This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike License. Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.
Feature Level Data

Outline

• Affymetrix GeneChip arrays
• Two color platforms
Non-specific Hybridization

Affymetrix GeneChip Design

GeneChip Feature Level Data

• MM features used to measure optical noise and non-specific binding directly
• More than 10,000 probesets
• Each probeset represented by 11-20 feature
  Note 1: Position of features are haphazardly distributed about the array.
  Note 2: There are between 10-15 chip types
• So we have $PM_{gi}$, $MM_{gi}$
  ($g$ is gene, $i$ is array and $j$ is feature)
• A default summary is the avg of the PM-MM
Two color platforms

- Common to have just one feature per gene
- Typically, longer molecules are used so non-specific binding not so much of a worry
- Optical noise still a concern
- After spots are identified, a measure of local background is obtained from area around spot

Local background

GenePix does something different these days

Two color feature level data

- Red and Green foreground and and background obtained from each feature
- We have $R_{gf}^{ij}$, $G_{gf}^{ij}$, $R_{gb}^{ij}$, $G_{gb}^{ij}$ ($g$ is gene, $i$ is array and $j$ is replicate)
- A default summary statistic is the log-ratio: $(R_f - R_b) / (G_f - G_b)$
Affymetrix Spike In Experiment

Spike-in Experiment

• Throughout we will be using Data from Affymetrix’s spike-in experiment

• Replicate RNA was hybridized to various arrays

• Some probesets were spiked in at different concentrations across the different arrays

• This gives us a way to assess precision and accuracy

• Done for HGU95 and HGU133 chips

• Available from Bioconductor experimental data package: SpikeIn

Spikein Experiment (HG-U95)

Probeset
Spikein Experiment (HG-U133)

- A similar experiment was repeated for a newer chip
- The 1024 picoMolar concentration was not used. 1/8 was used instead.
- No groups of 12
- Note: More spike-ins to come!

Background Effects
Experiments

Learn about optical effect and NSB

<table>
<thead>
<tr>
<th>label</th>
<th>sample type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empty</td>
<td>empty</td>
</tr>
<tr>
<td>NoRNA</td>
<td>no RNA</td>
</tr>
<tr>
<td>NoLabel</td>
<td>human</td>
</tr>
<tr>
<td>YeastDNA</td>
<td>yeast genomic DNA</td>
</tr>
<tr>
<td>polyC</td>
<td>poly C</td>
</tr>
<tr>
<td>polyG</td>
<td>poly G</td>
</tr>
</tbody>
</table>
The Background Effects

This are the no-label and Yeast DNA chips
Why Adjust for Background?

Notice local slope decrease as the nominal concentration becomes small.

Probe-specific NSB
Why not subtract MM,BG?

Why not subtract MM?

Why not subtract MM?
Solutions

Direct Measurement Strategy

The hope is that:

\[
\begin{align*}
PM &= B + S \\
MM &= B
\end{align*}
\]

\[PM - MM = S\]

But this is not correct!

Notice
- We care about ratios
- We usually take log of S

Stochastic Model

Better to assume:

\[
\begin{align*}
PM &= B_{PM} + S \\
MM &= B_{MM}
\end{align*}
\]

\[\text{Cov}(\log(B_{PM}), \log(B_{MM})) \approx 0.7\]

\[\text{Var}[\log(PM - MM)] \approx 1/S^2\]

Alternative solution:

\[E[S \mid PM]\]
Simulation

• We create some feature level data for two replicate arrays
• Then compute $Y = \log(PM - kMM)$ for each array
• We make an MA using the $Y$s for each array
• We make a observed concentration versus known concentration plot
• We do this for various values of $k$. The following “movie” shows $k$ moving from 0 to 1.
Real Data

PM = B + S

Observed: PM

Of interest: S

Pose a statistical model and use it to predict S from the observed PM.

The Basic Idea

PM = B + S

• A mathematically convenient, useful model

  – B ~ Normal (μ, σ)
  – S ~ Exponential (λ)

  \[ \hat{S} = E[S \mid PM] \]

  – No MM
  – Borrowing strength across probes
**MAS 5.0**

Notice improved precision but worst accuracy

**Problem**

- Global background correction ignores probe-specific NSB
- MM have problems
- Another possibility: Use probe sequence
Sequence effect


Affinity = \sum_{j,k} \sum_{i,j} \mu_{jk} = \text{smooth function of } k

General Model

PM_{\alpha} = Q^{\alpha} + \exp(h_{ij}(\alpha_{ij} + k_{ij}) + \exp(f_{ij}(\alpha_{ij} + \theta_{ij} + z_{ij}))

MM_{\alpha} = O^{\alpha} + \exp(b_{ij}(\alpha_{ij} + k_{ij} + \#_{ij} + \%_{ij}))

We can calculate: \( E[\theta_{ij} \mid PM_{\alpha}, MM_{\alpha}] \)

Explains Bimodality