Gene expression analysis: 
Introduction to microarrays

Adam Ameur
The Linnaeus Centre for Bioinformatics,
Uppsala University

February 15, 2006
Overview

• Introduction

• Part I: How a microarray experiment is carried out
  – Step by step description of microarray experiment
  – Possible error sources
  – What can we learn from a microarray experiment
  – What can we not learn from a microarray experiment

• Part II: Analysis of microarray data
  – Visualization
  – Correcting for systematic errors in the data
  – Detecting interesting genes
  – Biological validation/interpretation of results
Part I: Introduction - Microarrays

- Microarrays can be used for several purposes:
  - Measuring gene expression levels for thousands of genes simultaneously.
  - Detecting chromosomal aberrations (array-CGH)
  - Detecting \textit{in vivo} transcription factor binding sites (ChIP-chip)

- Different array types:
  - cDNA and long oligonucleotide arrays
    - Comparison of expression between samples (treated vs control)
    - Can be constructed 'in-house', commercial arrays available
    - Relatively Cheap
  - Short oligonucleotide arrays
    - Each sample is hybridized individually – simpler experimental design
    - Less technical variation
    - Expensive \textasciitilde3000 SEK/array
    - Most short oligo arrays are from Affymetrix
  - Data analysis somewhat different for the two types
    - Here we focus on cDNA and long oligo-arrays
Microarrays: Basic idea

RNA extraction

reverse transcription

cDNA Spotting
Printing or Coupling Oligos

label green & hybridize

label red & hybridize

2 Channel Array
The steps of a microarray experiment

A microarray experiment:

- Biological question
- Experimental design
- RNA extraction
- Hybridization
- Scanning
- Image quantification
- Data analysis presented later…
Biological question

• A microarray experiment starts with a biological question:
  – Some examples:
    • Gene expression changes in cell cycle
    • Response in gene expression in endothelial cells to growth factor VEGF
    • Stress response in yeast
    • Differences in gene expression between cancer cells and healthy cells
    • Differences in gene expression between tissues
  – More systematic studies.
    • Expression profiles for knockout strains in yeast
    • Bioinformatics applications – classification of genes or samples.

• The biological question influences the design of the experiment
Experimental design

- Given the biological question, decide on the following:
  - What type of array should be used?
  - How many replicates are required?
    - Biological vs. Technical
  - Which samples should be hybridized together?
    - Not a problem for short oligo arrays.

→ Cost and biological samples are usually limiting factors
Experimental design – Manufacturing the arrays (1)

- Design of cDNA and long oligo arrays
  - cDNA arrays
    - cDNA clones (up to 5000 bp)
  - Long oligos
    - Probes of about 100 bp
  - Duplicates of each spot possible
  - Printing in groups by a robot
  - There may be errors introduced during the printing of the array
Experimental design – Manufacturing the arrays (2)

- Short oligo arrays
  - Each probe is ~ 25 nucleotides long
  - 16-20 probes for each gene
  - Less cross hybridization

- Probes manufactured directly on the array
RNA extraction

• How much RNA is needed?
  – Microarray experiments require about 20 micrograms of mRNA, though there are some protocols that require as little as 5 micrograms.

• mRNA extraction
  – Commercial kits are available

• Problems
  – The samples should be representative for the population.
  – The RNA may become damaged/degraded during extraction
    • If the sample is contaminated with RNAse, the RNA is destroyed.
    • The sample quality can be assessed by running it through a gel.
Hybridization

• RNA is cloned to cDNA with reverse transcriptase.

• The cDNA is labeled.
  – Fluorescent labeling (e.g. Cy3 and Cy5) is most common, but radioactive labeling is also used.
  – Labeling may be incorporated in hybridization, or applied afterwards.

• Then, labeled samples are hybridized to the microarrays.

• Problems
  – Since different labels are used, this may influence the results:
    • Labels may react differently with the RNA and have different fluorescent patterns.
    • Hybridization may not be uniform across the array.
Scanning and image quantification

• Once the sample has been hybridized to the array, the array is scanned. The result is an image of the array:

![Image of array]

• The picture is then analyzed by a program to
  – Identify the spots
  – Find out which spot is associated with each gene
  – Quantify the spots
• Different scanners and programs may give different results
• Problems
  – Finding spots
  – Saturation: Impossible to detect differences between two very bright spots.
What questions can microarray experiments answer

- Microarrays give global measurements of all (or most) mRNA
  - Useful for scanning
    - If we don’t know exactly what we are looking for
  - Useful for getting a big picture
    - Which biological functions are common among the genes that change?
  - Useful for classification
    - Pattern of gene expression characteristic for the state of a cell.
  - Useful for finding genes are differentially expressed between ...
    - Cell types: tissues, cancer cells/healthy cells, different cell lines, different species
    - Developmental stages, cell cycle states
    - exposure to external signals: growth factors, stress, nutrients…
    - mutations, knockouts etc…
What question can microarray experiments not answer

• Microarrays have low sensitivity
  – Hard to separate changes in expression from noise.
  – Hard to detect low expression levels, and small changes

• We only get a snapshot of what is going on in the cell

• mRNA expression doesn’t tell us everything !!!
Part II: Analysis of Microarray data

The aim is to give an answer to the biological question.

• Requires that:
  – Microarrays are suited for answering the biological question
  – Correct experimental design have been chosen
  – Data has sufficiently high quality

The usual biological question:

‘Detect all genes that are significantly differentially expressed between different experimental conditions’

• Experimental conditions:
  – Different treatments
  – Different time points
Overview of microarray data analysis

Usually, many analysis steps are required

- Visualization
- Normalization
- Filtering
- Clustering
- Biological validation
- Statistical tests

Examples from the LCB-DWH follows:
Example Hybridization

Hypothesis: Most genes should be unaffected by the treatment

→ Same signal for Cy5 and Cy3
Visualization – RG plot

- Shows dependencies between red and green intensities
- Points expected to be centered around the diagonal (unchanged)
- Difficult to see systematic errors, since most spots have low intensities
Visualization – MA plot

MA plot:

- Transformation of the RG plot
- M is a measure of the spot color
- A is a measure of the spot intensity
- Points expected to be centered around M = 0 (unchanged)
- Useful for detecting systematic dye-related errors

\[ M = \log_2(\text{Red}/\text{Green}) \]

\[ A = \frac{1}{2} \log_{10}(\text{Red} \cdot \text{Green}) \]
Visualization – Array plots

Array plot:

- Displays the layout of the physical array
- Useful for detecting spatial variation
Visualization

Print-tip group plots:

- Plots each individual print-tip group
- Lines and boxes are expected to be at 0
- Useful for detecting spatial variation as well as systematic errors
Visualization: PCA

Principal Component Analysis:

- Data from each array is seen as an N (nr of genes) dimensional vector
- Projection onto a 2D plane consisting of directions of most variance
- Arrays with similar data are close to each other
- Only for visualization
Normalization

Normalization aims at removing systematic variation in the data.

**Assumption in normalization methods:**
‘Most genes are not differentially expressed between sample and reference’

Classes of (within slide) normalization methods:
- **Background subtraction methods**
  - Corrects spot intensities with respect to variation in background intensities.
- **Global normalization methods**
  - Forces all intensities on a slide to follow some criteria
    - Mean or median of log-ratios = 0
    - Local regrESSion (LOESS)
- **Local normalization methods**
  - Forces all intensities in a print-tip group to follow some criteria
    - Print-tip LOESS
- **Quality control methods**
  - Uses control spots for which the expected log-ratios are known
Normalization

Print-tip loess normalization:

- Intensities are corrected for each print-tip group
- Spots with poor quality (flagged) should have little effect
- A good result plot doesn’t mean all systematic variation is gone
Normalization

How to know if the normalization was successful?

- Not easy to tell, but we can:
  - Check the values of control spots (if there are such values)
  - Compare PCA-plots, before and after normalization
    - Replicates should lie close to each other
Detecting interesting genes

**Hypothesis testing** is used for detecting differentially expressed genes.

- **Input:**
  - Normalized and filtered microarray data
  - Information about the experiment design

- **Some hypothesis testing methods:**
  - Standard t-test
  - Significance Analysis of Microarrays (SAM)
  - B-statistics

- **Output:**
  - Every gene is associated with a score
  - Ranked gene list
  - Needs a cut-off to tell which genes are differentially expressed
  - The gene list may require further analysis
Hypothesis testing: B-statistics

An empirical Bayes method for detecting differential expression

- Each gene is associated with a B-score (level of differential expression)
- To get a high B-score a gene shall have:
  - Low variation between replicates
  - High variation between groups

![B-score vs M value plot](image)
Further analysis: Clustering

K-means clustering:

Hierarchical clustering:
Biological validation: Gene Ontology

The GO describes gene products. Consists of three classes:

- **MF: Molecular Function**
  - The tasks performed by individual gene products
  - Examples: carbohydrate binding and ATPase activity
  - 6957 terms

- **BP: Biological Process**
  - Broad biological goals that are accomplished by ordered assemblies of molecular functions
  - Examples: mitosis and purine metabolism
  - 9304 terms

- **CC: Cellular Component**
  - Sub-cellular structures, locations, and macromolecular complexes
  - Examples: nucleus, telomere and origin recognition complex
  - 1512 terms

The terms within the GO is organized in a hierarchical structure (DAG)
Example: A subset of the GO

GO

BP

MF

CC

Behaviour
BP unknown
Cellular process
Development
Physiological process
Regulation of BP
Viral life cycle

Membrane fusion
Cell communication
Cell differentiation
Cellular phys. process
Regulation of cellular process

Cell development

Sporulation

"part of"

"is a"
Genes and GO

- Mappings from genes to GO
  - Each gene is mapped to one or many GO terms
  - The terms are the most specific terms that are appropriate
  - Mappings of genes are not handled by the GO consortium, but by others
    - For example by the Gene Ontology Annotation (GOA) group
      Entrez gene ids (Locus Link) identifiers → GO terms
  - There is an evidence code for each individual mapping
    - Says something about the reason for the mapping
  - All appropriate GO terms for a gene is called the induced GO graph.

```
  BP
   / \
  /   \
/     \
/-------\
Behaviour  BP unknown  Cellular process  Development  Physiological process  Regulation of BP  Viral life cycle

  /          \                                      /                                      |
  /          \                                      /                                      |
/            \                                    /                                      |
/            \                                    /                                      |
/            \                                    /                                      |
Cell membrane fusion  Cell communication  Cell differentiation  Cellular phys. process  Regulation of cellular process

/                        \                                      /                                      |
/                        \                                      /                                      |
/                        \                                    /                                      |
/                        \                                    /                                      |
/                        \                                    /                                      |
Cell development  ...  Sporulation

```

GENE
GO and microarray data

We can do a **statistical test** using this data

**Input:**
- A set of differentially expressed genes
- A reference set of genes, e.g. all genes on the array
- The GO terms for all genes on the array

**Output:**
- GO terms that are either over- or under represented among the DE genes
- P-value for each GO-term
The GO significance test tool

- Example: Results at significance level 0.025
  - 50 Significant GO terms were observed (5.2 expected)

GO:0006869 lipid transport (1.03e-06)
GO:0042157 lipoprotein metabolism (1.03e-06)
GO:0008015 circulation (3.08e-06)
GO:0009653 morphogenesis (7.94e-05)

GO:0007186 G-protein coupled receptor protein signaling pathway (0.000254)
GO:0007600 sensory perception (0.000375)
GO:0009581 detection of external stimulus (0.000375)
GO:0007606 sensory perception of chemical stimulus (0.000578)
The GO significance test tool

- The results can also be presented in tables:

<table>
<thead>
<tr>
<th>Over represented (two sided p-values)</th>
<th>GO id/term</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006888 lipid transport</td>
<td>0.03e-06</td>
<td></td>
</tr>
<tr>
<td>GO:0042157 lipoprotein metabolism</td>
<td>0.03e-06</td>
<td></td>
</tr>
<tr>
<td>GO:0008016 circulation</td>
<td>3.00e-06</td>
<td></td>
</tr>
<tr>
<td>GO:0005653 morphogenesis</td>
<td>7.94e-05</td>
<td></td>
</tr>
<tr>
<td>GO:0018043 cell organization and biogenesis</td>
<td>0.00016</td>
<td></td>
</tr>
<tr>
<td>GO:0030029 actin filament-based process</td>
<td>0.000223</td>
<td></td>
</tr>
<tr>
<td>GO:0030036 actin cytoskeleton organization and biogenesis</td>
<td>0.000223</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Under represented (two sided p-values)</th>
<th>GO id/term</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:007138 G-protein coupled receptor protein signaling pathway</td>
<td>0.00054</td>
<td></td>
</tr>
<tr>
<td>GO:0007600 sensory perception</td>
<td>0.00075</td>
<td></td>
</tr>
<tr>
<td>GO:0009681 detection of external stimulus</td>
<td>0.00075</td>
<td></td>
</tr>
<tr>
<td>GO:0007638 sensory perception of chemical stimulus</td>
<td>0.00078</td>
<td></td>
</tr>
<tr>
<td>GO:0007630 perception of smell</td>
<td>0.00078</td>
<td></td>
</tr>
<tr>
<td>GO:0007601 visual perception</td>
<td>0.00213</td>
<td></td>
</tr>
<tr>
<td>GO:0052653 sensory perception of light</td>
<td>0.00213</td>
<td></td>
</tr>
<tr>
<td>GO:005790 regulation of enzyme activity</td>
<td>0.00254</td>
<td></td>
</tr>
<tr>
<td>GO:0007194 negative regulation of adenylyl cyclase activity</td>
<td>0.00366</td>
<td></td>
</tr>
<tr>
<td>GO:0051350 negative regulation of lyase activity</td>
<td>0.00366</td>
<td></td>
</tr>
<tr>
<td>GO:0007187 G-protein signaling, coupled to cyclic nucleotide second messenger</td>
<td>0.00371</td>
<td></td>
</tr>
<tr>
<td>GO:0007136 G-protein signaling, coupled to c-AMP nucleotide second messenger</td>
<td>0.00371</td>
<td></td>
</tr>
<tr>
<td>GO:0019932 second messenger-mediated signaling</td>
<td>0.00371</td>
<td></td>
</tr>
<tr>
<td>GO:0019933 c-AMP-mediated signaling</td>
<td>0.00371</td>
<td></td>
</tr>
<tr>
<td>GO:0019935 cyclic-nucleotide-mediated signaling</td>
<td>0.00371</td>
<td></td>
</tr>
</tbody>
</table>
Current analysis tools:

- **Normalization**
  - Background correction*
  - Normalize between arrays*
  - Normalize within arrays*
  - On all values in every column

- **Filtering**
  - B-statistics*
  - SAM*

- **Hypothesis testing**
  - Array plots*
  - Quality control plots*
  - PCA*

- **Visualization**
  - Hierarchical clustering
  - K-means clustering

- **Clustering**
  - GO visualization*
  - GO significance test*

- **Biological validation**
  - Merge assays*
  - Merge duplicate spots
  - Imputation

*R/Bioconductor functions