Background Correction and Normalization

Lecture 3
Computational and Statistical Aspects of Microarray Analysis
June 21, 2005 Bressanone, Italy
Feature Level Data
Outline

• Affymetrix GeneChip arrays
• Two color platforms
Affymetrix GeneChip Design

Reference sequence

...TGTGATGATGATGGGTACAGAAGGGCTCCCGATGCGCCGATTGAGAAT...

Perfect match

Mismatch

GTACTACCCAGTCTTCCCGGAGGCTA
GTACTACCCAGTGTCCCGGAGGCTA

5' 3'

NSB & SB

NSB
Before Hybridization

Sample 1

Sample 2

Array 1

Array 2
More Realistic

Array 1

Sample 1

Array 2

Sample 2
Non-specific Hybridization

Array 1

Array 2
Affymetrix GeneChip Design

Reference sequence:

...TGTGATGGGTGCATGATGGGTCAGAAGGCCTCCGATGCGCCGATTGAGAAT...

Perfectmatch (PM):

GTACTACCCAGTCTTCCCGGAGGCTA

Mismatch (MM):

GTACTACCCAGTGTTCGGGAGGCTA

NSB & SB

NSB
GeneChip Feature Level Data

- MM features used to measure optical noise and non-specific binding directly
- More than 10,000 probesets
- Each probeset represented by 11-20 feature

Note 1: Position of features are haphazardly distributed about the array.
Note 2: There are between 20-100 chip types

- So we have $PM_{gij}$, $MM_{gij}$
  ($g$ is gene, $i$ is array and $j$ is feature)

- A default summary is the avg of the PM-MM
Two color platforms

- Common to have just one feature per gene
- Typically, longer molecules are used so non-specific binding not so much of a worry
- Optical noise still a concern
- After spots are identified, a measure of local background is obtained from area around spot
Local background

---- GenePix
---- QuantArray
---- ScanAnalyze
Two color feature level data

- Red and Green foreground and background obtained from each feature
- We have $R_{f_{gij}}, G_{f_{gij}}, R_{b_{gij}}, G_{b_{gij}}$ ($g$ is gene, $i$ is array and $j$ is replicate)
- A default summary statistic is the log-ratio: 
  \[
  \frac{(R_f - R_b)}{(G_f - G_b)}
  \]
Affymetrix Spike In Experiment
Spike-in Experiment

- Throughout we will be using Data from Affymetrix’s spike-in experiment
- Replicate RNA was hybridized to various arrays
- Some probesets were spiked in at different concentrations across the different arrays
- This gives us a way to assess precision and accuracy
- Done for HGU95 and HGU133 chips
### Spikein Experiment (HG-U95)

#### Probeset

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>0</td>
<td>512</td>
<td>1024</td>
<td>256</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>0</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>512</td>
<td>64</td>
</tr>
<tr>
<td>C</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>0.5</td>
<td>0</td>
<td>0.25</td>
<td>128</td>
<td>256</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>1</td>
<td>0.25</td>
<td>0.5</td>
<td>0</td>
<td>256</td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>2</td>
<td>0.5</td>
<td>1</td>
<td>0.25</td>
<td>512</td>
</tr>
<tr>
<td>F</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>0.5</td>
<td>1024</td>
</tr>
<tr>
<td>G</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>16</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>0.25</td>
</tr>
<tr>
<td>I</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>32</td>
<td>8</td>
<td>16</td>
<td>4</td>
<td>0.5</td>
</tr>
<tr>
<td>J</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>64</td>
<td>16</td>
<td>32</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>K</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>128</td>
<td>32</td>
<td>64</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>L</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>256</td>
<td>64</td>
<td>128</td>
<td>32</td>
<td>4</td>
</tr>
</tbody>
</table>

---

**Array**

- **A**
- **B**
- **C**
- **D**
- **E**
- **F**
- **G**
- **H**
- **I**
- **J**
- **K**
- **L**
- **M**
- **N**
- **O**
- **P**
- **Q**
- **R**
- **S**
- **T**
Spikein Experiment (HG-U133)

• A similar experiment was repeated for a newer chip

• The 1024 picoMolar concentration was not used. 1/8 was used instead.

• No groups of 12

• Note: More spike-ins to come!
Introduction to Preprocessing
Composition of Feature Intensities

- Specific binding
- Non-specific binding and Optical noise
- Systematic variations
- Stochastic noise

- We will focus on Affymetrix GeneChip first
Three main steps

• Background adjustment of correction
• Normalization
• Summarization
Why is it difficult?

These are from replicate RNA. Left are empirical density estimates for 5 GeneChip arrays. Right is MA plot for one two-color array.
Approaches

• Stepwise: The most common approach is to solve these issues one by one and produce one measurement of expression (absolute or relative) for each gene
  – Typically the order is: correct for background, normalize, then summarize.

• Integrated: Statistical models that quantify the need for normalization as well as the amount of expression are used
Background Noise
Distribution of Optical and NSB Noise

Note: Material in this lecture are for Affymetrix Arrays
Spike-In Data

- Specific Binding: Linear with concentration

- Background Additive

\[ Y = B + aC \]
Effect of Background

\[
\frac{B + aC_1}{B + aC_2} \neq \frac{C_1}{C_2}
\]

\[
\log(B + aC_1) - \log(B + aC_2) \neq \log(C_1) - \log(C_2)
\]

\[
\log(B + aC) \approx \log(B)
\]

When \( a \) or \( C \) is small
Why background correct

a) Original scale

b) Log (base 2) scale
Different NSB for different features

![Graph showing different NSB for different features]
Direct Measurement Strategy

- Affymetrix’s tries to measure non-specific binding directly
- PM-MM is supposed to give us a good estimate of specific binding
- But it is not so simple
Sometimes MM larger then PM
Deterministic Model

\[ \text{PM} = O + N + S \]

\[ \text{MM} = O + N \]

\[ \text{PM} - \text{MM} = S \]

For reasons explained later, we want to take log of S
Deterministic model is wrong

- Do MM measure non-specific binding?
- Look at Yeast DNA hybridized to Human Chip
- Look at PM, MM log-scale scatter-plot
- $R^2$ is only 0.5
Stochastic Model
(Additive background/multiplicative error)

PM = \( O_{PM} + N_{PM} + S \),
MM = \( O_{MM} + N_{MM} \)

\[ \log(N_{PM}), \log(N_{MM}) \sim \text{Bivariate Normal} \ (\rho \approx 0.7) \]
S = \( \exp(s + a + \varepsilon) \)
s is the quantity of interest (log scale expression)

\[ E[PM - MM] = S \], but
\[ \text{Var}[\log(PM - MM)] \sim 1/S^2 \] (can be very large)

Note: Technically, var[ log(PM-MM) ] is not defined. We need to make sure we avoid taking logs of negative. Affymetrix’s current default does this.
Simulation

• We create some feature level data for two replicate arrays
• Then compute \( Y = \log(PM-kMM) \) for each array
• We make an MA using the \( Y \)s for each array
• We make a observed concentration verse known concentration plot
• We do this for various values of \( k \). The following “movie” shows \( k \) moving from 0 to 1.
Does this happen in practice?

MAS 5.0

RMA
RMA: The Basic Idea

$PM = B + S$

*Observed: PM*

*Of interest: S*

Pose a statistical model and use it to predict $S$ from the observed $PM$
The Basic Idea

\[ PM = B + S \]

- A mathematically convenient, useful model

\[ \hat{S} = E[S \mid PM] \]

- No MM
- Borrowing strength across probes
Normalization

- **Normalization** is needed to ensure that differences in intensities are indeed due to differential expression, and not some printing, hybridization, or scanning artifact.

- Normalization is necessary before any analysis which involves within or between slides comparisons of intensities, e.g., clustering, testing.

- Normalization is different in spotted/two-color and high-density-oligonucleotides (Affy) technologies.
Outline

• Why do we need to normalize
• Types of normalization
  – Affy
  – cDNA
• Case study
• Discussion
Technical Replicates

Different scanners were used! These is Affymetrix data
Empirical Densities for Replicates

Density of PM probe intensities for Spike-In chips
Self-self hybridization
Self-self hybridization

Two-color platform
Self-self hybridization

Robust local regression within sectors (print-tip-groups) of intensity log-ratio $M$ on average log-intensity $A$. 

$M$ vs. $A$
Boxplots by print-tip-group

Swirl 93 array: pre-normalization log-ratio $M$

Intensity log-ratio, $M$
What can we do?

• Throw away the data and start again? Maybe.
• Statistics offers hope:
  – Use control genes to adjust
  – Assume most genes are not differentially expressed
  – Assume distribution of expression are the same
Simplest Idea

• Assume all arrays have the same median log expression or relative log expression

• Subtract median from each array

• In two-color platforms, we typically correct the Ms. Median correction forces the median log ratio to be 0
  – Note: We assume there are as many over-expressed as under-expressed genes)

• For Affymetrix arrays we usually add a constant that takes us back to the original range.
  – It is common to use the median of the medians
  – Typically, we subtract in the log-scale
How does it work

Notice subtracting in the original scale won't work well
How does it work

The medians match but there are some discrepancies
What are the consequences

These are two technical replicates. Only the red are differentially expressed.
There appears to be a non-linear dependence on intensity

- Proposed solutions
  - Force distributions (not just medians) to be the same:
    - Amaratunga and Cabrera (2001)
    - Bolstad et al. (2003)
  - Use curve estimators such as splines to adjust for the effect:
    - Li and Wong (2001) Note: they also use a rank invariant set
    - Colantuoni et al (2002)
    - Dudoit et al (2002)
  - Use adjustments based on additive/multiplicative model:
    - Rocke and Durbin (2003)
Quantile normalization

- All these non-linear methods perform similarly
- Quantiles is my favorite because it's fast
- Basic idea:
  - order value in each array
  - take average across probes
  - Substitute probe intensity with average
  - Put in original order
Example of quantile normalization

<table>
<thead>
<tr>
<th>Original</th>
<th>Ordered</th>
<th>Averaged</th>
<th>Re-ordered</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 4 4</td>
<td>2 3 4</td>
<td>3 3 3</td>
<td>3 5 3</td>
</tr>
<tr>
<td>5 4 14</td>
<td>3 4 8</td>
<td>5 5 5</td>
<td>8 5 8</td>
</tr>
<tr>
<td>4 6 8</td>
<td>3 4 8</td>
<td>5 5 5</td>
<td>6 8 5</td>
</tr>
<tr>
<td>3 5 8</td>
<td>4 5 9</td>
<td>6 6 6</td>
<td>5 6 5</td>
</tr>
<tr>
<td>3 3 9</td>
<td>5 6 14</td>
<td>8 8 8</td>
<td>5 3 6</td>
</tr>
</tbody>
</table>
How does it work
How does it work

Does it wash away real differential expression?
These are two technical replicates. Only the red are differentially expressed.
After quantile normalization

log(PM) after normalization

We say more later.....
Two-color platforms

• Identify and remove the effects of systematic variation in the measured fluorescence intensities, other than differential expression, for example
  – different labelling efficiencies of the dyes;
  – different amounts of Cy3- and Cy5-labelled mRNA;
  – different scanning parameters;
  – print-tip, spatial, or plate effects, etc.
Normalization

• Much more complicated than Affy
• The need for normalization can be seen most clearly in self-self hybridizations, where the same mRNA sample is labeled with the Cy3 and Cy5 dyes.
• The imbalance in the red and green intensities is usually not constant across the spots within and between arrays, and can vary according to overall spot intensity, location, plate origin, etc.
• These factors should be considered in the normalization.
MA-plot by print-tip-group

The smooth curves were created using loess. Splines work just as well.

Intesity log ratio, M

Average log intensity, A
Example of Normalization

\[ \log_2 \frac{R}{G} \leftrightarrow \log_2 \frac{R}{G} - L(\text{intensity, sector, } \ldots) \]

- **Constant normalization:** \( L \) is constant
- **Adaptive normalization:** \( L \) depends on a number of predictor variables, such as spot intensity \( A \), sector, plate origin.
  - Intensity-dependent normalization.
  - Intensity and sector-dependent normalization.
  - 2D spatial normalization.
  - Other variables: *time of printing, plate, etc.*
  - Composite normalization. **Weighted average of several normalization functions.**
2D images of L values

- Global median normalization
- Global loess normalization
- Within-print-tip-group loess normalization
- 2D spatial normalization
2D images of normalized M-L

- Global median normalization
- Global loess normalization
- Within-print-tip group loess normalization
- 2D spatial normalization
Boxplots of normalized M-L
MA-plots of normalized M-L

Global median normalization

Global loess normalization

Within-print-tip-group loess normalization

2D spatial normalization
Comparison to other methods

MSP
Rank invariant
Housekeeping
Tubulin, GAPDH
What if we know most genes are up and down regulated?

- Then it will be hard to normalize
- Let us look at an example
Dilution Experiment

- cRNA hybridized to human chip (HGU95) in range of proportions and dilutions
- Dilution series begins at 1.25 µg cRNA per GeneChip array, and rises through 2.5, 5.0, 7.5, 10.0, to 20.0 µg per array. 5 replicate chips were used at each dilution
- A few control genes were spiked in at same concentration across all arrays.
- Notice we can only make usual assumption within groups of 5
Raw Data
If you can’t normalized your conclusions will likely be wrong!
Normalize with control genes

Looks almost as bad!
We can not median normalize

We wash out real biological effects
Or quantile normalize
But we can quantile normalize and then use controls
Look at the improvement
Look at the improvement
<table>
<thead>
<tr>
<th>Method</th>
<th>Med slope</th>
<th>Log SD</th>
<th>Std % change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>0.39</td>
<td>0.35</td>
<td>27%</td>
</tr>
<tr>
<td>Controls</td>
<td>0.52</td>
<td>0.25</td>
<td>19%</td>
</tr>
<tr>
<td>Median</td>
<td>-0.01</td>
<td>0.17</td>
<td>12%</td>
</tr>
<tr>
<td>Quantiles</td>
<td>0.00</td>
<td>0.14</td>
<td>10%</td>
</tr>
<tr>
<td>Quantiles/ctrls</td>
<td>0.55</td>
<td>0.14</td>
<td>10%</td>
</tr>
</tbody>
</table>
Why the non-linear dependence?

• The additive background noise plus multiplicative error models predicts these non-linear behavior.
• The multiplicative part explains the difference in over-all mean log-intensity
• Different background noise mean levels causes non-linear MA plots
Add a different constant
Add noise with different mean
Conclusions

- I still haven’t seen microarray data that doesn’t need to be normalized
- Non-linear methods are better than linear ones
- If you can’t assume most genes are not differentially expressed or that probe intensities have roughly the same distribution you need many control genes (hundreds maybe thousands)
- Control genes must cover the dynamic range