Babelomics: Microarray Data Analysis

Martina Marbà
mmarba@cipf.es

Bioinformatics and Genomics Department
Centro de Investigación Príncipe Felipe, Valencia, Spain

Barcelona, 12 de Marzo de 2012
1 Introduction to Microarrays

2 Tool Overview

3 Analysis

4 Example
Outline

1 Introduction to Microarrays
   - DNA Microarrays
   - Hybridization
   - The Data
   - Problems

2 Tool Overview

3 Analysis

4 Example
DNA Microarrays

- Paradigm of High Throughput Technologies
- Yield concentration measurements for:
  - Genes, SNP, exons, mRNA ...
- Measure cells in different biological conditions
- In a genomic scale
- Allow us conducting 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.
DNA Microarrays

- Introduction to Microarrays
- Tool Overview
- Analysis
- Example

DNA Microarrays

- Hybridization
- The Data
- Problems

DNA Microarrays

- Actual size of GeneChip® array
- Millions of DNA strands built up in each location
- 500,000 locations on each GeneChip® array
- Actual strand = 25 base pairs

Martina Marbà  mmarba@cipf.es

Babelomics: Microarray Data Analysis
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.
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

**Introduction to Microarrays**

**DNA Microarrays**

**Tool Overview**

**Analysis**

**Example**

**Problems**

**RNA extraction**

**Acknowledgments**

**Martina Marbà**

mmarba@cipf.es

**Babelomics: Microarray Data Analysis**
Labelling the Sample

Fluorescent (Dye)
Hybridization

RNA fragment hybridizes with DNA
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

- Measuring the intensity of the fluorescence in each spot, we get a measurement of the hybridization in each spot of the microarray.
The Data: (individual) biological sample

- We get intensity measurements for thousands of genetic transcripts.

- The measured intensity is used as an indicator of gene expression.
The Data: (several) biological sample

<table>
<thead>
<tr>
<th>Sample</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>31307_at</td>
<td>5.53</td>
</tr>
<tr>
<td>31308_at</td>
<td>7.07</td>
</tr>
<tr>
<td>31309_r_at</td>
<td>6.05</td>
</tr>
<tr>
<td>31310_at</td>
<td>7.42</td>
</tr>
<tr>
<td>31311_at</td>
<td>7.77</td>
</tr>
<tr>
<td>31312_at</td>
<td>9.47</td>
</tr>
<tr>
<td>31313_at</td>
<td>8.58</td>
</tr>
<tr>
<td>31314_at</td>
<td>7.14</td>
</tr>
<tr>
<td>31315_at</td>
<td>9.62</td>
</tr>
<tr>
<td>31316_at</td>
<td>5.45</td>
</tr>
<tr>
<td>31317_r_at</td>
<td>10.27</td>
</tr>
<tr>
<td>31318_at</td>
<td>5.7</td>
</tr>
<tr>
<td>31319_at</td>
<td>9.25</td>
</tr>
<tr>
<td>31320_at</td>
<td>11.5</td>
</tr>
<tr>
<td>31321_at</td>
<td>7.79</td>
</tr>
<tr>
<td>31322_at</td>
<td>6.98</td>
</tr>
<tr>
<td>31323_r_at</td>
<td>11.18</td>
</tr>
<tr>
<td>31324_at</td>
<td>7.97</td>
</tr>
<tr>
<td>31325_at</td>
<td>9.53</td>
</tr>
<tr>
<td>31326_at</td>
<td>9.67</td>
</tr>
<tr>
<td>31327_at</td>
<td>6.48</td>
</tr>
<tr>
<td>31328_at</td>
<td>8.92</td>
</tr>
<tr>
<td>31329_at</td>
<td>6.11</td>
</tr>
<tr>
<td>31330_at</td>
<td>14.44</td>
</tr>
<tr>
<td>31331_at</td>
<td>6.3</td>
</tr>
</tbody>
</table>

- DNA Microarrays
- Hybridization
- The Data
- Problems

Martina Marbà  mmarba@cipf.es
Babelomics: Microarray Data Analysis
### The Data: data matrix

<table>
<thead>
<tr>
<th>Gene</th>
<th>Array 1</th>
<th>Array 2</th>
<th>Array 3</th>
<th>Array 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.53</td>
<td>5.66</td>
<td>5.52</td>
<td>5.69</td>
</tr>
<tr>
<td>2</td>
<td>7.07</td>
<td>7.14</td>
<td>7.05</td>
<td>7.19</td>
</tr>
<tr>
<td>3</td>
<td>6.05</td>
<td>5.33</td>
<td>5.35</td>
<td>5.07</td>
</tr>
<tr>
<td>4</td>
<td>7.42</td>
<td>7.02</td>
<td>7.02</td>
<td>7.04</td>
</tr>
<tr>
<td>5</td>
<td>7.77</td>
<td>7.83</td>
<td>7.79</td>
<td>7.75</td>
</tr>
<tr>
<td>6</td>
<td>9.47</td>
<td>9.43</td>
<td>9.34</td>
<td>9.37</td>
</tr>
<tr>
<td>7</td>
<td>8.58</td>
<td>8.67</td>
<td>8.52</td>
<td>8.42</td>
</tr>
<tr>
<td>8</td>
<td>7.14</td>
<td>7.3</td>
<td>7.19</td>
<td>7.27</td>
</tr>
<tr>
<td>10</td>
<td>5.45</td>
<td>5.53</td>
<td>5.3</td>
<td>5.35</td>
</tr>
<tr>
<td>11</td>
<td>10.27</td>
<td>10.75</td>
<td>10.41</td>
<td>10.45</td>
</tr>
<tr>
<td>12</td>
<td>5.7</td>
<td>5.53</td>
<td>5.59</td>
<td>5.58</td>
</tr>
<tr>
<td>13</td>
<td>9.25</td>
<td>9.19</td>
<td>9.24</td>
<td>8.78</td>
</tr>
<tr>
<td>14</td>
<td>11.5</td>
<td>11.51</td>
<td>11.35</td>
<td>11.36</td>
</tr>
<tr>
<td>15</td>
<td>7.79</td>
<td>7.91</td>
<td>7.76</td>
<td>7.82</td>
</tr>
<tr>
<td>16</td>
<td>6.98</td>
<td>6.93</td>
<td>6.91</td>
<td>7.04</td>
</tr>
</tbody>
</table>
Problems

How to treat the data?

We have big amounts of data but we need tools to be able to transform all this data into information.
Outline

1. Introduction to Microarrays
2. Tool Overview
   - Motivation
   - Babelomics Suite
   - Definitions
   - Questions
3. Analysis
4. Example

Martina Marbà  mmarba@cipf.es  Babelomics: Microarray Data Analysis
Motivation: Extracting information from data

- Many new statistical considerations
  - Data normalization
  - Multiple test correction

- New analysis are possible
  - Predictors
  - Clustering

- Computational problems
  - Many gigas of data
  - How to process all this data

- Biological Databases
  - More than 1200 thousand
Babelomics: the Web

It’s a merge and rewrite of Babelomics-3 and GEPAS-4:

- Integration
- Speed up
- Development platform

More than 10 publications in NAR in the last 5 years

Babelomics was chosen by NAR editors as Featured Article in 2010
Babelomics: definitions

- One of the most complete integrated packages of tools for **microarray data analysis** available over the web
- A complete suite of web tools for **functional analysis** of genome-scale experiments
Questions

- Is there any **significant difference** in gene expression between tumor and healthy cells?
- Can we detect group of genes with **similar** expression profiles?
- Can we **classify new** biological samples based on gene expression patterns?
- Is there any **significant functional enrichment** in my gene list?
- Are these genes involved in the **same disease**?
Outline

1. Introduction to Microarrays
2. Tool Overview
3. Analysis
   - Analysis progress
   - Gene expression analysis
   - Functional Analysis
4. Example

Martina Marbà  mmarba@cipf.es

Babelomics: Microarray Data Analysis
Babelomics: Modules/Analysis progress

Gene Expression Analysis

Functional Analysis

Upload data
Processing data
Expression
Genomic
Functional analysis

Data
Normalization
Diff. Expression Clustering Predictors
FatiGO Fatiscan

Gene Expression ANALYSIS

Functional ANALYSIS

Martina Marbà  mmarba@cipf.es  Babelomics: Microarray Data Analysis
Characteristics of the data:

- We NEVER deal with individual arrays, we deal with collections of arrays obtained for a given experimental design.
- Most of the genes are not informative with respect to the trait we are studying (account for unrelated physiological conditions, etc.).
- Number of variables (genes) is several orders of magnitude larger than the number of experiments.
Babelomics: Modules

- **Gene Expression Analysis**
  - Normalization
  - Differential Expression
  - Clustering
  - Predictors

- **Functional Analysis**
  - FatiGO
  - FatiScan
Normalization

- Upload data
- Processing data
  - Edit
    - Modify your uploaded data.
  - Normalize
    - Expression
      - One-channel
        - Affymetrix
        - Normal
    - Agilent
  - Affy normalization
    - Online examples (test the form with example data)
    - Select your data
      - browse server
        - no data selected.
        - Or go to Upload Data form: Upload [affymetrix:expression: microarray]
    - Analysis
      - RMA
      - Plier
      - Present-absent calls
    - Job
      - Job name: sample name
      - Job description
    - Run
Normalization

To adjust for the effects that are due to variations in the technology rather than the biology.
There is always:

- **noise** from technical irregularities
- that produces **signal effects** not due to biological reason

**Background effects**

**Scale variability**
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 relative to the others.
- The number of genes over-expressed in a condition is similar to the number of genes under-expressed.

These assumptions should agree with your experimental context.
Differential Expression

To find genes that are differentially expressed between classes or conditions.

- Class comparison
- Correlation
- Survival
- Time / dosage series
Differential Expression: results

- We perform one hypothesis test for each gene
- There is an increased chance of finding false positives
- We need to adjust p-values to control
  - FDR (false discovery rate)
  - Bonferroni
  - Holm, ...
Differential Expression

Class Comparison

Two class comparison

- T-test
- Limma
- Fold-change
Differential Expression

**T-test for a gene expression**

For each gene, we check if its mean expression is equal or different across the two classes.

\[
H_0 : \mu_1 = \mu_2 \\
H_1 : \mu_1 \neq \mu_2
\]

- \(H_0\) : (Null hypothesis) mean expression is equal in both groups.
- \(H_1\) : (Alternative) mean expression is different between groups.

T-statistic: \( T = \frac{Y_1 - Y_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \)
Differential Expression

P-value:

- If we reject when p-value < 0.05 there is a 5% chance of getting a **false positive**.
- On average:
  - If you test 100 hypotheses 5 will be false positives (appear significantly wrong)
  - If you test 10000 hypotheses 500 will appear as false positives

**Multiple testing correction** is needed
Differential Expression

Multi Class Comparison

Ordered genes by statistic

Multi class tests
- Anova
- Limma
Differential Expression

Correlation coefficient:
- Pearson correlation coefficient
- Spearman correlation coefficient
- Linear regression

Arrays ranked according to the independent variable.
Genes ranked by correlation to the continuous variable.
Differential Expression

Survival data:
- Cox model coefficient

Arrays ranked according to the survival time.
Genes ranked by their relationships to the survival time.
To find genes with a changing pattern:

- along time
- or increasing dose concentrations

with different profile evolutions between different series.
Clustering

- Differential expression
  - Class comparison:
    - Study differential
  - Correlation
  - Study expression
  - Survival
    - Study the relation
  - Time/dosage series
    - The module finds

- Predictors
  - Class prediction
  - Builds prediction

- Clustering

Select your data
- browse server: no data selected.
- Or go to Upload Data form: Upload (datamatrix)

Select type of clustering: samples and/or genes
- Clustering of samples
- Clustering of genes

Select method
- UPGMA
- SOTA
- K-means
  - Number of sample-clusters (k-value): 5
  - Number of gene-clusters (k-value): 16

Select distance
- Euclidean (normal)
- Euclidean (square)
Clustering

Our interest is to detect patterns between genes or samples joining together these elements that are more similar among them.

Can we find groups of experiments with similar gene expression profiles?

• What genes co-express?
• How many different expression patterns do we have?
• What do they have in common?
• Etc.
Clustering

- Gene clusters are previously unknown
- Distance function
- Cluster gene expression patterns based uniquely on their similarities
Predictors

- Differential expression
  - Class comparison
  - Correlation
  - Study expression
- Survival
- Study the relation
- Time/dosage series
- The module finds

- Predictors
- Class prediction
- Builds prediction

- Clustering

Error estimation

- Validations
  - Leave-one-out
  - KFold
- repeats 10
- folds 5

Select your data

- Train
  - Train and test
- Train data (expression matrix)
  - browse server
  - no data selected.
  - Or go to Upload Data form: Upload [datamatrix]
- Class name: No classes available

Algorithms

- SVM
- KNN
- Random forest

Introduction to Microarrays
Tool Overview
Analysis
Example
Class Prediction

- How can different classes be distinguished based on the corresponding profiles of gene expression?
- How a phenotypic trait (resistance to drugs, survival, etc.) can be predicted?

Gene selection

- Which genes among the thousands analysed are relevant for the classification?
Predictors

Prediction or classification of new unlabelled samples from previous specific labelled samples

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is X, A or B?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diff (B, X)</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diff (A, X)</td>
<td>13</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unknown class/label
From data to functional interpretation
Functional Analysis

Babelomics, a suite of web tools for: statistical test, multiple test corrections, blast, ...

List of genes or ids, i.e.: differentially expressed genes, ...

Integrated Biological DB of Functional Annotation
(GO, KEGG, InterPro, Ensembl, SwissProt, Transcription Factors, MicroRNA, Cisred, Biocarta, Bioentities Literature)
FatiGO

Functional analysis

- Single enrichment analysis
  - FatiGO
  - Marmite

- Set enrichment analysis
  - Gene set analysis
    - Finds gene-sets with gene-set analysis
  - MarmiteScan
  - Ge5BAP
    - Gene set analysis (a with SNPs or CNVs)

Define your comparison

- Id list vs Id list
- Id List vs Rest of genome
- Id List vs Rest of ids contained in your annotations (complementary list)

Select your data

List 1:
- browse server
- Or go to Upload Data form: Upload [idlist]

List 2:
- browse server
- Or go to Upload Data form: Upload [idlist]

Options

- Fisher exact test: Two tailed
- Remove duplicates: Never

Martina Marbà mmarba@cipf.es

Babelomics: Microarray Data Analysis
FatiGO features

- Compares **two** lists of genes.
- Compares **one** list of genes against the rest of the genome.

- One statistical test (Fisher’s exact) for each **Block** of annotations.
  - Multiple testing context
  - Filtering of annotation is convenient (the less tests the best correction).
FatiGO features

- Compares **two** lists of genes.
- Compares **one** list of genes against the rest of the genome.

- One statistical test (Fisher’s exact) for each **Block** of annotations.
  - Multiple testing context
  - Filtering of annotation is convenient (the less tests the best correction).
FatiGO test

Genes in group A have significantly to do with biosynthesis, but not with sporulation.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biosynthesis</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>No biosynthesis</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

We do this for each GO, miRNA, Interpro, ... !!!
FatiScan

Functional analysis

- Single enrichment analysis
  - Fatigo
    - Provides significant
  - Marmite
    - Single enrichment analysis

- Set enrichment analysis
  - Gene set analysis
    - Finds gene-sets with gene-set analysis
  - MarmiteScan
    - Implements gene-set analysis
  - GeSBAP
    - Gene set analysis (a with SNPs or CNVs)

Select your ranked list

- Logistic model
- Fatiscan
  - Fisher exact test: Two tailed
  - Remove duplicates: Never

Databases

- Organism: Select an organism
  - GO biological process [options]
  - GO molecular function [options]
  - GO cellular component [options]
FatiScan features

- Interpret a **ranked list of genes.**
- There is not need for choosing a cut-off. All information is included.

  - One statistical test (Fisher’s exact) for each **Block** of annotations.
    - Multiple testing context (hundreds of annotations)
    - Filtering of annotation is convenient (the less tests the best correction).
FatiScan features

- Interpret a **ranked list of genes**.
- There is not need for choosing a cut-off. All information is included.

- One statistical test (Fisher’s exact) for each **Block** of annotations.
  - Multiple testing context (hundreds of annotations)
  - Filtering of annotation is convenient (the less tests the best correction).
FatiScan: testing along an ordered list

- Index ranking genes according to some biological aspect under study.
- Database that stores gene class membership information.
- **FatiScan** searches over the whole ordered list, trying to find runs of functionally related genes.
### FatiScan results

#### Significant results:

<table>
<thead>
<tr>
<th>Term</th>
<th>Term size</th>
<th>Term size (in genome)</th>
<th>Term annotation % per list</th>
<th>Annotated Ids</th>
<th>Odds ratio (log e)</th>
<th>pvalue</th>
<th>Adjusted pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>negative regulation of apoptosis (GO:0043066)</td>
<td>412</td>
<td>403</td>
<td>7.2%</td>
<td>list 1: 205229_at,20979... list 2: ENSG0000001084,ENSG...</td>
<td>1.5495</td>
<td>7.006e-13</td>
<td>7.65e-10</td>
</tr>
<tr>
<td>negative regulation of programmed cell death (GO:0043069)</td>
<td>418</td>
<td>409</td>
<td>7.2%</td>
<td>list 1: 205229_at,20979... list 2: ENSG0000001084,ENSG...</td>
<td>1.5334</td>
<td>1.074e-12</td>
<td>7.65e-10</td>
</tr>
<tr>
<td>cellular amino acid derivative metabolic process (GO:0006575)</td>
<td>182</td>
<td>173</td>
<td>4.8%</td>
<td>list 1: 209604_s_at,20979... list 2: ENSG0000001084,ENSG...</td>
<td>1.995</td>
<td>9.24e-13</td>
<td>7.65e-10</td>
</tr>
<tr>
<td>cellular amino acid and derivative metabolic process (GO:0006519)</td>
<td>447</td>
<td>447</td>
<td>7.4%</td>
<td>list 1: 209604_s_at,20979... list 2: ENSG0000001084,ENSG...</td>
<td>1.491</td>
<td>1.7e-12</td>
<td>9.082e-10</td>
</tr>
</tbody>
</table>

**Enriched class**

**Annotated genes par GO from each list**
Outline

1. Introduction to Microarrays
2. Tool Overview
3. Analysis
4. Example
   - Experiment description
   - Analysis
   - Results
Example

Running an Example ...
Example of Rheumatoid Arthritis and Osteoarthritis

Experiment description:

- 15 Affymetrix (HG-U133A), 38,500 genes
- Human
- 3 classes: control, osteoarthritis and rheumatoid arthritis
Example of Rheumatoid Arthritis and Osteoarthritis

Analysis:

- Normalization (rma)
- Clustering (by samples)
- Differential Expression (ND vs RA)
- FatiScan
Example: Results

Normalization

Clustering

Tables in txt format files with data results:
  rma.summary.txt

Clustering of samples:

- GSM4402
- GSM4401
- GSM4388
- GSM4392
- GSM4397
- GSM4394
- GSM4396
- GSM4393
- GSM4388
- GSM4385
- GSM4391
- GSM4393
- GSM4379

Martina Marbà  mmarba@cipf.es  Babelomics: Microarray Data Analysis
Example: Results

Differential Expression

FatiScan

<table>
<thead>
<tr>
<th>Term</th>
<th>Term size</th>
<th>Term size (in genome)</th>
<th>Term annotation</th>
<th>Annotated hits</th>
<th>Odds ratio (log2)</th>
<th>p-value</th>
<th>Adjusted p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>protein secretion (GO:0006936)</td>
<td>91</td>
<td>102</td>
<td>0.55%</td>
<td>5</td>
<td>-1.67</td>
<td>0.000247</td>
<td>0.00189</td>
</tr>
<tr>
<td>immune system development (GO:0012392)</td>
<td>493</td>
<td>308</td>
<td>3.32%</td>
<td>10</td>
<td>-3.66</td>
<td>0.001480</td>
<td>1.14E-10</td>
</tr>
<tr>
<td>cell cycle checkpoint (GO:000075)</td>
<td>36</td>
<td>80</td>
<td>-1.46%</td>
<td>2</td>
<td>-0.81</td>
<td>0.60536</td>
<td>0.60536</td>
</tr>
<tr>
<td>reproductive burst (GO:0046750)</td>
<td>21</td>
<td>19</td>
<td>-2.64%</td>
<td>1</td>
<td>-0.71</td>
<td>0.00727</td>
<td>0.002246</td>
</tr>
<tr>
<td>oocyte/ovarian metabolic process (GO:0000811)</td>
<td>21</td>
<td>27</td>
<td>-5.15%</td>
<td>2</td>
<td>-1.81</td>
<td>0.00410</td>
<td>0.00410</td>
</tr>
<tr>
<td>antigen processing and presentation of peptide or polysaccharide antigen via MHC class I (GO:0002094)</td>
<td>23</td>
<td>16</td>
<td>-0.02%</td>
<td>1</td>
<td>-0.01</td>
<td>0.9543</td>
<td>2.00E-05</td>
</tr>
<tr>
<td>mitotic spindle (GO:0000919)</td>
<td>177</td>
<td>188</td>
<td>-1.60%</td>
<td>1</td>
<td>-0.90</td>
<td>0.2434</td>
<td>0.22877</td>
</tr>
<tr>
<td>blood coagulation (GO:0007166)</td>
<td>171</td>
<td>182</td>
<td>-1.77%</td>
<td>1</td>
<td>-0.80</td>
<td>0.1830</td>
<td>0.020332</td>
</tr>
<tr>
<td>immune to hydrogen peroxide (GO:004342)</td>
<td>77</td>
<td>50</td>
<td>-5.53%</td>
<td>2</td>
<td>-2.01</td>
<td>0.001842</td>
<td>0.002457</td>
</tr>
<tr>
<td>regulation of actin filament length (GO:003833)</td>
<td>64</td>
<td>61</td>
<td>-0.47%</td>
<td>1</td>
<td>-1.50</td>
<td>0.01418</td>
<td>0.0002459</td>
</tr>
</tbody>
</table>

Tables in txt format files with data results:
img2/t_significative_table.txt
Acknowledgements

Fatima Al-Shahrour
Eva Alloza
José Carbonell
Ignacio Medina
Ana Conesa
Pablo Escobar
Francisco García

Stefan Götz
Martina Marbà
Pablo Mínguez
David Montaner
Javier Santoyo
Joaquín Tárraga
Luz García

http://bioinfo.cipf.es