GEPAS
Microarray Data Analysis
Normalization

May 2009 in Valencia
Ignacio Medina
imedina@cipf.es
http://bioinfo.cipf.es/imedina
Bioinformatics and Genomics Department
Centro de Investigacion Principe Felipe (CIPF)
(Valencia, Spain)
GEPAS/Babelomics
Definitions

- **GEPAS**
  - The *Gene Expression Profile Analysis Suite* is one of the most complete integrated packages of tools for *microarray data analysis* available over the web

- **Babelomics**
  - Babelomics is a complete suite of web tools for *functional analysis* of genome-scale experiments
GEPAS/Babelomics

Definitions

- **GEPAS**
  - The *Gene Expression Profile Analysis Suite* is one of the most complete integrated packages of tools for *microarray data analysis* available over the web

- **Babelomics**
  - Babelomics is a complete suite of web tools for *functional analysis* of genome-scale experiments
GEPA/Babelomics Pipeline

Herrero et al., 2003, 2004; Vazquez et al., 2005 NAR; Montaner et al., 2006 NAR; Al-Shahrour et al., 2003, 2006 NAR; 2005 Bioinformatics
GEPAS Pipeline

Herrero et al., 2003, 2004; Vaquerrizas et al., 2005 NAR; Montaner et al., 2006 NAR; Al-Shahour et al., 2003, 2006 NAR; 2005 Bioinformatics
GEPAS
Normalization

![GEPAS Normalization Form](image-url)

**Normalization Form**: Select file from computer, specify background correction method (e.g., `rma`), and enter job name (e.g., `affymetrix`).
GEPAS
Preprocessing

Preprocessing : form

- Select file from your computer
- Log-transform
  - Log base
- Merge replicates
  - Merge
  - Average
  - Median
- Filter missing values
  - Minimum percentage of existing values (%)
- Impute missing values
  - Impute
  - Fill with zeros
  - Fill with row average
  - Fill with row median
  - Use KNN impute
  - K value
  - Filter
GEPA S
Clustering

GEPA S
Gene Expression Pattern Analysis Suite

Clustering : form

[Form options and fields]

Hierarchical

Non-Hierarchical
GEPAS
Differential expression
Gene Expression Pattern Analysis Suite

Tools
- Normalization
- Preprocessing
- Clustering
- Differential expression
- Predictors

On-line examples

Train

Gene expression file
- Select file from your computer
- or select file from the server
- or enter your data

Class file
- Select file from your computer
- or select file from the server
- or enter your data

Number of genes to use in training
- All genes (max. 1000)
- Custom
- Number of genes (separated by blanks)
- 2 5 10 20 35 50 75 100

Gene selection
- F ratio
- Wilcoxon test
- None

Algorithms
- SVM
- KNN
- DLDA
- PAM
- SOM

Job name
- train

Predict

Exam inar...

Bioinformatics. 23(3):390-1
Babelomics

Herrero et al., 2003, 2004; Vequeiras et al., 2005 NAR; Montaner et al., 2006 NAR; Al-Shahrouq et al., 2003, 2006 NAR; 2005 Bioinformatics
1. Microarray Technology
2. Normalization
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 the mRNA molecules that could be transcribed in any biological process.
DNA Microarrays

Actual size of GeneChip® array

500,000 locations on each GeneChip® array

Millions of DNA strands built up in each location

Actual strand = 25 base pairs
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 procedures is to damage 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

When carefully applied, homogenization leaves most of the membrane-bounded organelles intact.

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

- Centrifugation separates cell components on the basis of size and density. The larger and denser components experience the greatest centrifugal force and move most rapidly. They sediment to form a pellet at the bottom of the tube, while smaller, less dense components remain in suspension above, called the supernatant.

LOW SPEED CENTRIFUGATION
- Supernatant 1
- Whole cells
- Nuclei
- Exoskeletons

MEDIUM SPEED CENTRIFUGATION
- Supernatant 2
- Mitochondria
- Lysosomes
- Peroxisomes

HIGH SPEED CENTRIFUGATION
- Supernatant 3
- Endoplasmic reticulum
- Ribosomes

VERY HIGH SPEED CENTRIFUGATION
- Supernatant 4
- Nuclear proteins

VELOCITY SEDIMENTATION
Subcellular components sediment at different speeds according to their size when carefully layered over a dilute salt solution. In order to separate the components effectively, 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 sedimenting 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 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.

At equilibrium, components have migrated to a region in the gradient where they match their own density.

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

We get a measurement of the hybridization in each spot of the microarray.

Measuring the intensity of the fluorescence in each spot
The Data

For each biological sample (individual) we get intensity measurements for thousands of genetic transcripts. The measured intensity is used as an indicator of gene expression.
<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Expression Level</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></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>10.03</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>
1. Microarray Technology
2. Normalization
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

Synthetic image

log$_{2} \left( \frac{\text{Red}}{\text{Green}} \right)$

<table>
<thead>
<tr>
<th>Gene 16</th>
<th>126.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 12</td>
<td>132.5</td>
</tr>
<tr>
<td>Gene 1</td>
<td>134.4</td>
</tr>
<tr>
<td>Gene 10</td>
<td>136.7</td>
</tr>
<tr>
<td>Gene 21</td>
<td>212.1</td>
</tr>
<tr>
<td>Gene 6</td>
<td>342.8</td>
</tr>
<tr>
<td>Gene 18</td>
<td>395.6</td>
</tr>
<tr>
<td>Gene 9</td>
<td>421.3</td>
</tr>
<tr>
<td>Gene 2</td>
<td>596.5</td>
</tr>
<tr>
<td>Gene 8</td>
<td>665.5</td>
</tr>
<tr>
<td>Gene 18</td>
<td>883.4</td>
</tr>
<tr>
<td>Gene 22</td>
<td>1018</td>
</tr>
<tr>
<td>Gene 17</td>
<td>1039</td>
</tr>
<tr>
<td>Gene 29</td>
<td>1070</td>
</tr>
<tr>
<td>Gene 5</td>
<td>1072</td>
</tr>
<tr>
<td>Gene 24</td>
<td>1283</td>
</tr>
<tr>
<td>Gene 11</td>
<td>1229</td>
</tr>
<tr>
<td>Gene 4</td>
<td>1239</td>
</tr>
<tr>
<td>Gene 23</td>
<td>1235</td>
</tr>
</tbody>
</table>

Gene 1 104.4
Gene 2 606.5
Gene 3 1066
Gene 4 1288
Gene 5 1072
Gene 6 512.9
Gene 7 1710
Gene 8 666.5
Gene 9 421.9
Gene 10 665.6
Gene 11 1205
Gene 12 1072
Gene 13 2606
Gene 14 1576
Gene 15 427.1
Gene 16 125.1
Gene 17 1038
Competitive hybridization

• Each slide is hybridized with two biological samples each labelled with a different dye.

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

Print head

Printing spots in blocks
Main-grid / Sub-grid

- **MetaRow**: 54 mm
- **MetaColumn**: 5 mm
- **Row**: 12
- **Column**: 19
- **SCp36 / 1**: 123456789

- **Cy3 landing light**
- **Cy5 landing light**
- **Probe**
- **Empty**
- **16s or 23s rRNA**
- **Scorecard controls**
- **Carry-over or negative controls**

Duplicate spots per probe, both in same sub-grid

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

Foreground
Mean measured intensity within the spot

Background
Median measured intensity around the spot
That is a noisy measurement process

- Irregularities in 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 ↔ + hybridization ↔ + 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

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

\[ A = \sqrt{\frac{\log R + \log G}{2}} = \log \sqrt{R \cdot G} \]
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)
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

CCCTTACCCAGTCTTTCCGGAGGCTA

• Mismatch changes one base in the sequence

CCCTTACCCAGTGTTTCCGGAGGCTA

Intended to correct for cross-hybridization

Intended to correct for cross-hybridization
Affymetrix technology
DIGESTION plots

RNA digestion plot

Mean intensity, shifted and scaled

Array1.CEL
Array2.CEL
Array3.CEL

Probe Number

5' ⏯ 3'
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

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

\[ A = \left( \frac{\log R + \log G}{\gamma} \right) = \log \sqrt{R \cdot G} \]
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)
http://www.gepas.org/