An Introduction to Affymetrix Microarrays

Naomi Altman
Format of an Affymetrix Array

• Each gene is represented by a “probe set”
• Each “probe set” is 11-20 pairs of oligos
• Each oligo is 25 nucleotides
• A PM (perfect match) probe matches a strand of cDNA
• The corresponding MM (mismatch) probe differs from the PM by a change in the central nucleotide
• The probe pairs are spatially dispersed
• Control probes are printed
Format of an Affymetrix Array

Figure 3: Oligonucleotide Probe Pair

http://cnx.rice.edu/content/m12388/latest/figE.JPG
Printing the Slide

Heuristics for “Probe Sets”

MM probe is supposed to control for:
• Variation in chemical composition
• Abundance of cross-hybridizing fragments from other genes

By combining PM and MM information from many probes, gene to gene differences should be minimized.
Heuristics for “Probe Sets”

Initially it was thought that this approach would make it possible to consider the measured intensities as a proxy for actual mRNA concentration.
Affy expression files

The scanner software produces 3 files for each array:

The DAT file is the intensity for each pixel on the array.

The CEL file, which is the summary of probe intensity for every probe on the array.

An expression file, which is the Affymetrix Suite compilation of the probes into gene expression values. (MAS method)

We will start by looking at the CEL file.
Affy package in Bioconductor

We are going to read in the data we downloaded.

We open R in the directory with the CEL files.

library(affy)

Clicking on "vignettes" in the GUI brings up the main documentation
data=ReadAffy()  # reads all the CEL files into data

gNames=geneNames(data)  # retrieves the probeset names

pm(data,gName[5])  # retrieves the positive match probe expression values for gName 5 for every array

pm(data)  # retrieves all the PM expression values

mm  # is the matching command for the mismatch probes
Quality Checks

image(data)
image(data,col=rainbow(6))
gnames[1:5]
gnames[5]

pm5=pm(data,gnames[5])
mm5=mm(data,gnames[5])

#Lets have a look at the PM and MM probes for the 5th probeset on the first array.

plot(mm5[,1],pm5[,1],xlab="Mismatch",ylab="Perfect Match", main=gnames[5])
boxplot(as.data.frame(t(pm5)))
boxplot(as.data.frame(t(mm5)), add=T, boxwex=.5)
Some other applications of the Affymetrix paradigm

Genotyping:

Exon:

Tiling:

7 or 14 array sets - you need a big budget!
Gene Expression

The objective is to combine information from the PM and MM probe set to quantify gene expression.

Comparability for a single gene across arrays is essential.

Comparability among genes on a single array is desirable.
In most cases the probes on different arrays should be similar

\[ \text{pairs}(\log(\text{PM}[1:2000,1:5],2)) \]
\[ \text{pairs}(\log(\text{MM}[1:2000,1:5],2)) \]
You might also plot MM vs PM
RNA Digestion Plot

plotAffyRNAdeg(AffyRNAdeg(AffyData),col=c(1,1,2,2,3,3,4,4,5,5,6,6))
legend(locator(),legend=paste(phenotype),col=c(1,1,2,2,3,3,4,4,5,5,6,6),pc
Once you are satisfied that the arrays are OK
Convert to gene expression values.
Steps in Data Extraction

An intensity is read for each probe.

The intensities are corrected for background noise.

The probe intensities are normalized.

The intensities are combined into probeset expression values.
Why Normalize?

Normalization is supposed to remove systematic measurement errors, both within and between arrays. These can be seen in the scatterplot matrices as deviations from linearity between arrays (especially replicates).

This should improve our power to detect differential expression.
Background/Signal Adjustment

• A method which does some or all of the following
  • Corrects for background noise, processing effects
  • Adjusts for cross hybridization
  • Adjust estimated expression values to fall on proper scale
• Probe intensities are used in background adjustment to compute correction (unlike cDNAarrays where area surrounding spot might be used)

from: http://stat-www.berkeley.edu/users/bolstad/talks/TalkCSUEB.pdf
Affymetrix Expression Values (MAS 5.0)

For gene i, we have pairs $pm_{ij}$, $mm_{ij}$. We form the background corrected values:

$$f_{ij} = \max(\log_2(pm_{ij}/mm_{ij}), d) \quad d > 0$$

We form the weighted average

$$f_i = \sum_{j=1}^{n \text{ probes}} w_{ij} f_{ij}$$

The weights are proportional to the distance between $f_{ij}$ and the median. The expression value returned is $2^{f_i}$
Affymetrix Expression Values (MAS 5.0)

Based on the 11 PM-MM pairs, a Wilcoxon test is done to determine if the gene expressed in the sample. This is called the "detection p-value".

\[ R_{ij} = \frac{PM_{ij} - MM_{ij}}{PM_{ij} + MM_{ij}} \]

**H₀**: median(\( R_{ij} \)) ≤ T

Rank all of the \(|R_{ij} - T|\) values : \( r_{ij} \).

Sum up the ranks for the values with \( R_{ij} - T > 0 \).

Compute the p-value on under the null hypothesis.

Gene is "detected" if we reject \( H₀ \).
Objections to Affymetrix Normalization

The Affymetrix approach is under attack because between 15% - 30% of the MM are greater than the PM.

```
sum(MM[,1]>PM[,1])/nrow(PM)
```

On some newer arrays, MM probes are not used, allowing for twice as many genes to be represented.

But many people like the "detection p-value"
Other Expression Extraction Methods

RMA (robust multi-array analysis)

After background correction

Probes are normalized using quantile method.

PM probes only are combined into gene expression values using “median polish” method that recognizes outliers for the probe across arrays and within the probe set on one array.
Looking at the R code: for each array, using only the PM probes:
Use the intensity histogram to identify the probes below the peak of $O=PM$. Use these to estimate $\mu$, $\sigma$.
Use the probes above the peak of the "background" probes, to estimate $\alpha$. 

$$E(S|O=o) = a + b \frac{\phi\left(\frac{a}{b}\right) - \phi\left(\frac{o-a}{b}\right)}{\Phi\left(\frac{a}{b}\right) + \Phi\left(\frac{o-a}{b}\right) - 1}$$

$$a = o - \mu - \sigma^2 \alpha, \quad b = \sigma$$
Quantile Normalization

Adjusts probes across arrays by:

- Sorting all probes from smallest to largest
  \( p_{i(j)} \) is the j-th largest probe on array i
- Replace \( p_{i(j)} \) by the median_{i} (\( p_{i(j)} \))
- In RMA, this is done on the background adjusted PM probes.
Probe distribution before background correction and quantile normalization.
Median Polish

For a selected probeset we have

\[ f_{ij} \] is the quantile normalized log2 expression value
of the \( j^{th} \) pm probe on the \( i^{th} \) array.

A "median" probeset effect is estimated over all
arrays.
A"median" array effect is estimated for the probeset
over all probes on the array.
The log2(expression) for the gene on the \( i^{th} \) array is
estimated as the sum of these effects.
Other Expression Extraction Methods

dChip=Li and Wong

Uses a robust regression of PM on MM for each probe set.
Other Expression Extraction Methods

GCRMA

Assumes that the probe intensity is a function of optical noise, non-specific binding, specific binding, whether PM or MM and the probe sequence.

Probe intensities are adjusted by this model.

Signal is then estimated using median polish.
Does the Normalization Matter For Assessing Absolute Expression?

RMA=rma(data)
MAS=log(mas5(data),2)

Even at the high end the results differ.
There are "non-detections" and "detections" throughout the range.
But what we care about is Differential Expression

This shows the log ratio for 2 replicate arrays.
More than 100 points have a 2-fold difference in log ratio.
The problem with RMA, gcRMA, dChip methods

You cannot add a new array without renormalizing all.

These will not work well if the arrays are not comparable. (e.g. RNA degradation experiments).

Is this reasonable? In comparing A and B, should it matter if condition C is included in the data?

This problem is not unique to normalization. It also occurs with any statistical method of determining differential expression.
Our experience at PSU

Affymetrix arrays are highly replicable.
Affymetrix arrays are widely used - hence there are some good informatics tools available, including freeware in Bioconductor.
Most studies that use data compiled from the www use Affy data - more comparable than other types.

The high cost on a per array basis is recouped by fewer failures and simpler processing
The Best Normalization

The "gold standard" is not clear. Most of the procedures are tested on a small number of experiments in which mRNA was spiked into samples.

There are some interesting papers and talks at
http://stat-www.berkeley.edu/users/bolstad
http://stat-www.berkeley.edu/users/terry
http://biosun01.biostat.jhsph.edu/~ririzarr