GEPAS

Microarray Data Analysis
Normalization

Cambridge, September 2009

Javier Santoyo-Lopez
jsantoyo@cipf.es
http://bioinfo.cipf.es

Bioinformatics and Genomics Department
Centro de Investigacion Principe Felipe (CIPF)
Valencia, Spain
DNA Microarrays

- Paradigm of high Throughput Technologies
- Measure gene expression (our variables)
  - of different cells
  - under different biological conditions
  - in a genomic scale
- Allow us to conduct biological experiments

So. How do they work?
Central Dogma of Molecular Biology

For a cell, at a particular time, thousands of mRNA are created and sent out of the nucleus to be translated into proteins.

Protein concentration regulates biological systems.
Human Genome Project

- Determined the sequences that make up human genome. (3 billion base pairs)
- Identified the sequence of all genes in human DNA. (20,000 – 25,000)
- Now we know the complementary sequence of all mRNA molecules that could be transcribed in any biological process.
DNA Microarrays

1.28 cm
Actual size of GeneChip® array

ATCATG
ATCATG
ATCATG
ATCATG

ATCATG
ATCATG
ATCATG
ATCATG

500,000 locations on each GeneChip® array

Millions of DNA strands built up in each location

Actual strand = 25 base pairs
Biological Sample / Cell Culture

We want to know which genes are expressed under particular biological conditions.

We can extract all mRNA molecules that are being translated within the cells and provide an expression level indicator of its concentration in the biological sample.
RNA Extraction

**BREAKING CELLS AND TISSUES**
The first step in the purification of most proteins is to disrupt tissues and cells in a controlled fashion.

- **Using gentle mechanical procedures, called homogenization**, the plasma membranes of cells can be ruptured so that the cell contents are released. Four commonly used procedures are shown here.
  - 1. **Break cells with high frequency sound**
  - 2. **Use a mild detergent to make holes** in the plasma membrane
  - 3. **Force cells through a small hole using high pressure**
  - 4. **Shear cells between a close-fitting rotating plunger and the thick walls of a glass vessel**

**THE CENTRIFUGE**

- **The centrifuge** is used to separate homogenates into parts or fractions. The homogenate is placed in test tubes and rotated at high speed in a centrifuge (sometimes called an ultracentrifuge). Present-day ultracentrifuges rotate at speeds up to 100,000 revolutions per minute and produce enormous forces, as high as 600,000 times gravity. At such speeds, centrifuge chambers must be refrigerated and evacuated so that friction does not heat up the homogenate. The centrifuge is surrounded by thick armor plating, since an unbalanced rotor can shatter with an explosive release of energy. A fixed-angle rotor can hold...

**DIFFERENTIAL CENTRIFUGATION**
Repeated centrifugation at progressively higher speeds will fractionate cell homogenates into their components.

- **Supernatant 1** (whole cells, nuclei, cytoplasmic isolation)
- **Pellet 1** (mitochondria, lysosomes, peroxisomes)
- **Supernatant 2** (microsomes, other small vacuoles)
- **Pellet 2** (microsomes, other small vacuoles)
- **Supernatant 3** (ribosomes, large macromolecules)
- **Pellet 3** (ribosomes, large macromolecules)

**VELOCITY SEDIMENTATION**

- Subcellular components sediment at different speeds according to their size when carefully layered over a dilute salt solution. In order to stabilize the sedimenting components against convective mixing in the tube, the solution contains a continuous shallow gradient of sucrose that increases in concentration toward the bottom of the tube. This is typically 5-20% sucrose. When sedimented through such a dilute sucrose gradient, different cell components separate into distinct bands that can, after an appropriate time, be collected individually.

**EQUILIBRIUM SEDIMENTATION**

- The ultracentrifuge can also be used to separate cellular components on the basis of their buoyant density, independently of their size or shape. The sample is usually either layered on top of, or dispersed within, a step density gradient that contains a very high concentration of sucrose or cesium chloride. Each subcellular component will move up or down when centrifuged until it reaches a position where its density matches its surroundings and then will move no further. A series of distinct bands will eventually be produced, with those nearest the bottom of the tube containing the components of highest buoyant density. This method is also called density gradient centrifugation...

**CENTRIFUGATION**

- **Before Centrifugation**
- **After Centrifugation**

The final bands can be collected from the base of the tube, as shown above.
Labeling the Sample
Labeling the Sample

Fluorescent Dye
Hybridization
Expression Measurement

If the fluorescent label is attached to one spot we know that the particular complementary gene transcript was present in our cell sample.

The greater the fluorescence the greater the concentration of the transcript.
We get a measurement of the hybridization in each spot of the microarray.
The Data

For each biological sample (individual)

We get intensity measurements for thousands of gene transcripts.

The measured intensity is used as an indicator of gene expression.
Several Microarrays

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Log Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>200000_s_at</td>
<td>134.4</td>
</tr>
<tr>
<td>200001_at</td>
<td>586.5</td>
</tr>
<tr>
<td>200002_at</td>
<td>1868.4</td>
</tr>
<tr>
<td>200003_s_at</td>
<td>1232.7</td>
</tr>
<tr>
<td>200004_at</td>
<td>1071.6</td>
</tr>
<tr>
<td>200005_at</td>
<td>312.8</td>
</tr>
<tr>
<td>200006_at</td>
<td>1712.6</td>
</tr>
<tr>
<td>200007_at</td>
<td>606.5</td>
</tr>
<tr>
<td>200008_s_at</td>
<td>421.9</td>
</tr>
<tr>
<td>200009_at</td>
<td>395.6</td>
</tr>
<tr>
<td>200010_at</td>
<td>1228.6</td>
</tr>
<tr>
<td>200011_s_at</td>
<td>132.5</td>
</tr>
<tr>
<td>200012_x_at</td>
<td>2606.3</td>
</tr>
<tr>
<td>200013_at</td>
<td>1572.9</td>
</tr>
<tr>
<td>200014_s_at</td>
<td>138.7</td>
</tr>
<tr>
<td>200015_s_at</td>
<td>124.1</td>
</tr>
<tr>
<td>200016_x_at</td>
<td>1058.7</td>
</tr>
<tr>
<td>200017_at</td>
<td>889.4</td>
</tr>
<tr>
<td>200018_at</td>
<td>3964.2</td>
</tr>
<tr>
<td>200019_s_at</td>
<td>1069.9</td>
</tr>
<tr>
<td>200020_at</td>
<td>212.1</td>
</tr>
<tr>
<td>200021_at</td>
<td>1018.1</td>
</tr>
<tr>
<td>200022_at</td>
<td>1254.8</td>
</tr>
<tr>
<td>200023_s_at</td>
<td>1202.8</td>
</tr>
<tr>
<td>200024_at</td>
<td>2460.6</td>
</tr>
</tbody>
</table>
### Several Microarrays

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>200000_s_at</td>
<td>134.4</td>
</tr>
<tr>
<td>200001_at</td>
<td>586.5</td>
</tr>
<tr>
<td>200002_s_at</td>
<td>421.9</td>
</tr>
<tr>
<td>200003_s_at</td>
<td>1232.7</td>
</tr>
<tr>
<td>200004_at</td>
<td>1712.6</td>
</tr>
<tr>
<td>200005_at</td>
<td>606.5</td>
</tr>
<tr>
<td>200006_at</td>
<td>1228.6</td>
</tr>
<tr>
<td>200007_at</td>
<td>132.5</td>
</tr>
<tr>
<td>200008_x_at</td>
<td>2606.3</td>
</tr>
<tr>
<td>200009_at</td>
<td>395.6</td>
</tr>
<tr>
<td>200010_at</td>
<td>1288.6</td>
</tr>
<tr>
<td>200011_s_at</td>
<td>132.5</td>
</tr>
<tr>
<td>200012_x_at</td>
<td>1572.9</td>
</tr>
<tr>
<td>200013_at</td>
<td>138.7</td>
</tr>
<tr>
<td>200014_s_at</td>
<td>124.1</td>
</tr>
<tr>
<td>200015_s_at</td>
<td>1058.7</td>
</tr>
<tr>
<td>200016_x_at</td>
<td>889.4</td>
</tr>
<tr>
<td>200017_at</td>
<td>3964.2</td>
</tr>
<tr>
<td>200018_s_at</td>
<td>1069.9</td>
</tr>
<tr>
<td>200020_at</td>
<td>212.1</td>
</tr>
<tr>
<td>200021_at</td>
<td>1018.1</td>
</tr>
<tr>
<td>200022_at</td>
<td>1254.8</td>
</tr>
<tr>
<td>200023_s_at</td>
<td>1202.8</td>
</tr>
<tr>
<td>200024_at</td>
<td>2460.6</td>
</tr>
<tr>
<td>200000_s_at</td>
<td>134.4</td>
</tr>
<tr>
<td>200001_at</td>
<td>586.5</td>
</tr>
<tr>
<td>200002_s_at</td>
<td>421.9</td>
</tr>
<tr>
<td>200003_s_at</td>
<td>1232.7</td>
</tr>
<tr>
<td>200004_at</td>
<td>1712.6</td>
</tr>
<tr>
<td>200005_at</td>
<td>606.5</td>
</tr>
<tr>
<td>200006_at</td>
<td>1228.6</td>
</tr>
<tr>
<td>200007_at</td>
<td>132.5</td>
</tr>
<tr>
<td>200008_x_at</td>
<td>2606.3</td>
</tr>
<tr>
<td>200009_at</td>
<td>395.6</td>
</tr>
<tr>
<td>200010_at</td>
<td>1288.6</td>
</tr>
<tr>
<td>200011_s_at</td>
<td>132.5</td>
</tr>
<tr>
<td>200012_x_at</td>
<td>1572.9</td>
</tr>
<tr>
<td>200013_at</td>
<td>138.7</td>
</tr>
<tr>
<td>200014_s_at</td>
<td>124.1</td>
</tr>
<tr>
<td>200015_s_at</td>
<td>1058.7</td>
</tr>
<tr>
<td>200016_x_at</td>
<td>889.4</td>
</tr>
<tr>
<td>200017_at</td>
<td>3964.2</td>
</tr>
<tr>
<td>200018_s_at</td>
<td>1069.9</td>
</tr>
<tr>
<td>200020_at</td>
<td>212.1</td>
</tr>
<tr>
<td>200021_at</td>
<td>1018.1</td>
</tr>
<tr>
<td>200022_at</td>
<td>1254.8</td>
</tr>
<tr>
<td>200023_s_at</td>
<td>1202.8</td>
</tr>
<tr>
<td>200024_at</td>
<td>2460.6</td>
</tr>
</tbody>
</table>

**Note:** The values represent expression levels or other measurements associated with each gene.
Several Microarrays

<table>
<thead>
<tr>
<th></th>
<th>Array1</th>
<th>Array2</th>
<th>Array3</th>
<th>Array4</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene1</td>
<td>10.23</td>
<td>9.98</td>
<td>10.41</td>
<td>10.55</td>
<td>10.65</td>
</tr>
<tr>
<td>gene2</td>
<td>10.51</td>
<td>9.74</td>
<td>10.65</td>
<td>10.63</td>
<td>10.43</td>
</tr>
<tr>
<td>gene3</td>
<td>9.89</td>
<td>10.02</td>
<td>9.89</td>
<td>11.03</td>
<td>10.21</td>
</tr>
<tr>
<td>gene4</td>
<td>10.25</td>
<td>10.83</td>
<td>8.94</td>
<td>10.16</td>
<td>10.49</td>
</tr>
<tr>
<td>gene...</td>
<td>.......</td>
<td>.......</td>
<td>.......</td>
<td>.......</td>
<td>.......</td>
</tr>
</tbody>
</table>
Single channel hybridization.

- Each slide is hybridized with a single biological sample labelled with a unique dye.

- Most new technologies follow this approach: Affymetrix, Agilent, Codelink.

- Measured fluorescent intensities ideally represent transcript levels in the sample. Nevertheless these are not calibrated.
Competitive hybridization

Two different biological samples

Green labelled sample

Red labelled sample

Hybridized in the same slide

Synthetic image

\[
\log_2 \left( \frac{\text{Red}}{\text{Green}} \right)
\]
Competitive hybridization

- Each slide is hybridized with two biological samples, each one labeled with a different dye.

- Log ratios of the two color intensities ideally represent the relative abundance of the transcripts in one sample compared to the transcripts in the other one.
Affymetrix technology
Affymetrix technology
Spotted arrays

Print head

Printing spots in blocks
Main-grid / Sub-grid

duplicate spots per probe, both in same sub-grid

spot-to-spot spacing 230 micron
all measurements are approximate
Image processing - Spot recognition
Image processing - Spot recognition

- Mean measured intensity whiting the spot
- Median measured intensity around the spot
That is a noisy measurement process

- Irregularities on the array surface
- Variations in the laboratory process
- Different scanner settings
- Dye effect
- Different DNA strands have different hybridization properties
- Several platforms
- Random noise
Background Effects
Scale Effects

A.CEL

B.CEL
Problem

+ intensity $\iff$ + hybridization $\iff$ + DNA/RNA concentration in the sample

• Every single spot in the array is measured in its own particular scale with different:
  – origin of the measurements
  – unit of change

• We cannot compare data that have been measured in different scales.

• First we need to calibrate the scales we use. They must have the same origin and unit of measurement.
Problems of measuring using different scales

- When comparing intensities from different spots within the same array
- When comparing intensities from the same spot between several arrays
Normalization starts here
Objective

• Achieve a measurement scale such that
  – It has the **same origin** (zero or other) for all spots
  – It uses the **same unit** for all spots and microarrays (homogeneous random)
  – It has a **linear relationship** with the DNA/RNA biological
  – It has **good statistical properties** (good for later analyses)

• Deal with the particular characteristics of each platform and experiment
  – Color differences
  – Reference sample
  – Summarize information of each gene
  – Deal with the Affymetrix PM-MM
Hypotheses

• Most normalization methodologies make two major assumptions about the data.

  – When comparing different samples, only few genes are over-expressed or under-expressed in one array relative to the others.
  – The number of genes over-expressed in a condition is similar to the number of genes under-expressed.

• This assumptions should agree with your experimental context.
Assessing the data. Some plots.

• Plots are useful tools to analyze both, raw and normalized microarray data.

• We use them to:
  – Unravel artifacts in the raw data which are not due to biological reasons.
  – Assess whether the normalization steps have succeeded in correcting them.
Box plots and histograms
Objective

MA plots & loess curves

\[ M = \log R - \log G = \log \left( \frac{R}{G} \right) \]

\[ A = \left( \frac{\log R + \log G}{2} \right) = \log \sqrt{R \cdot G} \]
General Steps

• **Background correction** (correcting the scale origin for spots)

• **Normalization** (standardizing the scale unit - rescaling)

• **Adjustments characteristics of each platform or experiment**
  – Perfect-Match Mismatch Adjustment (Affymetrix)
  – Correcting for different dye properties (in two color arrays)
  – Adjustments depending on the DNA strands

• **Summary of information from several spots into a single measure for each gene**
  – Averaging Affymetrix "probe sets"
  – Averaging duplicated spots
  – Calculating ratios
  – Taking logarithms
Background correction

Observed background intensities (HGMP 2b Cy5)

Observed spot intensities (HGMP 2b Cy5)
General Steps

• Background correction (correcting the scale origin)

• **Normalization** (correcting the scale unit - rescaling)

• Adjustments characteristics of each platform or experiment
  – Perfect-Match Mismatch Adjustment (Affymetrix)
  – Correcting for different dye properties (in two color arrays)
  – Adjustments depending on the DNA strands

• Summary of information from several spots into a single measure for each gene
  – Averaging Affymetrix "probe sets"
  – Averaging duplicated spots
  – Calculating ratios
  – Taking logarithms
Normalization
(rescaling & centering)
Normalization
(rescaling & centering)
General Steps

- Background correction (correcting the scale origin)
- Normalization (correcting the scale unit - rescaling)
- Adjustments characteristics of each platform or experiment
  - Perfect-Match Mismatch Adjustment (Affymetrix)
  - Correcting for different dye properties (in two color arrays)
  - Adjustments depending on the DNA strands
- Summary of information from several spots into a single measure for each gene
  - Averaging Affymetrix "probe sets"
  - Averaging duplicated spots
  - Calculating ratios
  - Taking logarithms
Probe-sets; PM - MM
Perfect Match – Mismatch

- Perfect Match probe has a known existent sequence
  `CCCTTACCCAGTCTTCCGGAGGCTA`
- Mismatch changes one base in the sequence
  `CCCTTACCCAGTGTTTCCCGGAGGCTA`
  Intended to correct for cross-hybridization

  Intended to correct for cross-hybridization
Affymetrix technology
DIGESTION plots
General Steps

• Background correction (correcting the scale origin)

• Normalization (correcting the scale unit - rescaling)

• Adjustments characteristics of each platform or experiment
  – Perfect-Match Mismatch Adjustment (Affymetrix)
  – Correcting for different dye properties (in two color arrays)
  – Adjustments depending on the DNA strands

• Summary of information from several spots into a single measure for each gene
  – Averaging Affymetrix "probe sets"
  – Averaging duplicated spots
  – Calculating ratios
  – Taking logarithms
MA plots & loess curves

\[ M = \log R - \log G = \log \left( \frac{R}{G} \right) \]

\[ A = \left( \frac{\log R + \log G}{2} \right) = \log \sqrt{R \cdot G} \]
Main-grid / Sub-grid

duplicate spots per probe, both in same sub-grid

spot-to-spot spacing 230 micron
all measurements are approximate
MA plots & print-tip loess curves

One loess line for each block (print-tip block)
Normalization of Affymetrix data

Affymetrix methods are those implemented in the *affy* package from Bioconductor.

- **Background correction:**
  - Rma, gcrma, mas.

- **Between arrays (rescaling):**
  - Quantiles, constant, spline, invariantset, loess, quantiles.robust.

- **PM-MM adjustment:**
  - Pmonly, mas, subtractmm.

- **Probe-set summary method:**
  - Medianpolish, avgdiff, liwong, mas.
Normalization of Two Color arrays

Two colors methods are those implemented in the limma package from Bioconductor.

- **Background correction:**
  - Rma, subtract, half, edwards.

- **Within array correction:**
  - Print tip loess, loess, median.

- **Between arrays standardization:**
  - Quantile, scale.