box)
- Infection was performed using a 3 ft settling tower with mesh
- Uninfected plants were kept in similar environmental chambers
- Three separate experiments were performed
- Triplicate samples were harvested for each experiment at each time point
- Samples-0 (uninfected), 6 hr, 1 d, 3 d, 5d, 7dpi, 7d (uninfected)
Erysiphe-infected plants @ 14 dpi.
The experiments, III

• Three experiments - effectively biological replicates - were conducted using the wt and mutant lines, and within each, 3 technical replicate series. We use one series from experiments I and III, and two from experiment II.

• The Affymetrix *Arabidopsis* 24K GeneChip® was used. In all 2(genotypes)x6(times)x4(experiments) = 48 chips were hybed. In addition, data from 2(genotypes)x4(experiments) = 8 chips for day 7 uninfected samples are plotted.

• Low level analysis (background, normalization, probe set summarization) done by RMA.
Statistical question:

Find genes whose expression profiles differ between genotypes?

Extra dots at day 7 from uninfected samples.
Gene selection with time-course data: Brief literature review

Clustering methods have been widely used in this context to find groups of genes with interesting and similar patterns.


Some drawbacks of clustering methods

They make no explicit use of the replicate information. They either use all the slides or means of the replicates.

Clustering does not provide a ranking for the individual genes based on the magnitude of change in expression levels over time.

When the number of genes becomes large, clustering methods may not provide clear group patterns.

Cluster analysis may fail to detect changing genes that belong to clusters for which most genes do not change (Bar-Joseph et al. 2003).

There is the perennial question: How many clusters?
Gene selection literature review: Some other approaches

Pairwise comparisons: $t$-tests, univariate posterior odds: e.g. the $LOD$ statistic, Lönnstedt and S (2002), Smyth (2004), the moderated $t$ statistic, Smyth (2004),

Treat time as a factor in anova: ordinary $F$ (many authors)

**Ordinary F (ordF)**

Treat time and genotypes as factors with \( k \) and 2 levels, resp., and form the ratio of the times x genotypes MS to residual MS, giving an \( F_{k-1, d} \) under the null, where \( d = 2k(n-1) \) are the residual d.f.. Since there is pairing, we might also fit ‘pairs’ on 3 d.f.

**Moderated F (modF)**

As with the ordinary F, but dividing the times x genotypes MS by a moderated residual MS, giving an \( F_{k-1, \nu+d} \) under the null, where \( \nu \) are the prior d.f.
What did we want?

We sought a formula to rank genes, in order to

• find those changing or not similarly expressed
• provide a cut off for clustering

We felt that this formula should be

• $t$-like, i.e. involve a standardized measure of an effect
• multivariate
• moderated
Why moderation?

• We sought genes with large overall amounts of change across time, relative to their replicate variances and covariances.

• Variances and covariance are poorly estimated in this context.

• Some sort of smoothing, borrowing strength, or empirical Bayes approach is called for.

• We use multivariate normals with conjugate priors, as we want usable formulae, and not to have to use MCMC.
Univariate moderated $t$ & posterior odds (LOD)

$$\tilde{s}^2 = \frac{(n-1)s^2 + \nu\lambda^2}{n-1 + \nu}$$

$$\tilde{t}^2 = \frac{M^2}{\tilde{s}^2 / n}$$

Moderated $s^2$ of $M$ values

Moderated $t$

$$B = LOD = c + \left( \frac{n + \nu}{2} \right) \log_{10} \left\{ \frac{\tilde{t}^2 + n - 1 + \nu}{\frac{1}{n} \frac{1}{n + 1} + n - 1 + \nu} \right\}$$

$B = \log_{10}$ of posterior odds against differential expression
Multivariate approaches

Here we treat one entire series as a random $k$-vector
Notation and models

We denote by $X_{g,1}, \ldots, X_{g,n}$ the replicate random $k$-vectors representing the observed time series for a single gene. For our Example, $n = 4$ and $k = 6$, and the $X_{g,i,t}$ are differences of log intensities, i.e. log ratios.

Our underlying model is that these $X_{g,i}$ are i.i.d. $N(\mu_g, \Sigma_g)$, and we make different assumptions about $\mu_g$ and $\Sigma_g$.

With our Example, we are interested in testing the null hypothesis $H: \mu_g = 0, \Sigma_g > 0$, against the alternative $K: \mu_g \neq 0, \Sigma_g > 0$. (Another null hypothesis of interest is that the mean is constant, but we do not discuss it here.)
For our empirical Bayes (EB) approach, we have priors for $\mu_g$ and $\Sigma_g$ reflecting the indicator status $l = l_g$ of the gene, where $l_g = 1$ if $\mu_g \neq 0$, and $l_g = 0$ if $\mu_g = 0$.

We suppose that $Pr(l_g = 1) = p$, independently for every gene, for a hyperparameter $p$, $0 < p < 1$.

From now on, we drop the subscripts $g$ wherever possible.
Notation and models, completed

With this background, our prior for $\Sigma$ is inverse Wishart with degrees of freedom $\nu$ and matrix parameter $(\nu \Lambda)^{-1}$, where $\Lambda > 0$ is positive definite. Our prior for $\mu$ is given by:

$$
\mu \mid \Sigma, I = 1 \sim \mathcal{N}(0, \eta^{-1} \Sigma), \quad \eta > 0, \text{ and}
$$

$$
\mu \mid \Sigma, I = 0 \sim \mathcal{N}(0,0).
$$

Finally, the data $X_1, \ldots, X_n$ are supposed i.i.d. given $I$, $\Sigma$ and $\mu$, with $X_i \mid I, \Sigma, \mu \sim \mathcal{N}(\mu, \Sigma)$.

The multivariate normality assumption is reasonable, but not precise. However, we judge our results by their utility, not on goodness-of-fit of the models.
Some distributional facts

The marginal distribution of the sample covariance matrix $S$ is Wishart, while that of $\bar{X}$ given $I=1$ is a multivariate $t$, and the distribution of $S$ given $I=1$ is a matrix generalized type II beta distribution, also called a Siegel distribution.

The joint distribution of $(\bar{X}, S)$ given $I=1$ is called a Student-Siegel, and is a little complicated. The same forms arise in the case $I=0$, but with different parameters.

These facts can be combined to obtain the simple formula on the next page for the posterior odds against a gene having zero mean vector.
Summary of results

Our moderated \( S \) is
\[
\tilde{S} = [E(\Sigma^{-1} \mid S)]^{-1} = \frac{(n-1)S + \nu \Lambda}{n - 1 + \nu},
\]
our moderated \( t \)-statistic is
\[
\tilde{t} = n^{1/2} \tilde{S}^{-1/2} \bar{X}.
\]

Finally,
\[
O = \frac{P(I = 1 \mid \text{data})}{P(I = 0 \mid \text{data})} = \left( \frac{p}{1 - p} \right) \frac{P(\tilde{t} \mid I = 1)}{P(\tilde{t} \mid I = 0)}
\]
is an increasing function of \( \tilde{T}^2 = \tilde{t}' \tilde{t} \).

We write \( MB = \log_{10} O \) for our multivariate \( B \)-statistic.
Likelihood Ratio statistic

For the likelihood ratio (LR) test, we simply test the null $H$ against the alternative $K$ in the usual way. We calculate:

$$LR = 2(l_K^{\text{max}} - l_H^{\text{max}}) = n \log(1 + \frac{n}{n-1} \overline{X}^T S^{-1} \overline{X})$$

$$= n \log(1 + T^2 / (n - 1))$$

where $S$ is assumed non-singular. Here $T^2$ is Hotelling's statistic. In our case, $n < k$ and $S$ is singular. If we plug in $\tilde{S}$, our moderated $S$, we get the moderated Hotelling statistic, $\tilde{T}^2$, just seen.
Hyperparameter estimation

There are $k(k+1)/2 + 3$ parameters in the prior: $\Lambda$, $p$, $\nu$, and $\eta$.

We simply choose $p = 0.02$, although clearly more could be done here. Neither $p$ nor $\eta$ enter into $\tilde{T}^2$.

Estimates of the hyperparameters $\nu$ and $\eta$ are developed using the univariate approach of Smyth (2004): $\eta$ using the $p/2$ genes with the highest $\tilde{T}^2$-values, and $\nu$ using all the genes. We omit the details.

$\Lambda$ is estimated by the method of moments using the formula

$$E(S) = (\nu-k-1)^{-1}\nu\Lambda.$$
The estimate of $\Lambda$

$100 \times \text{SD}: 14, 17, 15, 13, 16, 16.$

Correlation matrix

\[
\begin{array}{cccc}
1.00 & 0.15 & 0.12 & 0.05 \\
0.15 & 1.00 & 0.15 & 0.06 \\
0.12 & 0.15 & 1.00 & 0.02 \\
0.05 & 0.06 & 0.02 & 1.00 \\
\end{array}
\]
OPC-OD expt: av single gene covariance matrix

\[
\mathbf{S} = \begin{bmatrix}
0.10 & 0.06 & 0.05 & 0.04 & 0.03 & 0.03 & 0.03 & 0.02 \\
0.06 & 0.11 & 0.06 & 0.05 & 0.04 & 0.04 & 0.03 \\
0.05 & 0.06 & 0.11 & 0.05 & 0.04 & 0.04 & 0.03 \\
0.04 & 0.05 & 0.05 & 0.09 & 0.04 & 0.04 & 0.03 \\
0.03 & 0.04 & 0.04 & 0.04 & 0.09 & 0.04 & 0.03 \\
0.03 & 0.04 & 0.04 & 0.04 & 0.04 & 0.10 & 0.05 & 0.04 \\
0.03 & 0.04 & 0.04 & 0.04 & 0.05 & 0.09 & 0.04 \\
0.02 & 0.03 & 0.03 & 0.03 & 0.04 & 0.04 & 0.07 \\
\end{bmatrix}
\]

\[
\begin{align*}
14.6 & -4.6 & -3.1 & -1.5 & -0.7 & -0.3 & -0.1 & -0.3 \\
-4.6 & 15.8 & -3.3 & -3.0 & -1.4 & -1.1 & -1.3 & -0.4 \\
-3.1 & -3.3 & 16.4 & -3.1 & -1.4 & -1.5 & -1.4 & -1.7 \\
-1.5 & -3.0 & -3.1 & 18.3 & -3.0 & -2.1 & -2.1 & -1.5 \\
\end{align*}
\]

\[
\mathbf{S}^{-1} = \begin{bmatrix}
-0.7 & -1.4 & -1.4 & -3.0 & 16.9 & -2.5 & -3.0 & -1.0 \\
-0.3 & -1.1 & -1.5 & -2.11 & -2.5 & 17.6 & -4.8 & -4.2 \\
-0.1 & -1.3 & -1.4 & -2.1 & -3.0 & -4.8 & 21.0 & -5.9 \\
-0.3 & -0.4 & -1.7 & -1.5 & -1.0 & -4.3 & -5.9 & 22.1 \\
\end{bmatrix}
\]

If \(\mathbf{S} = U \Lambda V^T\), \(\Lambda = \text{diag}(0.38, 0.09, 0.06, 0.05, 0.05, 0.05, 0.04, 0.04)\),
\[
u_1 = \begin{bmatrix}
0.34 & 0.41 & 0.40 & 0.37 & 0.32 & 0.37 & 0.34 & 0.27
\end{bmatrix}_T
\]
Small simulation study based on a different example

100 data sets were simulated using parameters estimated from another data set with \( n=3 \) and \( k=8 \). The hypotheses were

\[ H: \mu_g = \mu_{g0}1, \Sigma > 0, \]

against

\[ K: \mu_g \neq \mu_{g0}1, \Sigma > 0. \]

Here \( 1 = 1_k \) is the \( k \)-vector of 1s. There were 20,000 genes in total, and with \( p=0.02 \), we had 400 non-constant genes.

The theory is just a bit more complicated than for the case

\[ H: \mu_g = 0, \Sigma > 0, \]

but we omit the details.

The simulation is straightforward, drawing a structured \( \Sigma \) first, then for genes with \( l=0 \) or \( l=1 \), draw \( \mu \), and finally, given \( \mu \) and \( \Sigma \), we draw \( X_1, X_2 \) and \( X_3 \) independently from \( N(\mu,\Sigma) \), and independently for all genes.

Of course this kind of simulation will not capture the complex dependence between gene expression patterns.
ROC curves

All eight statistics

No. False positives

No. False negatives

- MB
- MB, k independent samples
- MB, 100% moderation
- MB, 0% moderation
- F
- Moderated F
- One-sample moderated LR
- Variance
ROC curves for the best 6 statistics
Top-ranked genes for our Example

Next we display the profiles of the 12 highest ranked genes according to the MB (= modLR), modF and ordF statistics.

After that, we look at profiles of the top 4 ranked out of the $m \leq 200$, which are not in the top 200 of a different statistic.

Our aim is to exhibit the temporal profiles that get highly ranked by one of these statistics, but not another, in order to assess the efficacy of the statistics.
Top 4 by MB statistic
MB statistic: 9-12

All 12 (and many more) down in the mutant, relative to wt.
First 4 genes up in the mutant, by MB
Top 4 by modF = top4 by ordF
modF: 5-8
modF: 9-12
## Top 200 genes by MB, modF and ordF

<table>
<thead>
<tr>
<th>Pair</th>
<th># common</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>MB</td>
<td>mod F</td>
<td>72</td>
</tr>
<tr>
<td>MB</td>
<td>ordF</td>
<td>67</td>
</tr>
<tr>
<td>mod F</td>
<td>ordF</td>
<td>188</td>
</tr>
</tbody>
</table>
Top 4 of 128 selected by MB but not by mod F

MB rank=3
modF rank=299
ordF rank=359

MB rank=6
modF rank=279
ordF rank=330

MB rank=9
modF rank=2216
ordF rank =2654

MB rank=10
modF rank=670
ordF rank=805
Discussion

• It seems clear that the ordinary and moderated $F$ are very similar. This is probably because there are plenty of error d.f. (36 or 33), and so the moderation doesn’t have much effect. We could and probably should deliberately increase the amount of moderation in the $F$.

• The $MB$ ranking seems much better, but here it is probably due to the much greater amount of moderation rather than the incorporation of correlation, as there is not much here. In other examples, the serial correlation is much greater, and the difference between including and not including it greater.
Conclusions

• Methods which rank genes (e.g. the $MB$ statistic or the moderated Hotelling $T^2$) perhaps provide easier access to genes whose absolute or relative expression varies over time, than do multi-gene methods (e.g. cluster analysis).

• Among the single-gene methods, $MB$ performs no worse than other methods in both real data and simulated data comparisons, and better than the $F$.

• The Hotelling $T^2$ statistic is a viable alternative to $MB$, but we still need the moderated $S$.

• The $MB$ statistic may be able to select interesting genes which are missed by other methods.
Extensions completed, work in progress and future work

- Obtain MB statistics for ranking genes by evidence for non-constancy of gene expression
- Obtain MB statistics for two unpaired samples
- Extend the EB to linear models (almost completed), thereby dealing with >2 groups, etc.
- Software development/R package
- Change the EB model to deal with cross-sectional rather than longitudinal data
Future work, cont.

• Clustering: identification of co-regulated groups of genes
• Coupling this to cis-acting regulatory elements, leading to predictions and tests (e.g. WRKY38, candidate ethylene response factors)
• Using what we know, to help build up pathway knowledge, e.g. eds16-1 is a null mutant for isochorismate synthase 1, pathogenesis related protein 1 (PR1) is inactivated, PR2, PR5 are down, plant defensin 1.2 (PDF1.2) is up…etc.
• Overlaying data on AraCyc, to predict altered metabolic pathways, networks and physiology
• Looking at functional annotation of genes in clusters to get at functional responses.
• Build up a systems view of the host-pathogen system!
Acknowledgements

Yu Chuan Tai, UC Berkeley

Mary Wildermuth, UC Berkeley

Jason Dugas, Stanford

Moriah Szpara, UC Berkeley & members of the Ngai lab

Gordon Smyth, WEHI

Monica Nicolau, Stanford University