Normalization and Construction of Expression Measures for Affymetrix GeneChip Data

1/20/2009

Affymetrix .CEL Files

- A .CEL file contains one number representing signal intensity for each probe cell on a single GeneChip.
- .CEL files can be read with Affymetrix software or in R using the Bioconductor package *affy*.
- We will discuss two methods for normalizing and obtaining expression measures using data from Affymetrix .CEL files.

Methods

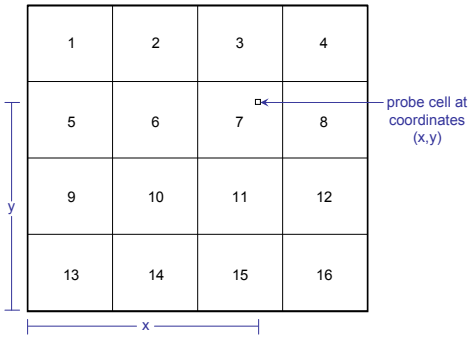
1. Microarray Analysis Suite (MAS) 5.0 Signal proposed by Affymetrix. *Statistical Algorithms Description Document* (2002) Affymetrix Inc.
2. Robust Multi-array Average (RMA) proposed by Irizarray et al. (2003) *Biostatistics* **4**, 249-264.

These are perhaps the two most popular of many methods for normalizing and computing expression measures using Affymetrix data. Currently over 50 methods are described and compared at http://affycomp.biostat.jhsph.edu/.

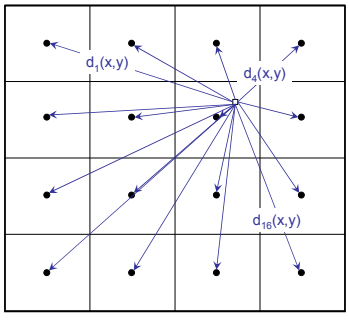
MAS 5.0 Signal: Background Adjustment

- Each chip is divided into 16 rectangular zones.
- The lowest 2% of intensities in each zone are averaged to form a zone-specific background value denoted bZ_k for zones k=1, 2, ..., 16.
- The standard deviation of the lowest 2% of intensities in each zone is calculated and denoted nZ_k for zones k=1, 2, ..., 16.
- Let $d_k(x,y)$ denote the distance from the center of zone k to a probe cell located at coordinates (x,y) on the chip.

GeneChip Divided into 16 Zones



16 Distances to Zone Centers for Each Probe Cell



MAS 5.0 Signal: Background Adjustment (continued)

- Let $w_k(x,y)=1/(d_k^2(x,y)+100)$.
- Denote the background for the cell located at coordinates (x,y) by

$$b(x,y)=\Sigma_{k=1}^{16} w_k(x,y)\, bZ_k / \Sigma_{k=1}^{16} w_k(x,y).$$

- Denote the "noise" for the cell located at coordinates (x,y) by

$$n(x,y)=\Sigma_{k=1}^{16} w_k(x,y)\, nZ_k / \Sigma_{k=1}^{16} w_k(x,y).$$

7

MAS 5.0 Signal: Background Adjustment (continued)

- Let I(x,y) denote the original intensity of the cell located at coordinates (x,y) on the chip. (75^{th} percentile of 36 pixel intensities in the center of the cell.)
- Let $I'(x,y)=\max(I(x,y) , 0.5)$.
- Define the background-adjusted intensity for the cell at coordinates (x,y) by

$$A(x,y)=\max\{ I'(x,y)-b(x,y) , 0.5n(x,y) \}.$$

- Henceforth these background-adjusted intensities will be referred to as either PM or MM for perfect match or mismatch cells, respectively.

8

MAS 5.0 Signal: Ideal Mismatch Computation

- MM values are supposed to provide measures of cross-hybridization and stray signal intensity that inflate the value of PM.
- In the simplest case, a PM value would be corrected simply by subtracting its corresponding MM value.
- However, some MM values are bigger than their corresponding PM values so that PM-MM would become negative.
- Because negative values do not make a lot of sense and would pose problems with subsequent steps in analysis, Affymetrix determines an *Ideal Mismatch* (IM) value for each probe pair that is guaranteed to be less than PM.

9

MAS 5.0 Signal: Ideal Mismatch Computation (continued)

For a given probe set containing n probe pairs, let PM_j and MM_j denote the perfect match and mismatch values of the j^{th} probe pair. The IM value from the j^{th} probe pair (IM_j) is determined as follows:

- If $PM_j > MM_j$, then $IM_j = MM_j$ and no further computation is needed.
- If $PM_j \le MM_j$, compute

$$M = TBW\{ \log_2(PM_1/MM_1),\ldots,\log_2(PM_n/MM_n) \}$$

where TBW denotes a one-step Tukey BiWeight (a special weighted average described later).

10

MAS 5.0 Signal: Ideal Mismatch Computation (continued)

- If $M > 0.03$, then $IM_j = PM_j / 2^M$.
- If $M \le 0.03$, then compute $P = \dfrac{0.03}{1+(\frac{0.03-M}{10})}$ and let

$$IM_j = PM_j / 2^P.$$

- Note that at M = 0.03, $IM_j = PM_j / 1.021012$ so that PM_j will be slightly larger than IM_j.
- As M gets larger, IM_j decreases. As M gets smaller, IM_j increases towards $PM_j / 1.020949$.

11

MAS 5.0 Signal: Signal Log Value Computation

- Let $V_j = \max(PM_j - IM_j , 2^{-20})$.
- Define the *probe value* for the j^{th} probe pair by $PV_j = \log_2(V_j)$.
- The *signal log value* for a given probe set is defined by

$$SLV = TBW(PV_1 , PV_2 , \ldots , PV_n)$$

where TBW denotes a one-step Tukey BiWeight

(a special weighted average to be discussed later).

12

MAS 5.0 Signal: Scaling and Signal Calculation

- Let SLV_i denote the signal log value for the i^{th} probe set on a single chip.
- Let I denote the number of probe sets on the chip.
- Let SF = 500/TrimMean(2^{SLV_1}, 2^{SLV_2}, ..., 2^{SLV_I} ; 0.02,0.98).

  The average of the values between the left parenthesis and the semicolon that are strictly between the 0.02 and 0.98 quantiles of the values between the left parenthesis and the semicolon.
- *MAS 5.0 Signal* for the i^{th} probe set is $Signal_i = SF * 2^{SLV_i}$.
- All computations are done separately for each chip to obtain a Signal value for each chip and probe set.

13

The One-Step Tukey BiWeight Estimator Used by Affymetrix

- Let $x_1, x_2, ..., x_n$ denote observations.
- Let m = median ($x_1, x_2, ..., x_n$).
- Let MAD = median ($|x_1 - m|, |x_2 - m|, ..., |x_n - m|$).
- For each i = 1, 2, ..., n; let $t_i = \frac{x_i - m}{5 * MAD + 0.0001}$.

  Factor Affymetrix uses to avoid division by 0.

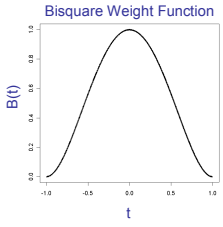
14

The One-Step Tukey BiWeight Estimator Used by Affymetrix (ctd.)

Recall the bisquare weight function defined as

$B(t) = (1 - t^2)^2$ for $|t| < 1$

$= 0$ for $|t| \geq 1$.



$$TBW (x_1, x_2, ..., x_n) = \frac{\sum_{i=1}^{n} B(t_i) x_i}{\sum_{i=1}^{n} B(t_i)}$$

15

An Example

Compute TBW (1, 7, 13, 15, 28, 1075).

Ignore the 0.0001 factor to make calculations easier.

m = (13 + 15) / 2 = 14.

MAD = median (|1-14|,|7-14|,|13-14|,|15-14|,|28-14|,|1075-14|)
= median (13, 7, 1, 1, 14, 1061)
= median (1, 1, 7, 13, 14, 1061)
= (7 + 13) / 2 = 10.

$t_1 = -13 / 50$ $t_2 = -7 / 50$ $t_3 = -1 / 50$

$t_4 = 1 / 50$ $t_5 = 14 / 50$ $t_6 = 1061 / 50$

16

An Example (continued)

$t_1 = -13 / 50$ $t_2 = -7 / 50$ $t_3 = -1 / 50$

$t_4 = 1 / 50$ $t_5 = 14 / 50$ $t_6 = 1061 / 50$

$B(t_1)=B(-0.26)=B(0.26)=(1 - 0.26^2)^2 = 0.8693698$
$B(t_2)=B(-0.14)=B(0.14)=(1 - 0.14^2)^2 = 0.9611842$
$B(t_3)=B(-0.02)=B(0.02)=(1 - 0.02^2)^2 = 0.9992002$
$B(t_4)=B(0.02)=(1 - 0.02^2)^2 = 0.9992002$
$B(t_5)=B(0.28)=(1 - 0.28^2)^2 = 0.8493466$
$B(t_6)=0$

$$\frac{0.8693698*1+ 0.9611842*7+0.9992002*13+0.9992002*15+0.8493466*28+0*1075}{0.8693698+ 0.9611842+0.9992002+0.9992002+0.8493466+0}$$

=12.68772.

17

Obtaining MAS5.0 Signal Values from Affymetrix .CEL Files

- MAS5.0 Signal values can be obtained from Affymetrix software.
- Approximate MAS5.0 Signal values can be computed with the *mas5* function that is part of the Bioconductor package *affy*.
- Use whichever method is easiest for you. The differences do not seem to be large enough to matter.

18

Installing R

- R is a free language and environment for statistical computing and graphics.
- Information about R including installation instructions and documentation can be found at

  www.r-project.org.

19

Installing Bioconductor

- Bioconductor is an open source and open development software project for the analysis and comprehension of genomic data.
- Information about Bioconductor can be found at www.bioconductor.org.
- To install Bioconductor type the following commands at the R prompt.

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

getBioC()
```

20

R Commands for Obtaining MAS5.0 Signal Values from Affymetrix .CEL Files

```
#
#Load the Bioconductor package affy.
#
library(affy)
#
#Set the working directory to the directory containing all the .CEL files.
#
setwd("C:/z/Courses/Smicroarray/AffyCel")
#
#Read the .CEL file data.
#
Data=ReadAffy()
#
#Compute the MAS5.0 Signal Values
#
signal=mas5(Data)
#
#Write the data to a tab-delimited text file.
#
write.exprs(signal, file="mydata.txt")
```

21

Robust Multi-array Average (RMA)

1. Background adjust PM values from .CEL files.
2. Take the base-2 log of each background-adjusted PM intensity.
3. Quantile normalize values from step 2 across all GeneChips.
4. Perform median polish separately for each probe set with rows indexed by GeneChip and columns indexed by probe.
5. For each row, find the average of the fitted values from step 4 to use as probe-set-specific expression measures for each GeneChip.

22

RMA: Background Adjustment

Assume PM = S + B where

signal S ~ Exp(λ) independent of

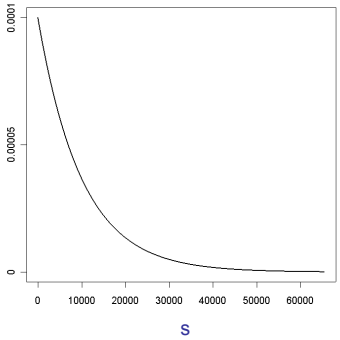
background B ~ $N^+(\mu,\sigma^2)$.

$N^+(\mu,\sigma^2)$ denotes $N(\mu,\sigma^2)$ truncated on the left at 0.

23

The Probability Density Function of the Exponential Distribution with Mean 1/λ = 10000

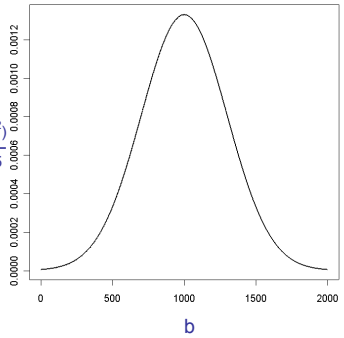
$\lambda e^{-\lambda s}$



24

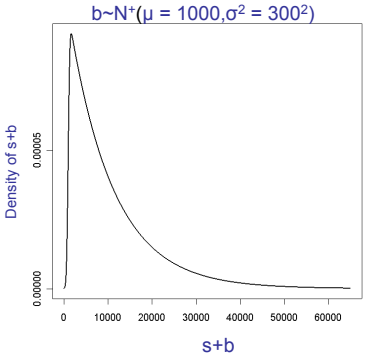
The Probability Density Function of the Normal Distribution with Mean μ = 1000 and Variance $\sigma^2 = 300^2$

$$\frac{e^{-(b-\mu)^2/(2\sigma^2)}}{(2\pi\sigma^2)^{0.5}}$$



25

The Probability Density Function of s + b where s~Exp(λ=1/10000) and $b \sim N^+(\mu = 1000, \sigma^2 = 300^2)$



26

RMA: Background Adjustment (continued)

N(0,1) density function

$$E(S|PM) = PM - \mu - \lambda\sigma^2 + \sigma \cdot \frac{\phi\left(\frac{PM-\mu-\lambda\sigma^2}{\sigma}\right) - \phi\left(\frac{\mu+\lambda\sigma^2}{\sigma}\right)}{\Phi\left(\frac{PM-\mu-\lambda\sigma^2}{\sigma}\right) + \Phi\left(\frac{\mu+\lambda\sigma^2}{\sigma}\right) - 1}$$

N(0,1) distribution function

Separately for each chip, estimate μ, σ, and λ from the observed PM distribution. Plug those estimates into the formula above to obtain an estimate of E(S|PM) for each PM value. These serve as background-adjusted PM values.

27

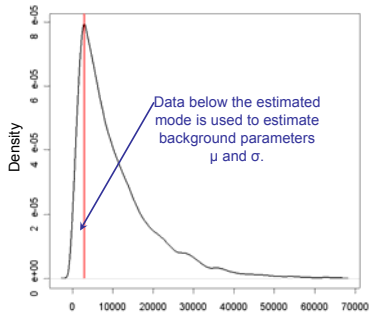
RMA: Background Adjustment (continued)

Obtaining Estimates of μ, σ, and λ
(unpublished description of the procedure)

- Estimate the mode of the PM distribution using a kernel density estimate of the PM density.
- Estimate the density of the PM values less than the mode. The mode of this distribution serves as an estimate of μ.
- Assume the data to the left of the estimate of μ are the background observations that fell below their mean. Use those observations to estimate σ.
- Subtract the estimate of μ from all observations larger than the estimate. The mode of this distribution estimates 1/λ.

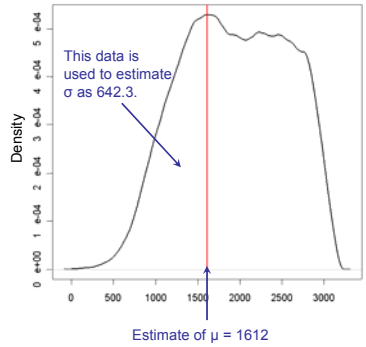
28

PM Density Estimate Based on Simulated Data



29

Density Estimate of PM Data below the Estimated Mode of the PM Distribution



30

Estimate of σ

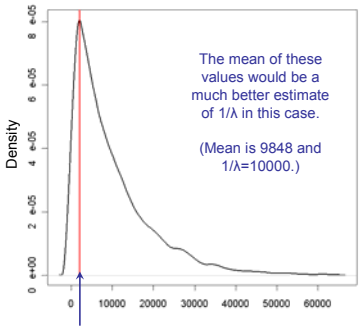
According to the RMA R code, σ is estimated as follows:

$$\hat{\sigma} = \sqrt{\frac{2\sum_{PM<\hat{\mu}}(PM-\hat{\mu})^2}{\#\{PM<\hat{\mu}\}-1}}$$

The purpose of the factor of 2 in the numerator is not clear.

31

Density Estimate of PM – $\hat{\mu}$ Values Greater than Zero



Estimate of 1/λ = 2019

32

RMA: Quantile Normalization

1. After background adjustment, find the smallest $\log_2$(PM) on each chip.
2. Average the values from step 1.
3. Replace each value in step 1 with the average computed in step 2.
4. Repeat steps 1 through 3 for the second smallest values, third smallest values,..., largest values.

33

RMA: Median Polish

- For a given probe set with J probe pairs, let y_{ij} denote the background-adjusted, base-2-logged, and quantile-normalized value for GeneChip i and probe j.
- Assume $y_{ij} = \mu_i + \alpha_j + e_{ij}$ where $\alpha_1 + \alpha_2 + ... + \alpha_n = 0$.

μ_i: gene expression of the probe set on GeneChip i

α_j: probe affinity affect for the j^{th} probe in the probe set

e_{ij}: residual for the j^{th} probe on the i^{th} GeneChip

- Perform Tukey's Median Polish on the matrix of y_{ij} values with y_{ij} in the i^{th} row and j^{th} column.

34

RMA: Median Polish (continued)

- Let $\hat{y}_{ij}$ denote the fitted value for y_{ij} that results from the median polish procedure.
- Let $\hat{\alpha}_j = \hat{y}_{\cdot j} - \hat{y}_{\cdot\cdot}$ where $\hat{y}_{\cdot j} = \frac{\sum_{i=1}^{I}\hat{y}_{ij}}{I}$ and $\hat{y}_{\cdot\cdot} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{y}_{ij}}{IJ}$ and I denotes the number of GeneChips.
- Let $\hat{\mu}_i = \hat{y}_{i\cdot} = \sum_{j=1}^{J}\hat{y}_{ij} / J$
- $\hat{\mu}_i$ is the probe-set-specific measure of expression for GeneChip i.

35

An Example

Suppose the following are background-adjusted, $\log_2$-transformed, quantile-normalized PM intensities for a single probe set. Determine the final RMA expression measures for this probe set.

GeneChip \ Probe	1	2	3	4	5
1	4	3	6	4	7
2	8	1	10	5	11
3	6	2	7	8	8
4	9	4	12	9	12
5	7	5	9	6	10

36

An Example (continued)

4	3	6	4	7	4
8	1	10	5	11	8
6	2	7	8	8	7
9	4	12	9	12	9
7	5	9	6	10	7

row medians

0	-1	2	0	3
0	-7	2	-3	3
-1	-5	0	1	1
0	-5	3	0	3
0	-2	2	-1	3

matrix after removing row medians

37

An Example (continued)

0	-1	2	0	3	0	4	0	0	0
0	-7	2	-3	3	0	-2	0	-3	0
-1	-5	0	1	1	-1	0	-2	1	-2
0	-5	3	0	3	0	0	1	0	0
0	-2	2	-1	3	0	3	0	-1	0
0	-5	2	0	3					

column medians

matrix after subtracting column medians

38

An Example (continued)

0	4	0	0	0	0
0	-2	0	-3	0	0
-1	0	-2	1	-2	-1
0	0	1	0	0	0
0	3	0	-1	0	0

row medians

0	4	0	0	0
0	-2	0	-3	0
0	1	-1	2	-1
0	0	1	0	0
0	3	0	-1	0

matrix after removing row medians

39

An Example (continued)

0	4	0	0	0	0	3	0	0	0
0	-2	0	-3	0	0	-3	0	-3	0
0	1	-1	2	-1	0	0	-1	2	-1
0	0	1	0	0	0	-1	1	0	0
0	3	0	-1	0	0	2	0	-1	0
0	1	0	0	0					

column medians

matrix after subtracting column medians

40

An Example (continued)

0	3	0	0	0
0	-3	0	-3	0
0	0	-1	2	-1
0	-1	1	0	0
0	2	0	-1	0

All row medians and column medians are 0.
Thus the median polish procedure has converged.
The above is the residual matrix that we will subtract from the original matrix to obtain the fitted values.

41

An Example (continued)

original matrix

$$\begin{pmatrix} 4 & 3 & 6 & 4 & 7 \\ 8 & 1 & 10 & 5 & 11 \\ 6 & 2 & 7 & 8 & 8 \\ 9 & 4 & 12 & 9 & 12 \\ 7 & 5 & 9 & 6 & 10 \end{pmatrix} -$$

residuals from median polish

$$\begin{pmatrix} 0 & 3 & 0 & 0 & 0 \\ 0 & -3 & 0 & -3 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 2 & 0 & -1 & 0 \end{pmatrix}$$

$=$

matrix of fitted values

$$\begin{pmatrix} 4 & 0 & 6 & 4 & 7 \\ 8 & 4 & 10 & 8 & 11 \\ 6 & 2 & 8 & 6 & 9 \\ 9 & 5 & 11 & 9 & 12 \\ 7 & 3 & 9 & 7 & 10 \end{pmatrix}$$

row means

$$\begin{pmatrix} 4.2 \\ 8.2 \\ 6.2 \\ 9.2 \\ 7.2 \end{pmatrix} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \\ \hat{\mu}_4 \\ \hat{\mu}_5 \end{pmatrix}$$

RMA expression measures for the 5 GeneChips

42

R Commands for Obtaining RMA Expression Measures from Affymetrix .CEL Files

```
#
#Load the Bioconductor package affy.
#
library(affy)
#
#Set the working directory to the directory containing all the .CEL files.
#
setwd("C:/z/Courses/Smicroarray/AffyCel")
#
#Read the .CEL file data.
#
Data=ReadAffy()
#
#Compute the RMA measures of expression.
#
expr=rma(Data)
#
#Write the data to a tab-delimited text file.
#
write.exprs(expr, file="mydata.txt")
```

43