Analysis of Microarray Data

By
H. Bjørn Nielsen
The DNA Array Analysis Pipeline

Question
Experimental Design

Array design
Probe design

Sample Preparation
Hybridization

Image analysis

Normalization

Comparable
Gene Expression Data

Statistical Analysis
Fit to Model (time series)

Advanced Data Analysis
Clustering
PCA
Classification
Promoter Analysis
Meta analysis
Survival analysis
Regulatory Network
What's the question?

Typically we want to identify differentially expressed genes.

Example:
Alcohol dehydrogenase is expressed at a higher level when alcohol is added to the media.

Alcohol dehydrogenase

Without alcohol

With alcohol
Intensities are not just mRNA concentrations

Sample preparation
• RNA degradation
• RNA purification
• Reverse transcription
• Amplification efficiency
• Dye effect (cy3/cy5)

Array / hybridization
• Spotting
• DNA-support binding
• Spatial effects (washing etc.)
• Image segmentation
Two kinds of variation

**Global variation**

- Amount of RNA in the biopsy
- Efficiencies of:
  - RNA extraction
  - Reverse transcription
  - Amplification
  - Labeling
  - Photodetection

**Gene-specific variation**

- Spotting efficiency,
  - Spot size
  - Spot shape
- Cross-/unspecific hybridization
- Biological variation
  - Effect
  - Noise

**Systematic**

**Stochastic**
The DNA Array Analysis Pipeline

Question
Experimental Design

Array design
Probe design

Sample Preparation
Hybridization

Image analysis

Normalization

Expression Index
Calculation

Comparable
Gene Expression Data

Statistical Analysis
Fit to Model (time series)

Advanced Data Analysis
Clustering PCA Classification Promoter Analysis
Meta analysis Survival analysis Regulatory Network

Buy Chip/Array
Calibration = Normalization = Scaling

Unnormalized data

Linear normalized data
Nonlinear normalization

Unnormalized data

Gspine normalized data
One of the most commonly utilized normalization techniques is the LOcally Weighted Scatterplot Smoothing (LOWESS) algorithm.
The DNA Array Analysis Pipeline

Question
Experimental Design

Array design
Probe design

Sample Preparation
Hybridization

Buy Chip/Array

Image analysis

Normalization

Comparable Gene Expression Data

Statistical Analysis
Fit to Model (time series)

Advanced Data Analysis
- Clustering
- PCA
- Classification
- Promoter Analysis
- Meta analysis
- Survival analysis
- Regulatory Network

Expression Index Calculation
Expression index value

- Some microarrays have multiple probes addressing the expression of the same gene
  - Affymetrix chips have 11-20 probe pairs pr. Gene
  
  - Perfect Match (PM)
  - MisMatch (MM)

PM:  CGATCAATTGCACACTATGTCATTTCT
MM:  CGATCAATTGCAGTATGTCATTTTCT
Why use gene expression index?

For most downstream analysis only one value pr. gene are used.

Therefore we collapse the intensities from many probes into one value: a gene expression index value
Expression index calculation

• Simplest method?

  Median

But more sophisticated methods exists:
dChip, RMA and MAS 5 (from Affymetrix)
All of this is implemented in...

- R

- In the BioConductor packages ‘affy’

(Gautier et al., 2003).
Sources of variation

**Systematic**
- Global variation
  - Similar effect on many measurements
  - Corrections can be estimated from data

**Stochastic**
- Gene-specific variation
  - Too random to be explicitly accounted for
  - “noise”

**Normalization**

**Statistical testing**
The DNA Array Analysis Pipeline

Question
Experimental Design

Array design
Probe design

Sample Preparation
Hybridization

Buy Chip/Array

Image analysis

Normalization

Comparable
Gene Expression Data

Expression Index
Calculation

Statistical Analysis
Fit to Model (time series)

Advanced Data Analysis
Clustering
PCA
Classification
Promoter Analysis
Meta analysis
Survival analysis
Regulatory Network
However, the measurements contain stochastic noise. There is no way around it.

Statistics
You can choose to think of statistics as a black box.

Noisy measurements $\rightarrow$ statistics $\rightarrow$ p-value

But, you still need to understand how to interpret the results.
The output of the statistics

**P-value**
The chance of rejecting the null hypothesis by coincidence

-----------------------------
For gene expression analysis we can say:
the chance that a gene is categorized as differentially expressed by coincidence
The statistics gives us a p-value for each gene

We can rank the genes according to the p-value

But, we can’t really trust the p-value in a strict statistical way!
Why not!

For two reasons:

1. We are rarely fulfilling all the assumptions of the statistical test

2. We have to take multi-testing into account
The $t$-test Assumptions

1. The observations in the two categories must be independent

2. The observations should be normal distributed

3. The sample size must be ‘large’ (>30 replicates)
What's inside the black box ‘statistics’

$t$-test or ANOVA
The \( t \)-test

Calculate \( T \)

\[
T = \frac{\overline{X}_2 - \overline{X}_1}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]

Lookup \( T \) in a table
The $t$-test tests for difference in means ($\mu$)

Intensity of gene x

Density

$\mu_{wt}$ wt

$\mu_{mut}$ mutant
Conclusion

- Array data contains stochastic noise
  - Therefore statistics is needed to conclude on differential expression
- We can’t really trust the p-value
- But the statistics can rank genes
- The capacity/needs of downstream processes can be used to set cutoff
- FDR can be estimated
- t-test is used for two category tests
- ANOVA is used for multiple categories