Exploration and Analysis of DNA Microarray and Protein Array Data

Dhammika Amaratunga
Senior Research Fellow in Nonclinical Biostatistics
Johnson & Johnson Pharmaceutical Research & Development

Javier Cabrera
Director of the Institute of Biostatistics
Rutgers University

A short course offered by the New York City Chapter of the ASA, March 2007
Exploration and Analysis of DNA Microarray and Protein Array Data
Dhammika Amaratunga & Javier Cabrera

Agenda

1. Microarrays: experimental objectives and procedures (DA)
2. Preprocessing microarray data (DA)
3. Finding differentially expressed genes (DA&JC)
4. Multivariate methods (JC)
5. Computational issues and software (JC)
Gene expression

♦ A *gene* is a segment of a DNA molecule that codes for a specific protein involved in a specific biological function.

♦ Generally a gene is inactive, but when it is active, it is *expressed* via the process:

\[
\text{DNA} \rightarrow \text{mRNA} \rightarrow \text{protein}
\]

- transcription
- translation

♦ Gene expression ← mRNA abundance
Research in functional genomics

♦ **Fact:** Specific genes express in specific cells under specific conditions $\rightarrow$ specific states for cell/organism.

♦ **Research:** Which genes are expressed in which cells and under what conditions.

♦ **Roadblock:** Large number of genes (e.g., human genome: $\sim 35,000$ genes).

♦ **Solution:** Robotics-driven high-throughput experiments.
DNA microarrays

♦ DNA microarrays are the most widely used tool to monitor the expression levels of many thousands of genes simultaneously.

♦ A DNA microarray is a tiny (glass) slide on which known DNA sequences (corresponding to genes) in solution have been robotically spotted in a rectangular array.
Prototypical microarray experiment

cDNA or oligonucleotide preparation

Glass slide

Biological sample

mRNA

Reverse transcribe and label

Print or synthesize

Microarray

Sample

Hybridize, wash and scan

Image

Quantify spot intensities

Gene expression data

5k-50k genes arrayed in rectangular grid; one spot per gene

+
Interpreting the scanned image

- High intensity spot $\Rightarrow$ the DNA at that spot corresponds to some mRNA in sample.
- Low intensity spot $\Rightarrow$ no mRNA in sample that corresponds to the DNA at that spot.
- Spot intensity $\sim$ mRNA abundance.
- For any gene, can compare intensities across different samples (but shouldn’t compare intensities for different genes for the same sample).
Comparing two scanned images

Control

- same genes on each slide
- only subset shown

Treatment
Technology differences

♦ pin spotting or photolithography or...

♦ multi-channel or single-channel

♦ almost-complete or subsequences
sequences (cDNA) (oligonucleotides)

cDNA array Affymetrix chip
cDNA microarrays

♦ Genes represented by long cDNA probes.

♦ For any sample:

   **Spot Intensity ~ gene expression level**

♦ Two-dye format: Take two mRNA samples, label each with a different fluorescent dye, then disperse the composite sample over microarray; measure the spot intensity for each dye.

   **Spot Intensity Ratio ~ differential expression**

♦ Advantage of two-dye format: Natural matching of samples; reduces spot-related bias.

♦ Disadvantage: Nonspecific hybridization.
Oligonucleotide arrays

♦ Each gene is represented by one or more short oligonucleotide probes.

♦ If one probe per gene:

Spot Intensity ~ gene expression level

♦ If multiple probes per gene:

Spot Intensity (summarized across probes) ~ gene expression level

♦ Advantage: Less nonspecific hybridization.

♦ Disadvantage: Lower sensitivity.
Affymetrix oligonucleotide arrays

♦ A gene is represented by a probe set of ~11 25bp-oligonucleotides called perfect matches (PM).

♦ Each PM is paired with a mismatch (MM) formed by switching the middle base of PM - MM acts as a (imperfect) control for nonspecific hybridization.

...CTGATGATCTCGAATAGCGTGCGCGAATGAT...

PM: ATGATCTCGAATAGCGTGCGCGAAT
MM: ATGATCTCGAATTTGCATCGCGCGAAT

♦ PM>>MM ⇒ gene expressed
PM≈MM ⇒ gene not expressed
Data
Probe set for gene $g$ in array $i$: $\{(PM_{ij}, MM_{ij}) : j$-probes$\} \rightarrow \theta_i$

Single array methods
- **Average Difference**: $\theta_i = \text{Mean}_j (PM_{ij} - MM_{ij})$
- **MAS5**: $\theta_i = \text{BiweightMean}_j (PM_{ij} - MM_{ij})$

Multi-array methods
- **Li-Wong model**: $MM_{ij} = \nu_j + \theta_i \alpha_j + \epsilon_{ij}$ and $PM_{ij} = \nu_j + \theta_i (\alpha_j + \phi_j) + \epsilon_{ij}$
- **Reduced Li-Wong model**: $Y_{ij} = PM_{ij} - MM_{ij} = \theta_i \phi_j + \epsilon_{ij}$
- **Linear model**: $\log(PM_{ij} - MM_{ij}) = \log(\theta_i) + \log(\phi_j) + \epsilon_{ij}$
Affymetrix summarization schemes (ctd)

- **RMA (robust multichip analysis):**
\[
\log(n(PM_{ij} - BG_i)) = \theta_i + \phi_j + \varepsilon_{ij}
\]

- **GC-RMA:**
\[
PM_{ij} = OPM_{ij} + NPM_{ij} + \theta_i \text{ and } MM_{ij} = OMM_{ij} + NMM_{ij} + \phi \theta_i.
\]

Signal: \( S = \text{quantity proportional to mRNA} \)

Proportion of signal picked up by MM: \( 0 < \phi < 1 \)

Optical noise: \( \log(OPM) \sim \text{normal}, \log(OMM) \sim \text{normal} \)

Nonspecific binding: \( (\log(NPM), \log(NMM)) \sim \text{normal} \)

- **FARMS:**
Factor analysis model
\[
\log(PM_{ij}) = (\theta_i \sigma + \mu) + (\theta_i \tau_j + \gamma_j) + \varepsilon_{ij}
\]
<table>
<thead>
<tr>
<th>Gene</th>
<th>Row</th>
<th>Col</th>
<th>Signal</th>
<th>Backgd</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>1</td>
<td>1</td>
<td>995</td>
<td>691</td>
<td>Good</td>
</tr>
<tr>
<td>G2</td>
<td>1</td>
<td>2</td>
<td>8742</td>
<td>696</td>
<td>Good</td>
</tr>
<tr>
<td>G3</td>
<td>1</td>
<td>3</td>
<td>2081</td>
<td>689</td>
<td>Good</td>
</tr>
<tr>
<td>G4</td>
<td>1</td>
<td>4</td>
<td>1284</td>
<td>693</td>
<td>Good</td>
</tr>
<tr>
<td>G5</td>
<td>1</td>
<td>5</td>
<td>1022</td>
<td>685</td>
<td>Good</td>
</tr>
<tr>
<td>G6</td>
<td>1</td>
<td>6</td>
<td>4410</td>
<td>683</td>
<td>Good</td>
</tr>
<tr>
<td>G7</td>
<td>1</td>
<td>7</td>
<td>1664</td>
<td>688</td>
<td>Good</td>
</tr>
<tr>
<td>G8</td>
<td>1</td>
<td>8</td>
<td>1052</td>
<td>680</td>
<td>Bad</td>
</tr>
<tr>
<td>G9</td>
<td>1</td>
<td>9</td>
<td>681</td>
<td>679</td>
<td>Weak</td>
</tr>
<tr>
<td>G10</td>
<td>1</td>
<td>10</td>
<td>963</td>
<td>676</td>
<td>Good</td>
</tr>
</tbody>
</table>

......
Image plot of a good array

Signal

Background
Image plot of a defective array

Signal

Background
### Case study

- Three groups of five patients each.
- RNA applied to microarray with 22,944 genes.

<table>
<thead>
<tr>
<th>Control Group:</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1 Group:</td>
<td>A1</td>
<td>A2</td>
<td>A3</td>
<td>A4</td>
<td>A5</td>
</tr>
<tr>
<td>Test 2 Group:</td>
<td>B1</td>
<td>B2</td>
<td>B3</td>
<td>B4</td>
<td>B5</td>
</tr>
</tbody>
</table>
### Single-channel spot intensity data

<table>
<thead>
<tr>
<th>Gene</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
</tr>
</thead>
<tbody>
<tr>
<td>G8521</td>
<td>1003</td>
<td>1306</td>
<td>713</td>
<td>1628</td>
<td>1268</td>
<td>1629</td>
</tr>
<tr>
<td>G8522</td>
<td>890</td>
<td>705</td>
<td>566</td>
<td>975</td>
<td>883</td>
<td>1005</td>
</tr>
<tr>
<td>G8523</td>
<td>680</td>
<td>749</td>
<td>811</td>
<td>669</td>
<td>724</td>
<td>643</td>
</tr>
<tr>
<td>G8524</td>
<td>262</td>
<td>311</td>
<td>336</td>
<td>1677</td>
<td>1286</td>
<td>1486</td>
</tr>
<tr>
<td>G8525</td>
<td>254</td>
<td>383</td>
<td>258</td>
<td>1652</td>
<td>1799</td>
<td>1645</td>
</tr>
<tr>
<td>G8526</td>
<td>81</td>
<td>140</td>
<td>288</td>
<td>298</td>
<td>241</td>
<td>342</td>
</tr>
<tr>
<td>G8527</td>
<td>4077</td>
<td>2557</td>
<td>2600</td>
<td>3394</td>
<td>2926</td>
<td>2755</td>
</tr>
<tr>
<td>G8528</td>
<td>2571</td>
<td>1929</td>
<td>1406</td>
<td>2439</td>
<td>1613</td>
<td>5074</td>
</tr>
<tr>
<td>G8529</td>
<td>55</td>
<td>73</td>
<td>121</td>
<td>22</td>
<td>141</td>
<td>24</td>
</tr>
<tr>
<td>G8530</td>
<td>1640</td>
<td>1693</td>
<td>1517</td>
<td>1731</td>
<td>1861</td>
<td>1550</td>
</tr>
<tr>
<td>G8531</td>
<td>168</td>
<td>229</td>
<td>284</td>
<td>220</td>
<td>310</td>
<td>335</td>
</tr>
<tr>
<td>G8532</td>
<td>323</td>
<td>258</td>
<td>359</td>
<td>345</td>
<td>308</td>
<td>325</td>
</tr>
<tr>
<td>G8533</td>
<td>12131</td>
<td>11199</td>
<td>14859</td>
<td>11544</td>
<td>11352</td>
<td>11506</td>
</tr>
<tr>
<td>G8534</td>
<td>2007</td>
<td>1878</td>
<td>1502</td>
<td>1758</td>
<td>2480</td>
<td>1731</td>
</tr>
</tbody>
</table>

...
Convert scanned image to spotted image

Run initial check of data quality

Adjust for background

Transform data

Normalize data

Run final check of data quality
Check array quality

Check consistency across arrays

- Scatterplot matrix
- Boxplots
- **Concordance correlation**
  (measures degree of agreement)

\[
\text{med}[\rho_c] = [0.90, 0.92, 0.95, 0.91, \ldots]
\]

- **Spearman correlation coefficient**
  (measures the degree of monotonicity, preservation of rank order)

\[
\text{med}[\rho_s] = [0.94, 0.97, 0.97, 0.93, \ldots]
\]
- Convert scanned image to spotted image
- Run initial check of data quality
- Adjust for background
- Transform data
- Normalize data
- Run final check of data quality
The signal at a particular spot is taken to be

\[ X_g \leftarrow \text{SpotIntensity} \]

or \[ X_g \leftarrow \text{SpotIntensity} - \text{Background} \]

or \[ X_g \leftarrow \text{SpotIntensity} - \text{SmoothedBackground} \]
Convert scanned image to spotted image
Run initial check of data quality
Adjust for background
Transform data
Normalize data
Run final check of data quality
Transformation

♦ Take logs (makes the range of the data more manageable, symmetrizes the within-gene distribution but does not eliminate the heterogeneity of variances across genes, reduces but does not eliminate the skewness of the across-gene distribution).

\[ X \leftarrow \log(X+\lambda) \text{ with } \lambda=0 \]

(Alternative: variance stabilizing transformation)
Convert scanned image to spotted image

Run initial check of data quality

Adjust for background

Transform data

Normalize data

Run final check of data quality
Normalization

♦ Often the signals on even identical microarrays tend to be on different scales (due to quality and quantity of RNA, labeling efficiency, scanner setting, etc) - this can be regarded as a sort of (nonlinear but monotone) array effect.

♦ The scales need to be normalized prior to further analysis, so that the arrays are on more directly comparable scales.
Two arrays

\( \rho \) (Concordance) = 0.90, \( \rho \) (Spearman) = 0.97
Normalization

To normalize arrays $C(1), \ldots, C(n)$:

♦ Calculate the “median mock array” $M$.

♦ Either use lowess (or spline) smoother to model the relationship between $C(i)$ and $M$ or fit a continuous monotone increasing function to the quantiles of $C(i)$ vs the quantiles of $M$.

♦ Back-predict to obtain the normalized values of $C(i)$. 
Two arrays (after normalization)

ρ (Concordance) = 0.98, ρ (Spearman) = 0.97
- Convert scanned image to spotted image
- Run initial check of data quality
- Adjust for background
- Transform data
- Normalize data
- Run final check of data quality
Final array quality check

Check consistency

♦ Scatterplot matrix
♦ Boxplots
♦ Concordance correlation coefficients

\[ \text{med}[\rho_c] = [0.90, 0.92, 0.95, 0.91, \ldots] \]
\[ \rightarrow \text{med}[\rho_c] = [0.98, 0.98, 0.96, 0.94, \ldots] \]

♦ Spearman correlation coefficients

\[ \text{med}[\rho_s] = [0.94, 0.97, 0.97, 0.93, \ldots] \]
\[ \rightarrow \text{med}[\rho_s] = [0.94, 0.97, 0.97, 0.93, \ldots] \]
Spectral map

♦ Principal components analysis / Biplot / Spectral map

\[ X = \text{matrix (} G \times N \text{) of rank } r \]

\[ X = U D V' \quad \text{(singular value decomposition)} \]

\[ D = \text{diag}(d_i) \text{ with } d_1 > d_2 > d_3 > \ldots > d_r > 0 \]

\[ U \text{ (and } V) = \text{eigenvectors of } XX' \text{ (and } X'X) \]

\[ X \cong X_2 = U_2 D_2 V_2' = [U_2 D_2^\alpha][V_2 D_2^{1-\alpha}]' = G_2 H_2' \]

BI PLOT: Plot the \( G \) \( 2 \)-vectors of \( G_2 \) and the \( N \) \( 2 \)-vectors of \( H_2 \) on the same graph

SPECTRAL MAP: Biplot of standardized \( X \).
Spectral map (example)
- Convert scanned image to spotted image
- Run initial check of data quality
- Adjust for background
- Transform data
- Normalize data
- Run final check of data quality
<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
</tr>
</thead>
<tbody>
<tr>
<td>G8521</td>
<td>6.89</td>
<td>7.18</td>
<td>6.60</td>
<td>7.40</td>
<td>7.15</td>
<td>7.40</td>
</tr>
<tr>
<td>G8522</td>
<td>6.78</td>
<td>6.55</td>
<td>6.37</td>
<td>6.89</td>
<td>6.78</td>
<td>6.92</td>
</tr>
<tr>
<td>G8523</td>
<td>6.52</td>
<td>6.61</td>
<td>6.72</td>
<td>6.51</td>
<td>6.59</td>
<td>6.46</td>
</tr>
<tr>
<td>G8524</td>
<td>5.67</td>
<td>5.69</td>
<td>5.88</td>
<td>7.43</td>
<td>7.16</td>
<td>7.31</td>
</tr>
<tr>
<td>G8525</td>
<td>5.64</td>
<td>5.91</td>
<td>5.61</td>
<td>7.41</td>
<td>7.49</td>
<td>7.41</td>
</tr>
<tr>
<td>G8526</td>
<td>4.63</td>
<td>4.85</td>
<td>5.72</td>
<td>5.71</td>
<td>5.47</td>
<td>5.79</td>
</tr>
<tr>
<td>G8527</td>
<td>8.28</td>
<td>7.88</td>
<td>7.84</td>
<td>8.12</td>
<td>7.99</td>
<td>7.97</td>
</tr>
<tr>
<td></td>
<td>7.81</td>
<td>7.58</td>
<td>7.24</td>
<td>7.79</td>
<td>7.38</td>
<td>8.60</td>
</tr>
<tr>
<td>G8530</td>
<td>7.36</td>
<td>7.45</td>
<td>7.31</td>
<td>7.46</td>
<td>7.53</td>
<td>7.35</td>
</tr>
<tr>
<td>G8531</td>
<td>5.30</td>
<td>5.36</td>
<td>5.70</td>
<td>5.41</td>
<td>5.73</td>
<td>5.77</td>
</tr>
<tr>
<td>G8532</td>
<td>5.84</td>
<td>5.48</td>
<td>5.93</td>
<td>5.81</td>
<td>5.73</td>
<td>5.73</td>
</tr>
<tr>
<td>G8534</td>
<td>7.57</td>
<td>7.55</td>
<td>7.30</td>
<td>7.48</td>
<td>7.82</td>
<td>7.46</td>
</tr>
</tbody>
</table>

$\rho_s, \rho_{cc}, ...$
Other preprocessing

♦ In some cases, other issues need to be addressed as well: e.g., when there are other high level sources of bias and variability such as hybridization date/time, print batch, operator. Check for global and gene-specific extraneous effects. Use linear modeling (global or gene-specific) to adjust for extraneous effects.

♦ Missing values, individual outliers, outlier arrays may also have to be dealt with.

♦ Pre-experiment: Design experiments using standard DOE principles (such as replication, balance, randomization).
Objectives of microarray experiments

♦ **Differential expression:** Identify those genes that are differentially expressed across two or more predefined classes.

♦ **Class prediction:** Develop multi-gene predictor ("signature") of class.

♦ **Pattern discovery:** Discover clusters among samples and/or genes.
Finding differentially expressing genes

♦ Identify those genes that are differentially expressed across two or more predefined classes (can compare gene expression patterns across classes multiple genes at a time):
  o Which genes are expressed in which cells and under what conditions.
  o Which genes are expressed differently in diseased cells compared to normal cells.
  o Which genes are expressed differently when a patient is administered a drug.
Data analysis approaches (1)

♦ **Fold change**: Seek genes that exhibit at least a certain specified fold increase or decrease in mean expression level (e.g., 5-fold).
♦ Statistical analysis of individual genes:
Seek genes that exhibit a statistically significant difference across the groups
(via e.g., t, permutation test, Ct, SAM, limma, Bayes/EmpiricalBayes procedures).

♦ Adjust for multiplicity:
pFDR = Average ( #FalsePositives / #Positives ).
Decision rule says “reject if $T>c$” $\rightarrow h_0$
$\rightarrow$ permute $\rightarrow h_1$ $\rightarrow$ permute $\rightarrow h_2$ $\rightarrow$ … $\rightarrow h_m$
$\rightarrow$ average=$h^*$ $\rightarrow$ pFDR=$h^*/h_0$ $\rightarrow$ refine
Data analysis approaches (3)

♦ Analysis of gene combinations: Seek combinations of genes that separate the groups (via e.g., lda, random forest).
Data analysis approaches (4)

♦ Incorporate biological information: Seek significant gene functions.

<table>
<thead>
<tr>
<th>GO #</th>
<th>GO term</th>
<th>Gene</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>reproduction</td>
<td>Clu</td>
<td>0.0109</td>
</tr>
<tr>
<td>3</td>
<td>reproduction</td>
<td>Mst1</td>
<td>0.6563</td>
</tr>
<tr>
<td>3</td>
<td>reproduction</td>
<td>Ppp2ca</td>
<td>0.0290</td>
</tr>
<tr>
<td>3</td>
<td>reproduction</td>
<td>Prkag1</td>
<td>0.9026</td>
</tr>
<tr>
<td>3</td>
<td>reproduction</td>
<td>Serpina1</td>
<td>0.6514</td>
</tr>
</tbody>
</table>

Gene function scoring: \( \{p\} \rightarrow \text{mean (-log } p\) = 1.8023*  

Over-representation analysis: 2/5 vs 167/12567 \( \rightarrow \) odds ratio=1.56*  

Significance is determined by permuting the \( p \)-values across the genes (thereby preserving correlations among gene sets) and calculating the proportion of artificial gene sets of the same size that have a test statistic value greater than or equal to that observed.
Assess biological significance

♦ Data analysis ➔ list of differentially expressed genes ➔ ?

♦ Confirm by RT-PCR or similar technique.

♦ Biological and bioinformatics considerations.
Summary of challenges for data analysis

♦ Explosive growth in dataset size, but ...
  • not necessarily very many replicates
  • instead many many variables
♦ How to explore very high-dimensional space → dimension reduction by variable and view selection
♦ Few replicates → borrow strength
♦ Data not Normal and not all that clean → quality checks, semi-parametric methods
♦ Computational considerations
♦ How to validate/confirm findings and avoid a large number of false discoveries
Wrap up

♦ Reference:

♦ Website:
www.geocities.com/damaratung
http://www.rci.rutgers.edu/~cabrera/DNAMR

♦ Email:
damaratu@prdus.jnj.com
cabrera@stat.rutgers.edu

Thank you!