Methodologies for Pre-processing Microarray Data

Ben Bolstad

bmb@bmbolstad.com
http://bmbolstad.com

Mar 31, 2006
Outline

• Introduction
• Affymetrix GeneChips (a popular single channel array)
• Pre-processing using the RMA framework and extensions
Workflow for a typical microarray experiment

1. **Biological Question**
2. **Experimental Design**
3. **Microarray Experiment**
   - Images
4. **Image Quantification**
5. **Background Adjustment**
6. **Normalization**
7. **Summarization**
8. **Quality Assessment**

**Pre-processing Low-level analysis**

**Expression Values**

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Array 1</th>
<th>Array 2</th>
<th>Array 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10.05</td>
<td>9.58</td>
<td>9.76</td>
</tr>
<tr>
<td>Gene 2</td>
<td>4.12</td>
<td>4.16</td>
<td>4.05</td>
</tr>
<tr>
<td>Gene 3</td>
<td>6.05</td>
<td>6.04</td>
<td>6.08</td>
</tr>
</tbody>
</table>

**High-level analysis**

- Estimation
- Testing
- Annotation
- Clustering
- Discrimination

**Biological verification and interpretation**
Introduction to preprocessing

- Pre-processing typically constitutes the initial (and possibly most important) step in the analysis of data from any microarray experiment
- Often ignored or treated like a black box (but it shouldn’t be)
- Consists of:
  - Data exploration
  - Background correction, normalization, summarization
  - Quality Assessment
- These are interlinked steps
Background Correction/Signal Adjustment

• A method which does some or all of the following:
  ▪ Corrects for background noise, processing effects on the array
  ▪ Adjusts for cross hybridization (non-specific binding)
  ▪ Adjust estimated expression values to fall across an appropriate range
Normalization

“Non-biological factors can contribute to the variability of data ... In order to reliably compare data from multiple probe arrays, differences of non-biological origin must be minimized.”1

• Normalization is the process of reducing unwanted variation either within or between arrays. It may use information from multiple chips.

• Typical assumptions of most major normalization methods are (one or both of the following):
  ▪ Only a minority of genes are expected to be differentially expressed between conditions
  ▪ Any differential expression is as likely to be up-regulation as down-regulation (ie about as many genes going up in expression as are going down between conditions)

1 GeneChip 3.1 Expression Analysis Algorithm Tutorial, Affymetrix technical support
Summarization

• Reducing multiple measurements on the same gene down to a single measurement by combining in some manner.
Quality Assessment

• Need to be able to differentiate between good and bad data.

• Bad data could be caused by poor hybridization, artifacts on the arrays, inconsistent sample handling, ..... 

• An admirable goal would be to reduce systematic differences with data analysis techniques.

• Sometimes there is no option but to completely discard an array from further analysis. How to decide .....
Affymetrix GeneChip

- Commercial mass produced high density oligonucleotide array technology developed by Affymetrix [http://www.affymetrix.com](http://www.affymetrix.com)
- Single channel microarray
- Today's talk relates to arrays designed for expression analysis

Image courtesy of Affymetrix Press Website.
Probes and Probesets

Typically 11 probe(pairs) in a probeset
Latest GeneChips have as many as:
  54,000 probesets
  1.3 Million probes
Two Probe Types

Reference Sequence

TAGGTCTGTATGACAGACACACAAAAGAAAGATG

CAGACATAGTGTG\textcolor{red}{C}TGTGTTTTCTTCTTCT

CAGACATAGTGTG\textcolor{red}{G}TGTGTTTTCTTCTTCT

PM: the Perfect Match

MM: the Mismatch

Note that about 30% of MM probe intensities are brighter than corresponding PM probe intensities.
Hybridization to the Chip

Sample of Fragmented Labeled RNA

Labeling molecule that fluoresces

Before Hybridization

After Hybridization
The Chip is Scanned

Shining a laser light at GeneChip causes tagged DNA fragments that hybridized to glow
Image Analysis
Boxplot raw intensities
Density plots
Comparing arrays

Comparing 2 arrays

Array2 vs Array 1

Bad
Comparing arrays

Comparing 2 arrays

Log2 Array2 vs Log2 Array 1

Better
Comparing arrays

$M = \log_2(\text{Array2}/\text{Array1})$

Vs

$A = \frac{1}{2} \log_2(\text{Array2} \times \text{Array1})$

Best

$M = \text{Minus}$

$A = \text{Average}$
Typical MA-plot

Comparing 2 arrays

Loess smoother
Pairwise MA plots

\[ M = \log_2 \frac{\text{array}_i}{\text{array}_j} \]
\[ A = \frac{1}{2} \log_2 (\text{array}_i \times \text{array}_j) \]
Boxplots comparing M
RMA Background Approach

- Convolution Model

\[ E(S \mid PM = pm) = a + b \frac{\phi\left(\frac{a}{b}\right) - \phi\left(\frac{pm - a}{b}\right)}{\Phi\left(\frac{a}{b}\right) + \Phi\left(\frac{pm - a}{b}\right) - 1} \]
GCRMA Background Approach

- \( PM = O_{pm} + N_{pm} + S \)
- \( MM = O_{mm} + N_{mm} \)

- \( O \) – Optical noise
- \( N \) – non-specific binding
- \( S \) – Signal

- Assume \( O \) is distributed Normal
- \( \log(N_{pm}) \) and \( \log(N_{mm}) \) are assumed bi-variate normal with correlation 0.7
- \( \log(S) \) assumed exponential(1)
GCRMA continued

• An experiment was carried out where yeast RNA was hybridized to human chips, so all binding expected to be non specific.
• Fitted a model to predict log intensity from sequence composition gives base and position effects

![Graph showing sequence effects](image)

• Uses these effects to predict an affinity for any given sequence call this $A$. The means of the distributions for the $N_{pm}$, $N_{mm}$ terms are functions of the affinities.
Non-Biological variability is a problem for single channel arrays.
Normalization

• In case of single channel microarray data this is carried out only across arrays.

• Could generalize methods we applied to two color arrays, but several problems:
  ▪ Typically several orders of magnitude more probes on an Affymetrix array then spots on a two channel array
  ▪ With single channel arrays we are dealing with absolute intensities rather than relative intensities.

• Need something fast
Quantile Normalization

- Normalize so that the quantiles of each chip are equal. Simple and fast algorithm. Goal is to give same distribution to each chip.
Density of PM probe intensities for Spike-In chips

$\log(\text{PM})$
It works!!

Unnormalized

Scaling

Quantile Normalization
It Reduces Variability

Expression Values

Fold change for Non differential genes

Unnormalized Quantile Scaling

Variance Unnormalized

Log2 Fold Change

Also no serious bias effects. For more see Bolstad et al (2003).
Problem: Calculating gene expression values.

How do we reduce the 11-20 probe intensities for each probeset on to a gene expression value?

Our Approach

- RMA – a robust multi-chip linear model fit on the log scale

Some Other Approaches

- Single chip
  - AvDiff (Affymetrix) – no longer recommended for use due to many flaws
  - Mas 5.0 (Affymetrix) – use a 1 step Tukey-biweight to combine the probe intensities in log scale

- Multiple Chip
  - MBEI (Li-Wong dChip) – a multiplicative model on natural scale
General Probe Level Model

\[ y_{kij} = f(X) + \epsilon_{kij} \]

- Where \( f(X) \) is function of factor (and possibly covariate) variables (our interest will be in linear functions)

- \( y_{kij} \) is a pre-processed probe intensity (usually log scale)

- Assume that \( \text{Var}[\epsilon_{kij}] = \sigma_k^2 \)
Parallel Behavior Suggests Multi-chip Model

Differentially expressing

Non Differential
Probe Pattern Suggests Including Probe-Effect

Differentially expressing

Non Differential

![Graphs showing PM probe intensity over probe number for differentially expressing and non-differential cases.](image)
Also Want Robustness

Differentially expressing

Non Differential

Differentially expressing

Non Differential
The RMA model

\[ y_{kij} = m_k + \alpha_{ki} + \beta_{kj} + \varepsilon_{kij} \]

where

- \( y_{kij} = \log_2 N(B(PM_{kij})) \)
- \( \alpha_{ki} \) is a probe-effect \( i=1,\ldots,I \)
- \( \beta_{kj} \) is chip-effect \( (m_k + \beta_{kj} \) is log2 gene expression on array \( j \) ) \( j=1,\ldots,J \)
- \( k=1,\ldots,K \) is the number of probesets
Median Polish Algorithm

\[
\begin{array}{ccc|c}
 y_{11} & \cdots & y_{1J} & 0 \\
 \vdots & \ddots & \vdots & \vdots \\
 y_{I1} & \cdots & y_{IJ} & 0 \\
 \hline
 0 & \cdots & 0 & 0
\end{array}
\]

Sweep Rows

Sweep Columns

Iterate

\[
\begin{align*}
\hat{\varepsilon}_{11} & \quad \cdots \quad \hat{\varepsilon}_{1J} & \hat{\alpha}_1 \\
\vdots & \ddots & \vdots \\
\hat{\varepsilon}_{I1} & \cdots & \hat{\varepsilon}_{IJ} & \hat{\alpha}_I \\
\hat{\beta}_1 & \cdots & \hat{\beta}_J & \hat{m}
\end{align*}
\]

\[
\text{median } \alpha_i = \text{median } \beta_j = 0
\]
RMA mostly does well in practice

Detecting Differential Expression
Not noisy in low intensities

A

Fold change (Affymetrix)

RMA

MAS 5.0
One Drawback

RMA

MAS 5.0

Linearity across concentration. GCRMA fixes this problem.
GCRMA improve linearity
An Alternative Method for Fitting a PLM

- Robust regression using M-estimation
- In this talk, we will use Huber’s influence function. The software handles many more.
- Fitting algorithm is IRLS with weights dependent on current residuals $\psi(r_{kij})$
Variance Covariance Estimates

- Suppose model is $Y = X\beta + \varepsilon$
- Huber (1981) gives three forms for estimating variance covariance matrix

\[
\kappa^2 \frac{1}{(n-p)\sum_i \psi(r_i)^2} \left[ \frac{1}{n} \sum_i \psi'(r_i) \right]^2 \left( X^T X \right)^{-1}
\]

\[
\kappa \frac{1}{n\sum_i \psi'(r_i)} W^{-1}
\]

\[
\frac{1}{\kappa} \frac{1}{(n-p)\sum_i \psi(r_i)^2} W^{-1} \left( X^T X \right) W^{-1}
\]

We will use this form
We Will Focus on the Summarization PLM

- Array effect model

$$y_{kij} = \alpha_{ki} + \beta_{kj} + \epsilon_{kij}$$

Pre-processed Log PM intensity

With constraint

$$\sum_{i=1}^{I} \alpha_{ki} = 0$$
Quality Assessment

• Problem: Judge quality of chip data

• Question: Can we do this with the output of the Probe Level Modeling procedures?

• Answer: Yes. Use weights, residuals, standard errors and expression values.
Chip pseudo-images
An Image Gallery

“Ring of Fire”

“Crop Circles”

“Tricolor”

http://PLMImageGallery.bmbolstad.com
NUSE Plots

Normalized
 Unscaled
 Standard
 Errors

\[ NUSE(\beta_{kj}) = \frac{SE(\beta_{kj})}{\text{med}_j(SE(\beta_{kj}))} \]
RLE Plots

Relative Log Expression

\[ RLE(\beta_{kj}) = \beta_{kj} - \text{med}(\beta_{kj}) \]
Acknowledgements

• Terry Speed
• Rafael Irizarry
• Julia Brettschneider
• Francois Colin
• Zhijin (Jean) Wu

• Any one else …
References