Design And Analysis of 2-Channel Microarray Experiments

Jane Chang
Bowling Green State University
Joint With
Jason C. Hsu And Tao Wang
The Ohio State University
Graybill Conference 2003
Uses of Microarrays

A Microarray is a bio-technology for studying gene expressions.

- Designer medicine
  - Screen genes to build diagnostic or prognostic chip

- Patient targeting
  - Eliminate patient subgroup prone to serious side effects

- Drug discovery
  - Find proteins to synthesize or suppress to treat the disease
Research Goals

- **Biomedical**: Infer, from genes differentially expressed in normal vs diseased tissues, proteins involved in a disease (drug discovery).

- **Statistical**: Find a good design to efficiently and effectively identify the genes that are differentially expressed between 2 (or more) treatments.
Outline of Presentation

- Introduction to 2-channel microarray experiments
- Modeling Approach
- Simultaneous Confidence Intervals
- The Design ($t=2$)
- A- Optimality for gene allocation
- The Design ($t>2$)
What Is A 2-channel Microarray?

- A glass slide on which thousands of spots of cDNA representing different genes are printed using a robotic arrayer.
2-channel Microarray Experiment

Culture bacteria
30°C  42°C

mRNA

Reverse transcription

Cy3
Cy5

cDNA

mix samples

Hybridize

Glass slide microarray

Scan & analyze
## 2-Channel Microarray Data

<table>
<thead>
<tr>
<th>Dye</th>
<th>Array 1</th>
<th>Array 2</th>
<th>……</th>
<th>Array a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>1</td>
<td>3123.40</td>
<td>8143.71</td>
<td>……</td>
</tr>
<tr>
<td>Dye</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>4522.56</td>
<td>4722.66</td>
<td>……</td>
<td>4522.86</td>
</tr>
<tr>
<td>Red</td>
<td>1</td>
<td>6123.27</td>
<td>4123.47</td>
<td>……</td>
</tr>
<tr>
<td>Dye</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>9521.35</td>
<td>7522.54</td>
<td>……</td>
<td>4522.23</td>
</tr>
</tbody>
</table>
Microarray Experiment

- Fabricate cDNA array
- Prepare mRNA sample
- Hybridize sample (mRNA) to probe (cDNA)
- Scan image
- Quantify image
Control of Error Rate

- **False Discovery Rate (FDR)**
  - Control expected proportion of false positives

- **Familywise Type I Error (FWER)**
  - Control probability of at least one false positive
    - a. Tests of equality
      - Individual P-values
      -Multiplicity Adjusted P-values
    - b. Confidence intervals
      - Individual confidence intervals (*Kerr et al, 2000*)
      - Simultaneous confidence intervals
Model for Gene Expression Data

\[ x_{ijklm} = \text{observed gene expression} \]

\[ y_{ijklm} = \log(x_{ijklm}) = \mu + T_k + G_l + (TG)_{kl} + A_i + D_j + \varepsilon_{ijklm} \]

- \( T_k \): treatment effect
- \( G_l \): block effect
- \( A_i \): block effect
- \( D_j \): block effect
- \( \varepsilon_{ijklm} \): noise
Choice of “Control” Genes

- All genes
  - Average of all genes (Kerr, Churchill et al)

- Housekeeping genes (expressed in all cells)
  - Median of housekeeping genes (Amaratunga and Cabrera)
  - Average of housekeeping genes (Hsu, Chang, Wang)
If $H$ denotes the set of housekeeping genes, then parameters of interest are

$$[(TG)_{2g} - (TG)_{1g}] - [\Sigma_{g^* \in H} (TG)_{2g^*} / |H| - \Sigma_{g^* \in H} (TG)_{1g^*} / |H|]$$
Each gene can be spotted **unequal number of times** on each array.

The number of replications of each gene remains constant across all arrays.
Least Squares Estimation

- With i.i.d. normal errors, estimating $(TG)_{tg}$ by its sample mean results in BLUE for

$$(TG)_{2g} - (TG)_{1g} - \left[ \sum_{g* \in H} \frac{(TG)_{2g*}}{|H|} - \sum_{g* \in H} \frac{(TG)_{1g*}}{|H|} \right]$$
Other Models Handled by $2 \times a$

Generalized Latin Square Design

- **Orthogonality**

  $$Y_{ijklm} = \mu + T_k + G_l + (TG)_{kl} + A_i + D_j + (AG)_{il} + (DG)_{jl} + \varepsilon_{ijklm}$$

- **Mixed effect model**
  - Array is random
Optimal Allocation of Genes

\(K\) spots on each array (fixed)
\(g\) target genes (fixed)
\(h\) housekeeping genes (fixed)
\(n\) replications of each target gene
\(m\) replications of each housekeeping gene

A-optimality

Minimize average variance of differential estimates

\[ m / n = g^{1/2} / h \]
### 2×a Generalized Youden Design

<table>
<thead>
<tr>
<th>Dye\Array</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>$T_1$</td>
<td>$T_2$</td>
<td>$T_1$</td>
<td>$T_3$</td>
<td>$T_2$</td>
<td>$T_3$</td>
</tr>
<tr>
<td>Red</td>
<td>$T_2$</td>
<td>$T_1$</td>
<td>$T_3$</td>
<td>$T_1$</td>
<td>$T_3$</td>
<td>$T_2$</td>
</tr>
</tbody>
</table>

- **Array** - Balanced Incomplete Block Design
- **Dye** - Complete Block Design
- Each gene appears equal number of times on each array
Summary

- Model Approach
- Simultaneous Confidence Intervals (FWER)
- Housekeeping genes for control
- $2 \times a$ Generalized Latin Square Design ($t=2$)
- A- Optimality for gene allocation
- $2 \times a$ Generalized Youden Design ($t>2$)