Background and Normalization: Investigating the effects of preprocessing on gene expression estimates

Ben Bolstad
Group in Biostatistics
University of California, Berkeley
bolstad@stat.berkeley.edu
http://www.stat.berkeley.edu/~bolstad
Outline

- “Background” correction methods
- Normalization methods
  - Key results from Bolstad et al (2003)
- Expression Summarization
- Expression Calculation as a 3 step process.
- Case Study: Affymetrix Latin Square
- Conclusions
What is background?

- A measurement of signal intensity caused by auto fluorescence of the array surface and non specific binding.
- Since probes are so densely packed on chip must use probes themselves (rather than region adjacent to probes as in cDNA arrays) to calculate the background.
- In theory the MM should serve as a biological background correction for the PM.
Background correction is:

- A method for removing background noise from signal intensities using information from only one chip
RMA Background

- \( O = S + N \)
- \( O \) – Observed, \( S \) - Signal, \( N \)-noise
- Model \( S \) by Exponential, \( N \) by Normal
- Parameters estimated in ad-hoc way using both PM (and sometimes MM) but worry only about correcting PM
- Background correction is given by \( E(S|O) \)
RMA Background Math

- Observed PM intensity \((O)\)
- Model as sum of signal \((S)\) and background \((N)\)
- Assume \(S\) is exponential \(\alpha\)
- Assume \(N\) is Normal \(\mu, \sigma\)
- Background corrected values are then

\[
E (S \mid O = o) = a + b \quad \frac{\phi \left( \frac{a}{b} \right) - \phi \left( \frac{o - a}{b} \right)}{\Phi \left( \frac{a}{b} \right) + \Phi \left( \frac{o - a}{b} \right) - 1}
\]

\[
a = o - \mu - \sigma^2 \alpha, \quad b = \sigma
\]
MAS 5.0 Background

Use the background correction method as described in Affymetrix “Statistical Algorithm Description Document”

Synopsis:

Break chip into k (k=1..16) rectangular regions

- lowest 2% is chosen as background for that region B_k
- Standard deviation for lowest 2% is chosen as noise for that zone N_k
The background adjustment to be used for cell at \((x,y)\) is weighted average of the \(B_k\), where the weights depend on the distance between \((x,y)\) and the centroids of the regions. \(b(x,y)\)
A noise adjustment is computed in a similar way using $N_k$ rather than $B_k$ : $n(x,y)$

The Background adjusted intensity is given by

$$A(x,y)=\max(I(x,y)-b(x,y),\text{NoiseFrac} \times n(x,y))$$

- Where $\text{NoiseFrac} = 0.5$
MAS 5.0 Mismatch correction

- The way Affymetrix make use of MM in MAS5.0
- Define biweight specific background (SB) for probe pair \( j \) in probeset \( l \) as
  - \( SB_i = T_{bi}(\log_2(PM_i,j) – \log_2(MM_i,j)) : j = 1, \ldots, n_i \)
  - \( T_{bi} \) is Tukey biweight function
MAS 5.0 Mismatch (Cont)

- IM\_i,j is the ideal mismatch
- If MM\_i,j < PM\_i,j
  - IM\_i,j = MM\_i,j
- If MM\_i,j >= PM\_i,j and SB\_i > contrasttau
  - IM\_i,j = PM\_i,j/2^{SB\_i}
- If MM\_i,j >= PM\_i,j and SB\_i <= contrasttau
  - IM\_i,j = PM\_i,j/2^{\text{contrasttau}/(1 + (\text{contrasttau} - SB\_i)/\text{scaletau})}

- contrasttau = 0.03, scaletau=10
- Corrected PM\_i,j is PM\_i,j – MM\_i,j
What is 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."

Normalization is a process of reducing unwanted variation across chips, may use information from multiple chips.
Quantile Normalization

- See Bolstad et al (2002)
- Quantile normalization is a method to make the distribution of probe intensities the same for every chip.
- The normalization distribution is chosen by averaging each quantile across chips.
- The diagram that follows illustrates the transformation.
Normalization Distribution

Raw Data

Density Function

Distribution Function

\[ x = F^{-1}(G(x)) \]
Quantile Normalization (cont)

- The two distribution functions are effectively estimated by the sample quantiles.
- Quantile normalization is fast.
- After normalization, variability of expression measures across chips reduced.
Normalization in MAS

- Compares a collection of experimental array with a baseline array, and normalizes the average intensity of the experimental array to the average intensity of the baseline array during normalization (sometime use a trimmed mean).
- We refer to this method as scaling.
- MAS documentation applies normalization after summarization, we will use it before.
Other prominent methods

- Nonlinear – method used in dChip
  - pick a baseline chip then fit non linear relations (smoothing spines, running medians) between baseline chips and other chips

- Contrast, Cyclic loess
  - generalized M vs A loess normalization methods

- Compares normalization methods in context of RMA measure
- Classifies normalization methods into two classes:
  - Complete Data Methods
    - Quantile
    - Contrast
    - Cyclic Loess
  - Baseline methods
    - Scaling
    - Non-linear
Bolstad et al (2003) cont

- Quantile normalization reduces between chip variability favorably when compared to other methods
Bolstad et al (2003) cont

- Quantile method also found to perform well on the issue of bias (this was measured by using spike-in data)
- Complete data methods recommended over using a baseline
Expression Summarization

Given a set of background corrected and normalized PM probe intensities for each probeset compute a single number for that probeset intended to represent gene expression on an array.
RMA: Robust Multichip Average

Suppose we have $j=1,\ldots,J$ arrays and $i=1,\ldots,I$ probes for a given probeset.

Fit a robust linear model with probe and chip effects to log transformed data.

\[ y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \]

Where $\alpha_i$ is probe-effect and is $\beta_j$ chip effect.

Expression is then (on a log scale) and given by

\[ \mu + \beta_j \]
RMA continued

- Method is compared with MAS 5.0 and Li-Wong MBEI in Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and Terence P. Speed (2002) Summaries of Affymetrix GeneChip Probe Level Data. Accepted to Nucleic Acids Research

- Found to outperform other methods in most regards

- Current implementations use median-polish to fit the linear model, other robust linear model fitting procedures are being explored.
MAS 5.0: “the statistical algorithm”

- Using log-scale data for the probes related to the probeset on single chip.
- Suppose $P_i$ for $i=1,\ldots, I$ are preprocessed probe values.
- Then expression is given by

$$E = T_{bi}(P_1, \ldots, P_I)$$

- Where $T_{bi}()$ is the 1 step Tukey Biweight.
Other expression measures

- Not explored here
- AvDiff – the old Affymetrix method. Found wanting for a number of reasons
- Li-Wong MBEI (Model Based Expression Index) – implemented in the dChip software
Expression Calculation as a 3 step process.

- Suppose x are probe intensities, then 3 step process is
- Background correct $B(x)$
- Normalize $N(x)$
- Transform and Summarize $S(\log_2(x))$
- Put the three together to get $S(\log_2(N(B(x))))$
- In the case of RMA: $B(x)$ is the RMA background correction, $N(x)$ is quantile normalization and $S(x)$ is the robust model fit.
- In the case of MAS 5.0: $B(x)$ is the MAS 5.0 Background followed by IMM subtraction, $N(x)$ leaves the data untouched and $S(x)$ is the tukey bi-weight
Affymetrix has made available the dataset which was used in the development of the MAS 5.0 algorithm.

The Latin square design for the Human data set consists of 14 spiked-in gene groups in 14 experimental groups. The concentration of the 14 gene groups in the first experiment is 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, and 1024pM.

Total of 59 CEL files.
### Case Study: Analysis setup

- We will mix and match the background, normalization and summarization steps, then compare the resulting Gene Expressions.

- In particular we will use:

<table>
<thead>
<tr>
<th>Background</th>
<th>Normalization</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>None</td>
<td>RMA - medianpolish</td>
</tr>
<tr>
<td>RMA Background</td>
<td>Quantile</td>
<td>Tukey Biweight</td>
</tr>
<tr>
<td>MAS5.0 Background</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAS5.0 + IMM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note MAS 5 implementations are based upon available documentation, may not completely agree with MAS 5.0 software.
Computing relative expression

In case of spike-in experiments, average in log–scale across spikein concentration replicates.

If \( E_{i,j} \) is expression of probeset \( i \) in group \( j \), then expression difference between grp 1 and 2 is

\[
M_i = E_{i,1} - E_{i,2}
\]
What does preprocessing do to non differential probesets?

To answer this we look at the relative expression between groups of non differential probesets.

For example if there is 14 dilution groups then there is $14 \times 13 / 2 = 91$ different comparisons for each probeset.
What about the Spike-ins?

- Plot Observed versus Truth in relative expression.
- Also fit linear regression of Observed on truth
Reconciling Results

- We look at how many spike-in relative expressions lie outside various quantiles of all non spike-in relative expressions.
- Probably a little conservative, an improved method would be to see how many spikeins are outside non-spikeins on each group to group comparison.
<table>
<thead>
<tr>
<th></th>
<th>100%</th>
<th>98%</th>
<th>95%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMA</td>
<td>326</td>
<td>1237</td>
<td>1249</td>
<td>1269</td>
</tr>
<tr>
<td>w/o N</td>
<td>325</td>
<td>1161</td>
<td>1191</td>
<td>1253</td>
</tr>
<tr>
<td>w/o B</td>
<td>418</td>
<td>1240</td>
<td>1255</td>
<td>1269</td>
</tr>
<tr>
<td>w/o B&amp;N</td>
<td>424</td>
<td>1153</td>
<td>1171</td>
<td>1254</td>
</tr>
<tr>
<td>w/ MB</td>
<td>287</td>
<td>1226</td>
<td>1239</td>
<td>1270</td>
</tr>
<tr>
<td>w/MB -</td>
<td>287</td>
<td>1153</td>
<td>1191</td>
<td>1259</td>
</tr>
<tr>
<td>imm</td>
<td>232</td>
<td>1073</td>
<td>1142</td>
<td>1267</td>
</tr>
<tr>
<td>imm-</td>
<td>231</td>
<td>1063</td>
<td>1130</td>
<td>1265</td>
</tr>
<tr>
<td>mb/imm</td>
<td>84</td>
<td>904</td>
<td>995</td>
<td>1260</td>
</tr>
<tr>
<td>mb/imm-</td>
<td>76</td>
<td>902</td>
<td>985</td>
<td>1258</td>
</tr>
</tbody>
</table>
What about using Tukey Bi-weight?

For brevity similar plots not shown, but conclusions about effects of preprocessing similar
Conclusions

- Preprocessing can help you, but it can also harm you.
- RMA does pretty good, the MAS IMM helps predict true fold change, but reduces sensitivity in detecting outliers.
Useful Data Sets

- **Affymetrix**
  - “Affymetrix® Latin Square Data for Expression Algorithm Assessment”
    - [http://www.affymetrix.com/analysis/download_center2.affx](http://www.affymetrix.com/analysis/download_center2.affx)

- **Genelogic**
  - Spike in and dilution datasets
Useful software

- The R “affy” package which is a component of the bioconductor project
  [http://www.bioconductor.org](http://www.bioconductor.org)
  - Provides
    - Framework for doing low level analysis and expression computation of GeneChip data.
    - Fast RMA expression computation (as of version 1.1)
    - Ability to mix and match background, normalization and expression summary methods.
Some Papers


- Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and Terence P. Speed (2002) *Summaries of Affymetrix GeneChip Probe Level Data*. Accepted to *Nucleic Acids Research*
Acknowledgements

- Terry Speed (UC Berkeley, WEHI)
- Rafael Irizarry (JHU)
- Francois Collin (Genelogic)