Probe-level analysis of Affymetrix GeneChip Microarray Data using BioConductor

Ben Bolstad

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

May 22, 2006
Outline

• Introduction to probe-level analysis
• Probe-level analysis using the RMA framework and extensions
• Case study using BioConductor tools
Workflow for a typical microarray experiment

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

**Pre-processing**
- Low-level analysis
  - Quality Assessment

**Expression Values**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Array 1</th>
<th>Array 2</th>
<th>Array 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.05</td>
<td>9.58</td>
<td>9.76</td>
</tr>
<tr>
<td>2</td>
<td>4.12</td>
<td>4.16</td>
<td>4.05</td>
</tr>
<tr>
<td>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 Probe-Level Analysis

- Also known as “Pre-processing” or “low-level analysis”
- 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
Non-Biological variability is a problem for single channel arrays.

Liver tissue Unnormalized PM

Log2 PM intensity

5 scanners for 6 dilution groups
Summarization

• Reducing multiple measurements on the same gene down to a single measurement by combining in some manner. ie take each of the multiple probe intensities for a probeset and derive a single number representing probeset expression value.
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 …..
What's RMA?

- **Robust Multi-array Analysis**
  - Background correction using a convolution model (GCRMA modifies this stage)
  - Quantile Normalization across arrays
  - Multi-array probe-level model fit to each probeset
  - Quality assessment
RMA Background Approach

- Convolution Model

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

• 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.

• 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.
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.

\[ x = F^{-1}(G(x)) \]
Density of PM probe intensities for Spike-In chips

- After Quantile Normalization
Summarization

• Need to take the normalized background corrected probe intensities and reduce to sensible gene expression measures.
• RMA uses a multi-array model fit to logarithmic scale data.
Parallel Behavior Suggests Multi-chip Model

Differentially expressing

Non Differential

PM probe intensity vs. Array
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

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

\[ y_{11} \ldots y_{1J} \mid 0 \]
\[ \vdots \ldots \vdots \mid \vdots \]
\[ y_{I1} \ldots y_{IJ} \mid 0 \]
\[ 0 \ldots 0 \mid 0 \]

Sweep Rows

Sweep Columns

Iterate

median \( \alpha_i = median \beta_j = 0 \)

\[ \hat{\epsilon}_{11} \ldots \hat{\epsilon}_{1J} \mid \hat{\alpha}_1 \]
\[ \vdots \ldots \vdots \mid \vdots \]
\[ \hat{\epsilon}_{I1} \ldots \hat{\epsilon}_{IJ} \mid \hat{\alpha}_I \]
\[ \hat{\beta}_1 \ldots \hat{\beta}_J \mid \hat{m} \]
RMA mostly does well in practice

Detecting Differential Expression Not noisy in low intensities

![Graph showing Fold change (Affymetrix) vs False Positives for RMA and MAS 5.0]

- RMA
- MAS 5.0
One Drawback

RMA

MAS 5.0

Linearity across concentration. GCRMA fixes this problem
GCRMA improve linearity
• See affycomp for more comparisons between RMA, GCRMA, MAS5 and many other expression measures.

• http://affycomp.biostat.jhsph.edu/
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 $\psi(r_{kij})$
We Will Focus on the Summarization PLM

- Array effect model

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

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

\[ NUSE(\hat{\beta}_{kj}) = \frac{SE(\hat{\beta}_{kj})}{med_j(SE(\hat{\beta}_{kj}))} \]
RLE Plots

Relative Log Expression

\[ RLE(\hat{\beta}_{kj}) = \hat{\beta}_{kj} - med_j(\hat{\beta}_{kj}) \]
• Based on the R language
• Approx 160 packages (at 1.8 Release Apr 2006)
• Microarray data is a major focus, but also currently some software for dealing with Mass Spec data, Cell Based Assays (Flow Cytometry), with others application areas areas planned and expected.

• http://www.bioconductor.org
Installing BioConductor

source("http://www.bioconductor.org/biocLite.R")
biocLite()

- Installs a small (approx 20) subset of the packages
- Additional packages can be installed

biocLite(c("simpleaffy","makecdfenv"))

- This handles all the (inter) dependencies between the different packages
Dealing with Affymetrix Data

- **affy** – Data structures for storing probe intensity data. Supplies RMA, general functionality for combining different background, normalization, summarization schemes. Basic methods for examining probe intensity data.
- **affyPLM** – Methods for fitting probe level models.
- **gcrma** – provides the GCRMA expression measure and background correction
- **simpleaffy** – provides Affymetrix standard QC
Affymetrix Meta-data Packages

- *cdfenv packages* – contain processed CDF information
- *Probe packages* – contain probe sequence information
- *Annotation packages* – contain annotation information created using public data repositories

- eg for u133A chips these would be
  - `hgu133acdf`
  - `hgu133aprobe`
  - `hgu133a`

- Automatically downloaded and installed on first use.
Case Study

• Data retrieved from a public repository, GEO
• Data Series GSE2603
• 121 HG-U133A microarrays
library(affyPLM)
### loads requisite packages including
### affy, Biobase, gcrma etc
Reading in the data

abatch.raw <- ReadAffy()

- Reads the all the CEL files in current directory into an R S4 object known as an **AffyBatch**
- Note we don’t need to supply the CDF file. Instead a processed version of it will get automatically downloaded if needed on the first use of that chip type.
- An **AffyBatch** is an object which can store probe-intensities, along with meta-data such as phenotypic data, for a set of arrays.
- Accessor functions like `pm()`, `mm()` allow access to the PM or MM probe intensities.
- Other functions can be used to visually examine the data

......
boxplot(abatch.raw)
hist(abatch.raw)
MAplot(abatch.raw, plot.method="smoothScatter", which=c(1,56,94,104))
Mbox(abatch.raw)
Computing RMA

eset.rma <- rma(abatch.raw)

- The function `rma()` returns an `exprSet` (in the future this likely to be replaced by the `eSet`) containing RMA values.
- An `exprSet` stores expression values and related meta-data. Many BioConductor functions for high-level analysis accept these as input.
- `gcrma()` can be used to get GCRMA values
boxplot(eset.rma)
MAplot(eset.rma, plot.method="smoothScatter",
          which=c(1, 56, 94, 104))
Carrying out QC Assessment

\[
P_{\text{set}} \leftarrow \text{fitPLM}(\text{abatch.raw})
\]

- A \text{PLMset} object is the return value of fitPLM(). It stores parameter estimates and their standard errors. Also residuals and weights from the IRLS procedure.
NUSE (Pset)
RLE(Pset)
image(Pset, which=99)
image(Pset, which=99, type="resids")
image(Pset, which=99, type="pos.resids")
image(Pset, which=99, type="neg.resids")
image(Pset, which=99, type="sign.resids")
Acknowledgements

- Terry Speed
- Rafael Irizarry
- Julia Brettschneider
- Francois Colin
- Zhijin (Jean) Wu
- Robert Gentleman
- Wolfgang Huber

- Any one else I happened to forget …
References


Supplemental Material
Affymetrix GeneChip

- Commercial mass produced high density oligonucleotide array technology developed by Affymetrix
  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

TAGGTCTGTATGACAGACACAAAGAAGATG
CAGACATAGTGTGTCGTTTTCTTCT
CAGACATAGTGTGTTTTCTTCT

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.

Non-hybridized DNA

Hybridized DNA
Image Analysis
Boxplot raw intensities
Density plots
Comparing arrays

Array2

vs

Array 1

Bad
Comparing 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

Array 1  Array 2  Array 3  Array 4
It works!!

- Unnormalized
- Scaling
- Quantile Normalization
It Reduces Variability

Expression Values

Fold change for Non differential genes

Also no serious bias effects. For more see Bolstad et al (2003)
Summarization

• 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

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

- 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 \( \text{Var}[\varepsilon_{kij}] = \sigma_k^2 \)
Probe Pattern Suggests Including Probe-Effect

Differentially expressing

Non Differential
Variance Covariance Estimates

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

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

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

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

We will use this form