Microarray experimental design and analysis

Sandrine Dudoit and Robert Gentleman

Bioconductor short course
Summer 2002

Outline

• Experimental design for cDNA microarray experiments.
• Combining data across slides for cDNA microarray experiments.
• Multiple testing.
• A 2x2 factorial microarray experiment.

Combining data across slides

Data on \( G \) genes for \( n \) hybridizations

\[ G \times n \] genes-by-arrays data matrix

Arrays

\[
\begin{array}{c|cccccc}
\text{Gene1} & \text{Array1} & \text{Array2} & \text{Array3} & \text{Array4} & \text{Array5} \\
0.46 & 0.30 & 0.80 & 1.51 & 0.90 & \ldots \\
\text{Gene2} & -0.10 & 0.49 & 0.24 & 0.06 & 0.46 & \ldots \\
\text{Gene3} & 0.15 & 0.74 & 0.04 & 0.10 & 0.20 & \ldots \\
\text{Gene4} & -0.45 & -1.03 & -0.79 & -0.56 & -0.32 & \ldots \\
\text{Gene5} & -0.06 & 1.06 & 1.35 & 1.09 & -1.09 & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\end{array}
\]

\[
M = \log_2(\text{Red intensity} / \text{Green intensity})
\]
expression measure, e.g., RMA
Combining data across slides

... but columns have **structure**

*How can we design experiments and combine data across slides to provide accurate estimates of the effects of interest?*

**Experimental design**

- Proper experimental design is needed to ensure that questions of interest can be answered and that this can be done **accurately**, given experimental constraints, such as cost of reagents and availability of mRNA.

**Experimental design**

- Design of the array itself
  - which cDNA probe sequences to print;
  - whether to use replicated probes;
  - which control sequences;
  - how many and where these should be printed.
- Allocation of target samples to the slides
  - pairing of mRNA samples for hybridization;
  - dye assignments;
  - type and number of replicates.
Graphical representation

Multi-digraph
- **Vertices**: mRNA samples;
- **Edges**: hybridization;
- **Direction**: dye assignment.

The structure of the graph determines which effects can be estimated and the precision of the estimates.
- Two mRNA samples can be compared only if there is a path joining the corresponding two vertices.
- The precision of the estimated contrast then depends on the number of paths joining the two vertices and is inversely related to the length of the paths.

Direct comparisons **within slides** yield more precise estimates than indirect ones between slides.

Comparing K treatments

(i) Common reference design   (ii) All-pairs design

**Question**: Which design gives the most precise estimates of the contrasts A1-A2, A1-A3, and A2-A3?

Comparing K treatments

**Answer**: The all-pairs design is better, because comparisons are done **within slides**.

For the same precision, the common reference design requires three times as many hybridizations or slides as the all-pairs design.

In general, for K treatments

Relative efficiency

\[ \frac{2K}{K-1} = 4, 3, \frac{8}{3}, \ldots \rightarrow 2. \]

For the same precision, the common reference design requires \( 2K/(K-1) \) times as many hybridizations as the all-pairs design.
2 x 2 factorial experiment
two factors, two levels each

Scaled variances of estimated effects

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main effect A</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4/3</td>
<td>1</td>
</tr>
<tr>
<td>Main effect B</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Interaction AB</td>
<td>3</td>
<td>3</td>
<td>4/3</td>
<td>8/3</td>
<td>2</td>
</tr>
<tr>
<td>Contrast A-B</td>
<td>2</td>
<td>2</td>
<td>4/3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Time course

Possible designs
1) All samples vs. common pooled reference
2) All samples vs. time 1
3) Direct hybridizations between timepoints

From Yee Hwa Yang (2002)

Experimental design

- In addition to experimental constraints, design decisions should be guided by the knowledge of which effects are of greater interest to the investigator.
  E.g. which main effects, which interactions.

- The experimenter should thus decide on the comparisons for which he wants the most precision and these should be made within slides to the extent possible.
Experimental design

- N.B. Efficiency can be measured in terms of different quantities
  - number of slides or hybridizations;
  - units of biological material, e.g. amount of mRNA for one channel.

Issues in experimental design

- Replication.
- Type of replication:
  - within or between slides replicates;
  - biological or technical replicates
    i.e., different vs. same extraction:
    generalizability vs. reproducibility.
- Sample size and power calculations.
- Dye assignments.
- Combining data across slides and sets of experiments:
  regression analysis ... next.

2 x 2 factorial experiment

two factors, two levels each

Study the joint effect of two treatments (e.g. drugs), A and B, say, on the gene expression response of tumor cells.

There are four possible treatment combinations

AB: both treatments are administered;
A : only treatment A is administered;
B : only treatment B is administered;
O : cells are untreated.

For each gene, consider a linear model for the joint effect of treatments A and B on the expression response.

\[
\mu_{AB} = \mu + \alpha + \beta + \gamma
\]

\[
\mu_A = \mu + \alpha
\]

\[
\mu_B = \mu + \beta
\]

\[
\mu_O = \mu
\]

\(\mu\): baseline effect;
\(\alpha\): treatment A main effect;
\(\beta\): treatment B main effect;
\(\gamma\): interaction between treatments A and B.
2 x 2 factorial experiment

Log-ratio M for hybridization estimates

\[ \mu_{AB} - \mu_A = \beta + \gamma \]

Log-ratio M for hybridization estimates

\[ \mu_B - \mu_A = \beta - \alpha \]

+ 10 others.

Regression analysis

- For parameters \( \theta = (\alpha, \beta, \gamma) \), define a design matrix \( X \) so that \( E(M) = X\theta \).
- For each gene, compute least squares estimates of \( \theta \).

\[
\hat{\theta} = (X'X)^{-1}X'M
\]

Regression analysis

- Combine data across slides for complex designs - can “link” different sets of hybridizations.
- Obtain unbiased and efficient estimates of the effects of interest (BLUE).
- Obtain measures of precision for estimated effects.
- Perform hypothesis testing.
- Extensions of linear models – generalized linear models; – robust weighted regression, etc.

Regression analysis

- Use estimated effects in clustering and classification

\[
\text{genes x arrays matrix} \rightarrow \text{genes x estimated effects matrix}
\]
Multiple testing

Differential gene expression

• Identify genes whose expression levels are associated with a response or covariate of interest
  – clinical outcome such as survival, response to treatment, tumor class;
  – covariate such as treatment, dose, time.
• Estimation: estimate effects of interest (e.g. difference in means, slope, interaction) and variability of these estimates.
• Testing: assess the statistical significance of the observed associations.

Hypothesis testing

• Test for each gene the null hypothesis of no differential expression, e.g. using t- or F-statistic.
  Two types of errors can be committed
• Type I error or false positive
  – say that a gene is differentially expressed when it is not, i.e.
    – reject a true null hypothesis.
• Type II error or false negative
  – fail to identify a truly differentially expressed gene, i.e.
    – fail to reject a false null hypothesis.

Multiple hypothesis testing

• Large multiplicity problem: thousands of hypotheses are tested simultaneously!
  – Increased chance of false positives.
  – E.g. chance of at least one p-value < \( \alpha \) for \( G \) independent tests is \( 1 - (1 - \alpha)^G \)
    and converges to one as \( G \) increases.
  For \( G=1,000 \) and \( \alpha = 0.01 \), this chance is 0.9999568!
  – Individual p-values of 0.01 no longer correspond to significant findings.
• Need to adjust for multiple testing when assessing the statistical significance of the observed associations.
Multiple hypothesis testing

- Define an appropriate Type I error or false positive rate.
- Develop multiple testing procedures that
  - provide strong control of this error rate,
  - are powerful (few false negatives),
  - take into account the joint distribution of the test statistics.
- Report adjusted p-values for each gene which reflect the overall Type I error rate for the experiment.
- Resampling methods are useful tools to deal with the unknown joint distribution of the test statistics.

Type I error rates

- Per-family error rate (PFER). Expected number of false positives, i.e.,
  \[ \text{PFER} = E(V). \]
- Per-comparison error rate (PCER). Expected value of (# false positives / # of hypotheses), i.e.,
  \[ \text{PCER} = E(V)/G. \]
- Family-wise error rate (FWER). Probability of at least one false positive, i.e.,
  \[ \text{FWER} = p(V > 0). \]

From Benjamini & Hochberg (1995)

Type I error rates

- False discovery rate (FDR). The FDR of Benjamini & Hochberg (1995) is the expected proportion of false positives among the rejected hypotheses, i.e.,
  \[ \text{FDR} = E(Q), \]
  where by definition
  \[ Q = V/R, \text{ if } R > 0, \]
  \[ 0, \text{ if } R = 0. \]
Strong control

- N.B. Expectations and probabilities above are conditional on which hypotheses are true.
- **Strong control.** Control of the Type I error rate under any combination of true and false hypotheses.
- **Weak control.** Control of the Type I error rate under only the complete null hypothesis, i.e., when all null hypotheses are true.
- **Strong control** is essential in microarray experiments.

Comparison of error rates

- In general, for a given multiple testing procedure, \( \text{PCER} \leq \text{FWER} \leq \text{PFER} \)
  and \( \text{FDR} \leq \text{FWER} \)
  with \( \text{FDR} = \text{FWER} \) under the complete null.
- Thus, for a fixed criterion \( \alpha \) for controlling the Type I error rates, the order reverses for the number of rejected hypotheses \( R \): procedures controlling the FWER are generally more conservative than those controlling either the FDR or PCER.

Adjusted p-values

- Given any test procedure, the adjusted p-value for a single gene \( g \) can be defined as the level of the entire test procedure at which gene \( g \) would just be declared differentially expressed.
- Adjusted p-values reflect for each gene the overall experiment Type I error rate when genes with a smaller p-value are declared differentially expressed.
- Can be estimated by resampling, e.g. permutation or bootstrap.

Multiple testing procedures

- **Strong control of FWER**
  - Bonferroni: single-step;
  - Holm (1979): step-down;
  - Hochberg (1986)*: step-up;
- **Strong control of FDR**
  - Benjamini & Hochberg (1995)*: step-up;

*some distributional assumptions required.
Multiple testing procedures

- Golub et al. (1999): neighborhood analysis
  - weak control only, problematic definition of error rate.

- Tusher et al. (2001): SAM
  - t- or F-like statistics;
  - similar to univariate test with asymmetric cut-offs;
  - permutation procedure controlling PCER;
  - the SAM estimate of the FDR is $E_0(V)/R$ -- can be greater than one.

A FAQ

- **Q:** What about pre-screening to reduce the number of tests with the aim of increasing power?
- **A:** Type I error is controlled in situations where
  - we only focus on a subset of genes that are of interest
    - selected before looking at the data;
  - the statistic used for screening is independent of the test statistic under the null.
- Other situations still need to be better understood.

Discussion

- Microarray experiments have revived interest in multiple testing
  - lots of papers;
  - old methods with new names;
  - new methods with inadequate or unknown control properties;
  - a lot of confusion!
- New proposals should be formulated precisely, within the standard statistical framework, to allow a clear assessment of the properties of different procedures.
R multiple testing software

- Bioconductor R `multtest` package.
- Multiple testing procedures for controlling
  - FWER: Bonferroni, Holm (1979), Hochberg (1986), Westfall & Young (1993) maxT and minP.
- Tests based on t- or F-statistics for one- and two-factor designs.
- Permutation procedures for estimating adjusted p-values.
- Fast permutation algorithm for minP adjusted p-values.
- Documentation: tutorial on multiple testing.

More detailed slides and references in

**Multiple testing in DNA microarray experiments**

available at [www.bioconductor.org](http://www.bioconductor.org)

A 2x2 factorial microarray experiment

Robert Gentleman, Denise Scholtens
Arden Miller, Sandrine Dudoit

Complexity of genomic data

- The functioning of cells is a complex and highly structured process.
- In the next slide we show a stylized biochemical pathway (adapted from Wagner, 2001).
- There are transcription factors, protein kinase and protein phosphatase reactions.
- Tools are being developed that allow us to explore this functioning in a multitude of different ways.
Overview

- Wagner (2001) suggests that the holy grail of functional genomics is the reconstruction of genetic networks.
- In this tutorial we examine some methods for doing this in factorial genome wide RNA expression experiments.
- Such experiments are easy to carry out and are becoming widespread. Tools for analyzing them are badly needed.

Gene effects

- A factor can either inhibit or enhance the production of mRNA for any gene.
- The inhibition or enhancement of mRNA production for any given gene can affect transcription for other genes either through inhibition or enhancement.

Targets

- We define a target of a factor to be a gene whose expression of mRNA is altered by the presence of the factor.
- A primary target is a target that is directly affected by the factor.
- A secondary target is a target whose transcription is altered only via the effects of some other genes, i.e., can be traced back to one or more primary targets.
Factorial experiments

- We assume that there are two factors of interest, \( F_1 \) and \( F_2 \).
- A 2x2 microarray experiment can be used to measure the expression response (mRNA level) of each gene under the four conditions:
  - nothing
  - \( F_1 \) alone
  - \( F_2 \) alone
  - \( F_1 \) and \( F_2 \).

Factorial experiments

- Experimental units depend on the population of interest (i.e., for which the inference is desired). They may be cells from the same cell line, patients, or different inbred model organisms.
- Questions of interest often involve identifying which genes are directly affected by the two factors \( F_1 \) and \( F_2 \).

CX experiment

- There are two factors
  - Estrogen, \( E \): known to affect transcription of various genes (some known, some unknown).
  - Cyclohexamide, \( CX \): known to stop all translation (with very few exceptions).
- The design is a classical 2x2 factorial design, with two replicates.
- We are interested in the main effects and interactions for \( E \) and \( CX \).
**CX experiment**

- We identify as **targets** all genes whose expression of mRNA is affected by the application of E.
- A target can be either primary or secondary
  - primary if E directly affects expression of mRNA.
  - secondary if mRNA production is affected by some other gene and can be traced back to a primary target.

---

**Scenario 1**

- Assume that there are two related genes, B and D, where
  - B is a primary target of E,
  - D is a secondary target only via B.
- Neither is expressed initially.
- E causes B to be expressed and this in turn causes D to be expressed.
- The addition of CX by itself may not affect expression of either B or D.
Scenario 1

- In the presence of both CX and E we see increased expression of mRNA_B but not of mRNA_D.
- CX stops translation of B and hence transcription of D.
- This will be one of the principles we can use to differentiate between primary targets of E (such as B) and secondary targets of E (such as D).

Interpretation: Scenario 1

<table>
<thead>
<tr>
<th></th>
<th>mRNA_B</th>
<th>mRNA_D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nothing</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>E</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>CX</td>
<td>Low(?)</td>
<td>Low(?)</td>
</tr>
<tr>
<td>E and CX</td>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>

Scenario 1

- Note that while we show a direct relationship between the expression of B and of D we cannot detect such a relationship from these data (the purpose of this scenario is purely pedagogical).
- Other scenarios include
  - Suppression of D by B, enhancement of B by E.
  - Enhancement of D by B, and suppression of B by E.
**CX experiment**

- Assume the following linear model for the observed expression response (possibly on transformed data) of any given gene

\[ y_{ig} = \mu_g + \beta_{Eg} x_{i1} + \beta_{Cg} x_{i2} + \beta_{E|C|g} x_{i1} x_{i2} + \epsilon_{ig} \]

- \( i \) indexes chips and \( g \) indexes genes.
- \( x_1 \) indicates the presence of E and \( x_2 \) indicates the presence of CX.

**Inference**

- The 2x2 CX microarray experiment measures the expression response of each gene under each of the four factor combinations.

- But there is a difference, B is a primary target of E, while D is a secondary target of E.

**Inference**

- If gene X is any target for E, the level of mRNA_X might not change when E is added.
- mRNA_X might already be being made as fast as possible, so addition of E has no effect.
- Production of mRNA_X might already be suppressed by some other compound.
- A true baseline would help in resolving these situations.

**Inference**

- The introduction of CX provides a form of baseline.

- Since (among other things) CX halts translation we should be able to use the presence or absence of CX to find out about primary versus secondary targets.
Inference

- For any gene we can interpret the coefficients in the linear model as follows.
- The parameter $\beta_E$ can be interpreted as the main effect of E.
- Genes for which $\beta_E$ is different from zero are potential targets.
- As noted previously, not all targets will have $\beta_E$ different from zero.

Inference

- The parameter $\beta_{CX}$ can be interpreted as the main effect of CX.
- If $\beta_{CX}$ is different from zero, this suggests that production of mRNA is translationally regulated.
- The interpretation of the interaction $\beta_{E:CX}$ is more difficult.

Primary targets

- Consider the case where we have only CX and CX+E.
- Since CX halts all translation, then any differences between the condition where CX alone is present and CX+E is present should indicate primary targets of E.
- This is equivalent to testing the hypothesis

$$H_0: \mu + \beta_E + \beta_{CX} + \beta_{E:CX} = \mu + \beta_{CX}$$

i.e.,

$$H_0: \beta_E + \beta_{E:CX} = 0$$

Primary targets

- Genes for which the hypothesis

$$H_0: \mu + \beta_E + \beta_{CX} + \beta_{E:CX} = \mu + \beta_{CX}$$

is rejected are candidates for primary targets.
- Those with $\beta_E$ different from zero, but for which we do not reject $H_0$, are secondary targets.
- It seems likely that some inference may be drawn from the relationship between $\beta_E$ and $\beta_{E:CX}$, their signs and their significance levels.
Scenario 1

<table>
<thead>
<tr>
<th></th>
<th>Primary</th>
<th>Secondary</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_E$</td>
<td>$&gt; 0$</td>
<td>$&gt; 0$</td>
</tr>
<tr>
<td>$\beta_{CX}$</td>
<td>$= 0$</td>
<td>$= 0$</td>
</tr>
<tr>
<td>$\beta_{E:CX}$</td>
<td>$= 0$</td>
<td>$- \beta_E$</td>
</tr>
</tbody>
</table>

Limitations

- While we may identify genes that are potentially primary targets and those that are potentially secondary targets we cannot identify gene—gene interactions, or feedback loops.
- We can observe the effects but not attribute them.
- The use of relevant metadata, biological and publication, seems pertinent and could help resolve some of the interactions.

Factorial experiments

- These experiments can be contrasted with those proposed by Wagner (2001).
- He proposes perturbing each gene in the genome of interest and observing the gene specific effects.
- We consider very few experiments and observe genome wide changes and hence less specific information.
- The two methods can be complementary since the results of the genome wide study could be used to design several single gene experiments.

Methylation experiments

- Methylation inhibits transcription of specific genes.
- If a factor that demethylates the genome were available, then one could, in principle, determine which genes were methylated (or affected by methylated genes).
- However, we could not determine which genes were primary and which were secondary targets.
Phosphorylation experiments

- Many cellular reactions are carried out using energy that is provided by the ADP ATP phosphorylation mechanism.
- If a simple mechanism was available for halting this process then that could be used as a factor in these experiments and genes whose transcription is affected by phosphorylation could be identified.