Selection of Good Two–Color Microarray Designs using Genetic Algorithms

Mahbub Latif
Queen Mary, University of London

(E Brunner (Goettingen) and F Bretz (Novartis))

DEMA, August 12, 2008
Outline

- Introduction
- Statistical Model
- Design Selection Criteria: Optimality and Robustness Criteria
- Search for Good Microarray Designs
- Summary
Introduction

Allocations of competing RNA samples to the arrays

- which pair of treatments are hybridized in the same array?
- how RNA samples should be labelled?
- how many replications of each type of array?
Introduction

Allocations of competing RNA samples to the arrays
- which pair of treatments are hybridized in the same array?
- how RNA samples should be labelled?
- how many replications of each type of array?

Good Microarray Design
- efficient estimates of the effects of interest
- robust against missing arrays
## Introduction

### Allocations of competing RNA samples to the arrays
- which pair of treatments are hybridized in the same array?
- how RNA samples should be labelled?
- how many replications of each type of array?

## Good Microarray Design
- efficient estimates of the effects of interest
- robust against missing arrays

## Before conducting the experiment, it is known
- number of completing samples, $K$
- number of arrays, $n$
- research questions of interest
Statistical Model

Gene–specific model for the normalized log–intensity $Y_{ij(k)}$

$$Y_{ij(k)} = \mu + A_i + \delta_j + \tau_k + \eta_{ij}$$
Statistical Model

Gene–specific model for the normalized log–intensity $Y_{ij(k)}$

$$Y_{ij(k)} = \mu + A_i + \delta_j + \tau_k + \eta_{ij}$$

dye \quad treatment \quad array \quad error

Difference of log–intensities

Treatments $k$ and $k'$ are hybridized on array $i$ and $k \ (k')$ is labelled with dye $j \ (j')$

$$Z_i = Y_{ijk} - Y_{ij'k'}$$

$$= \delta_g - \delta_r + \tau_k - \tau_{k'} + \epsilon_i$$

$E(\epsilon_i) = 0$ and $V(\epsilon_i) = \sigma^2$. 
In matrix notation

\[ Z = X\beta + \epsilon \]  

\[ Z = (Z_1, \ldots, Z_n)', \beta = (\delta_g, \delta_r, \tau_1, \ldots, \tau_K), \text{ and } X \text{ is a design matrix} \]
Statistical Model

In matrix notation

\[ Z = X\beta + \epsilon \] (1)

\( Z = (Z_1, \ldots, Z_n)' \), \( \beta = (\delta_g, \delta_r, \tau_1, \ldots, \tau_K) \), and \( X \) is a design matrix.

Design Matrix: an example

\[
X = \begin{pmatrix}
\delta_g & \delta_r & \tau_1 & \tau_2 & \tau_3 & \tau_4 \\
1 & -1 & 1 & -1 & 0 & 0 \\
1 & -1 & 0 & 1 & -1 & 0 \\
1 & -1 & 0 & 0 & 1 & -1 \\
1 & -1 & -1 & 0 & 0 & 1
\end{pmatrix}
\]

Each arrow represents an array; arrow head (tail) is labelled with red (green) dye.
Optimality Criteria

Linear combinations of $\beta \rightarrow$ effect of interest, e.g.,

- Pair–wise treatment comparison

\[ C'\beta = (0, 0, 1, -1, 0, 0)\beta = \tau_1 - \tau_2 \]

compares treatments 1 and 2.

- Main effects, interaction
Optimality Criteria

The effect $\mathbf{C}'\beta$ is estimable if and only if

$$\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}(\mathbf{X}'\mathbf{X}) = \mathbf{C}'$$ (2)

If $\mathbf{C}'\beta$ is estimable

- $\mathbf{C}'\hat{\beta} = \mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Z}$ is the BLUE of $\mathbf{C}'\beta$
- $\text{Var}(\mathbf{C}'\hat{\beta}) = \mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C} \sigma^2$

Variance factor

$\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}$ depends only on research questions and design
## Optimality Criteria

### The $E$–optimality criterion

$$
\Psi(X_n, C'\beta) = \frac{tr(C'C)}{\lambda_{\text{max}}\left\{C'(X_n'X_n)^{-1}C\right\}}
$$

(3)

\[
\lambda_{\text{max}}(V) : \text{largest eigenvalue of } V, \ X_n \text{ design matrix of } \xi_n
\]

### Large value of $\Psi$ indicates more optimum design

$\Psi(X_{n,1}, C'\beta) > \Psi(X_{n,2}, C'\beta)$ indicates the design $\xi_{n,1}$ is more efficient than $\xi_{n,2}$ for estimating $C'\beta$

### For more than one research questions, $C_1'\beta, \ldots, C_Q'\beta$

1. $\frac{1}{Q} \sum_{q=1}^{Q} \Psi(X_n, C'_q\beta)$
2. $\sum_{q=1}^{Q} w_q \Psi(X_n, C'_q\beta) / \sum_q w_q$
Robustness Criteria

- Residual Designs → Possible designs that can be constructed with a specific number of missing arrays

\[ R_m(\mathbf{X}_n) = \{ \mathbf{X}_{n,1}^{(-m)}, \ldots, \mathbf{X}_{n,w}^{(-m)} \}, \ w = \binom{n}{m} \]

- \( \mathbf{X}_{n,t}^{(-m)} \) → design matrix corresponding to \( t^{\text{th}} \) residual design

- \( R^*_m(\mathbf{X}_n, \mathbf{C}'\beta) \subseteq R_m(\mathbf{X}_n) \)

\[ \mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}(\mathbf{X}'\mathbf{X}) = \mathbf{C}', \ \forall \ \mathbf{X} \in R^*_m(\mathbf{X}_n, \mathbf{C}'\beta) \] (4)
Robustness Criteria

Breakdown Number (BdN)

The minimum number of missing observations that lead to at least one residual design for which the effect of interest is not estimable anymore.

\[
BdN(X_n, C'\beta) = \min_m \left\{ R^*_m(X_n, C'\beta) \subset R_m(X_n) \right\}
\]
Robustness Criteria

**Breakdown Number (BdN)**

The minimum number of missing observations that lead to at least one residual design for which the effect of interest is not estimable anymore.

\[
BdN(X_n, C'\beta) = \min_m \left\{ \mathcal{R}_{m}^{*}(X_n, C'\beta) \subset \mathcal{R}_{m}(X_n) \right\}
\]

**Average Optimality Criterion**

For \( m \) missing arrays,

\[
\Psi_m(X_n, C'\beta) = \begin{cases} 
\sum_{x \in \mathcal{R}_{m}(X_n)} \Psi(X, C'\beta), & \text{if } BdN(X_n, C'\beta) < m \\
\infty, & \text{otherwise.}
\end{cases}
\]
Example: Optimality and Robustness Criteria

- Competing designs for comparing treatments, 1, 2, 3, 4

\[ \mathbf{C}' \mathbf{\beta} = \tau_1 - \tau_2 \]

<table>
<thead>
<tr>
<th>Design</th>
<th>$\Psi_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_1$</td>
<td>2.00</td>
</tr>
<tr>
<td>$D_2$</td>
<td>5.33</td>
</tr>
<tr>
<td>$D_3$</td>
<td>5.33</td>
</tr>
</tbody>
</table>
Example: Optimality and Robustness Criteria

- Competing designs for comparing treatments, 1, 2, 3, 4

![Diagram of competing designs](image)

- Effect of interest: $\mathbf{C}'\beta = \tau_1 - \tau_2$

<table>
<thead>
<tr>
<th>Design</th>
<th>BdN</th>
<th>$\Psi_0$</th>
<th>$\Psi_1$</th>
<th>$\Psi_2$</th>
<th>$\Psi_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_1$</td>
<td>2</td>
<td>2.00</td>
<td>1.67</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$D_2$</td>
<td>2</td>
<td>5.33</td>
<td>4.45</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$D_3$</td>
<td>4</td>
<td>5.33</td>
<td>4.45</td>
<td>3.58</td>
<td>2.67</td>
</tr>
</tbody>
</table>
Search for Good Designs

Interest is in comparing $K$ treatments and the effects of interest are $C_1', \ldots, C_Q'$.

- $N = 2 \cdot {K \choose 2} \rightarrow$ the number of possible arrays
- $T = {N+n-1 \choose n} \rightarrow$ the number of possible designs with $n$ arrays out of $N$ arrays
- $\mathcal{H} \rightarrow$ set of design matrices corresponding to $T$ designs
Search for Good Designs

Interest is in comparing $K$ treatments and the effects of interest are $\mathbf{C}_1'\beta, \ldots, \mathbf{C}_Q'\beta$

$$N = 2 \cdot \binom{K}{2} \rightarrow \text{the number of possible arrays}$$

$$T = \binom{N+n-1}{n} \rightarrow \text{the number of possible designs with } n \text{ arrays out of } N \text{ arrays}$$

$\mathcal{H} \rightarrow \text{set of design matrices corresponding to } T \text{ designs}$

Optimization Problem

$$\max_{\mathbf{X} \in \mathcal{H}} \frac{\sum_{q=1}^{Q} \Psi(\mathbf{X}, \mathbf{C}_q'\beta)}{Q}$$

subject to $\mathbf{C}_q'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}) = \mathbf{C}_q', \ \forall \ q$
Search for Good Designs

Interest is in comparing $K$ treatments and the effects of interest are $C'_1\beta, \ldots, C'_Q\beta$

$$N = 2 \cdot \binom{K}{2} \rightarrow \text{the number of possible arrays}$$

$$T = \binom{N+n-1}{n} \rightarrow \text{the number of possible designs with } n \text{ arrays out of } N \text{ arrays}$$

$\mathcal{H} \rightarrow \text{set of design matrices corresponding to } T \text{ designs}$

Optimization Problem

$$\max_{X \in \mathcal{H}} \sum_{q=1}^{Q} \frac{\psi(X, C'_q\beta)}{Q}$$

subject to

$$C'_q(X'X)^{-1}(X'X) = C'_q, \forall q$$

- Complete search is computationally infeasible
- Search space is discrete
Search for Good Designs

- GA mimics some important features of natural evolution to find at least near-optimal solutions of the problem
  - Important features → Selection, Inheritance, Variability, etc.

- Data structure
  - Chromosome: A solution of the search space
  - Gene: Element of Chromosome

  Quality of a Chromosome is assessed by Fitness function

- Microarray Experiments → Genetic Algorithm
  - Design → Chromosome
  - Array → Gene
  - $E$-optimality criterion → Fitness
Search for Good Designs

- Pseudo Code for Genetic Algorithm
  create initial population
  evaluate fitness
  repeat
    selection
    inheritance (crossover)
    variability (mutation)
  evaluate fitness
  until terminating condition

- Flowchart of Genetic Algorithm
Genetic Algorithms

A typical run of a GA

![Graph showing the average fitness over generations.](image-url)
Example: $3 \times 2$ Factorial Experiment

- Two–Factor factorial experiment
- Factor A has three levels and B has two levels
  - Total number of treatments, $K = 3 \cdot 2 = 6$
  - Number of possible arrays, $N = 2 \cdot \binom{6}{2} = 30$
- Different combinations of main effects and interaction could be the effects of interest
Example: $3 \times 2$ Factorial Experiment

\[ \begin{array}{ccc}
11 & 12 & 21 \\
21 & 31 & 32 \\
31 & 32 & 22 \\
\end{array} \]

\[ \begin{array}{ccc}
D_6 & &  \\
& D_{15} & \\
& & D_{18} \\
\end{array} \]

\[
\begin{array}{c|ccc|}
\text{Design} & A & B & A \times B \\
\hline
D_6 & 6.00 & 4.00 & 2.00 \\
D_{15} & 12.00 & 5.60 & 12.00 \\
D_{18} & 12.00 & 8.00 & 16.00 \\
\end{array}
\]
Example: $3 \times 2$ Factorial Experiment

\[ D_{12} \]

\[ D_{18} \]

<table>
<thead>
<tr>
<th>Effect</th>
<th>( D_{12} )</th>
<th>( D_{18} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \psi_0 )</td>
<td>( \psi_1 )</td>
</tr>
<tr>
<td>( A )</td>
<td>12.0</td>
<td>9.0</td>
</tr>
<tr>
<td>( B )</td>
<td>4.0</td>
<td>3.4</td>
</tr>
<tr>
<td>( A \times B )</td>
<td>8.0</td>
<td>6.1</td>
</tr>
</tbody>
</table>
Simulation Study

- A–optimum designs for $K$, $K + 2$, and $2K$ arrays are known (all pairwise comparisons as the effects of interest)

- Followings are the results of 500 simulations

<table>
<thead>
<tr>
<th>$K$</th>
<th>$n$</th>
<th>min</th>
<th>mean</th>
<th>max</th>
<th>sd</th>
<th>Rel. Eff.</th>
<th>Generations</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>6</td>
<td>1.643</td>
<td>1.810</td>
<td>1.827</td>
<td>0.034</td>
<td>99.06</td>
<td>73</td>
</tr>
<tr>
<td>8</td>
<td>2.681</td>
<td>2.743</td>
<td>2.773</td>
<td>0.024</td>
<td>98.91</td>
<td>114</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>4.549</td>
<td>4.619</td>
<td>4.640</td>
<td>0.039</td>
<td>99.54</td>
<td>116</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.402</td>
<td>1.464</td>
<td>1.482</td>
<td>0.013</td>
<td>98.77</td>
<td>133</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2.139</td>
<td>2.204</td>
<td>2.213</td>
<td>0.012</td>
<td>99.60</td>
<td>152</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>4.096</td>
<td>4.211</td>
<td>4.327</td>
<td>0.038</td>
<td>97.31</td>
<td>290</td>
<td></td>
</tr>
</tbody>
</table>
Two criteria are introduced for quantifying robustness characteristics of a microarray design.

GA–based search procedures perform well to find near optimal microarray designs.

Other design criteria (e.g., D–optimality criteria) can be used in the search procedure.