Statistical Methods and Software for the Analysis of Microarray Experiments

Nicholas P. Jewell and Sandrine Dudoit
Division of Biostatistics, UC Berkeley

Mathematical Biosciences Institute
Ohio State University, Columbus, OH
September 20--24, 2004
Acknowledgments

Slides from
Bioconductor Short Courses
www.bioconductor.org

Sandrine Dudoit, Robert Gentleman, Rafael Irizarry, and Yee Hwa Yang.
Outline

• Basic principles.

• Two-color spotted DNA microarrays.

• Affymetrix oligonucleotide chips.
Basic principles
Differential expression

• Each cell contains a complete copy of the organism's genome.
• Cells are of many different types and states E.g. Blood, nerve, and skin cells, dividing cells, cancerous cells, etc.
• What makes the cells different?
  • Differential gene expression, i.e., when, where, and how much each gene is expressed.
• On average, 40% of our genes are expressed at any given time.
Central dogma

The **expression** of the genetic information stored in the DNA molecule occurs in two stages:

- (i) **transcription**, during which DNA is transcribed into mRNA;
- (ii) **translation**, during which mRNA is translated to produce a protein.

**DNA $\rightarrow$ mRNA $\rightarrow$ protein**

Other important aspects of gene regulation: methylation, alternative splicing, etc.
Central dogma
Central dogma
Functional genomics

- The various **genome projects** have yielded the complete DNA sequences of many organisms.
  
  E.g. Human, mouse, yeast, fruitfly, etc.
  
  Human: 3 billion base-pairs, ~30-40 thousand genes.

- Challenge: **go from sequence to function**, i.e., define the role of each gene and understand how the genome functions as a whole.
DNA microarrays
DNA microarrays

• DNA microarray experiments are high-throughput biological assays for measuring the abundance of DNA or RNA sequences in different types of cell samples for thousands of sequences simultaneously.

• Exploit the availability of sequence data to get information on gene expression in different types of cells.
DNA microarrays

• DNA microarrays rely on the hybridization properties of nucleic acids to monitor DNA or RNA abundance on a genomic scale in different types of cells.

• The ancestor of cDNA microarrays: the Northern blot.
Hybridization

- Hybridization refers to the annealing of two nucleic acid strands following the base-pairing rules.

- Nucleic acid strands in a duplex can be separated, or denatured, by heating to destroy the hydrogen bonds.
Hybridization

DNA is denatured by heating

Renaturation on cooling

Hybridization

Nucleic Acid Hybridization
Hybridization
Gene expression assays

- Spotted cDNA arrays (Brown/Botstein);
- Short oligonucleotide arrays (Affymetrix);
- Long oligonucleotide arrays (Agilent Inkjet);
- Fibre optic arrays (Illumina);
- Serial analysis of gene expression (SAGE);
- Etc.
Applications of microarrays

- Measuring transcript abundance (cDNA arrays);
- Genotyping;
- Estimating DNA copy number (CGH);
- Determining identity by descent (GMS);
- Measuring mRNA decay rates;
- Identifying protein binding sites;
- Determining sub-cellular localization of gene products;
- Etc.
Transcriptome

- mRNA or transcript levels sensitively reflect the state of a cell.
- Measuring protein levels (translation) would be more direct but more difficult.
Transcriptome

• The transcriptome reflects
  – Tissue source: cell type, organ.
  – Tissue activity and state:
    • Stage of development, growth, death;
    • Cell cycle;
    • Disease vs. healthy;
    • Response to therapy, stress.
Applications of microarrays

• Cancer research: Molecular characterization of tumors on a genomic scale
  → more reliable diagnosis and effective treatment of cancer.

• Immunology: Study of host genomic responses to bacterial infections.

• Etc.
Applications of microarrays

• Compare mRNA (transcript) levels in different types of cells, i.e., vary
  – Tissue: liver vs. brain;
  – Treatment: drugs A, B, and C;
  – State: tumor vs. non-tumor, development;
  – Organism: different yeast strains;
  – Timepoint;
  – etc.
Two-color spotted DNA microarrays
Spotted DNA microarrays

Prepare cDNA target

"Normal" Tumor

RT / PCR

Label with Fluorescent Dyes

Combine Equal Amounts

Hybridize target to microarray

Microarray Technology

Prepare Microarray

SCAN

www.accessexcellence.com/AB/GG/
Spotted DNA microarrays

• The relative abundance of a spotted DNA sequence in two DNA or RNA samples may be assessed by monitoring the differential hybridization of these two samples to the sequence on the array.

• **Probes**: DNA sequences spotted on the array, immobile substrate.

• **Targets**: Nucleic acid samples hybridized to the array, mobile substrate.
Spotted DNA microarrays

• The ratio of the red and green fluorescence intensities for each spot is indicative of the relative abundance of the corresponding DNA probe in the two nucleic acid target samples.
Spotted DNA microarrays

\[ M = \log_2 \frac{R}{G} = \log_2 R - \log_2 G \]

- **M < 0**, gene is over-expressed in green-labeled sample compared to red-labeled sample.
- **M = 0**, gene is equally expressed in both samples.
- **M > 0**, gene is over-expressed in red-labeled sample compared to green-labeled sample.
The process

Building the microarray:
- MASSIVE PCR
- PCR PURIFICATION AND PREPARATION
- PREPARING SLIDES
- PRINTING

RNA preparation:
- CELL CULTURE AND HARVEST
- RNA ISOLATION
- cDNA PRODUCTION

Hybing the array:
- ARRAY HYBRIDIZATION AND SCANNING
- TARGET LABELING
- POST PROCESSING

DATA ANALYSIS

Dudoit & Jewell
MBI, September 20-24, 2004
The arrayer

Ngai Lab arrayer, UC Berkeley

Print-head
Print-tips collect cDNA from wells

96-well plate
Contains cDNA probes

Glass slide
Array of bound cDNA probes
4x4 blocks = 16 print-tip-groups

Print-tip group 1

cDNA clones

Print-tip group 7
Sample preparation

human sample collection
tissue banks

model systems

pathology

microdissection

cell lines

sources of RNA
Hybridization

Binding of cDNA target samples to cDNA probes on the slide

Hybridize for 5-12 hours
Hybridization chamber

- 3XSSC
- HYB CHAMBER
- ARRAY
- LIFTER SLIP
- SLIDE
- SLIDE LABEL

- Humidity
- Temperature
- Formamide
  (Lowers the Tmp)
Scanning

Detector

PMT

Cy5: 635nm
Cy3: 532nm

Duplicate spots

Image
RGB overlay of Cy3 and Cy5 images
Raw data

• Pairs of 16–bit TIFFs, one for each dye.
• E.g. Human cDNA arrays:
  – ~43K spots;
  – ~ 20Mb per channel;
  – ~ 2,000 x 5,500 pixels per image;
  – spot separation: ~ 136um.
• For a “typical” array, the spot area has
  – mean = 43 pixels;
  – med = 32 pixels;
  – SD = 26 pixels.
Animation

http://www.bio.davidson.edu/courses/genomics/chip/chip.html
Affymetrix oligonucleotide chips

www.affymetrix.com
Terminology

- Each gene or portion of a gene is represented by 11 to 20 oligonucleotides of 25 base-pairs.

- **Probe**: an oligonucleotide of 25 base-pairs, i.e., a 25-mer.
- **Perfect match (PM)**: A 25-mer complementary to a reference sequence of interest (e.g., part of a gene).
- **Mismatch (MM)**: same as PM but with a single homomeric base change for the middle (13th) base (transversion purine <-> pyrimidine, G <->C, A <->T).
- **Probe-pair**: a (PM,MM) pair.
- **Probe-pair set**: a collection of probe-pairs (11 to 20) related to a common gene or fraction of a gene.
- **Affy ID**: an identifier for a probe-pair set.
- The purpose of the MM probe design is to measure non-specific binding and background noise.
 Probe-pair set

GeneChip® Expression Array Design

mRNA reference sequence

Reference sequence

Spaced DNA probe pairs

Fluorescence Intensity Image

Perfect match probe cells

Mismatch probe cells

Figure 1-3 Expression tiling strategy
## Spotted vs. Affymetrix arrays

<table>
<thead>
<tr>
<th>Spotted arrays</th>
<th>Affymetrix arrays</th>
</tr>
</thead>
<tbody>
<tr>
<td>One probe per gene</td>
<td>11 – 20 probe-pairs per gene</td>
</tr>
<tr>
<td>Probes of varying length</td>
<td>Probes are 25-mers</td>
</tr>
<tr>
<td>Two target samples per array</td>
<td>One target sample per array</td>
</tr>
</tbody>
</table>
Oligonucleotide chips

- Millions of copies of a specific oligonucleotide probe
- >200,000 different complementary probes
- Single stranded, labeled RNA target

Image of Hybridized Probe Array

Compliments of D. Gerhold

www.affymetrix.com
Oligonucleotide chips

• The probes are synthesized *in situ*, using combinatorial chemistry and photolithography.

• **Probe cells** are square-shaped features on the chip containing millions of copies of a single 25-mer probe. Sides are 18-50 microns.
Oligonucleotide chips

The manufacturing of GeneChip® probe arrays is a combination of photolithography and combinational chemistry.
Oligonucleotide chips

www.affymetrix.com

Single feature
Oligonucleotide chips

www.affymetrix.com

Hybridization

www.affymetrix.com
Oligonucleotide chips

www.affymetrix.com

Hybridized GeneChip®
Image analysis

• About 100 pixels per probe cell.
• These intensities are combined to form one number representing the expression level for the probe cell oligo.
• → CEL file with PM or MM intensity for each cell.

www.affymetrix.com
Expression measures

• Most expression measures are based on differences of PM-MM.
• The intention is to correct for background and non-specific binding.
• E.g. MarrayArray Suite® (MAS) v. 4.0 uses Average Difference Intensity (ADI) or 
  AvDiff = average of PM-MM.
• Problem: MM may also measure signal.
• More on this in lecture Pre-processing DNA Microarray Data.
What is the evidence?

Biological question

Experimental design

Microarray experiment

Image analysis

Expression quantification

Normalization

Pre-processing

Estimation

Testing

Clustering

Prediction

Biological verification and interpretation
Statistical computing

Everywhere …

• Statistical design and analysis:
  – image analysis, normalization, estimation, testing, clustering, prediction, etc.

• Integration of experimental metadata with biological metadata from WWW-resources
  – gene annotation (GenBank, LocusLink);
  – literature (PubMed);
  – graphical (pathways, chromosome maps).
Integration of experimental and biological metadata

• Phenotypes, microarray gene expression measures, sequence, structure, annotation, literature.

• Integration will depend on our using a common language and will rely on database methodology as well as statistical analyses.

• This area is largely unexplored.
WWW resources

• Complete guide to “microarraying”
  [Link](http://cmgm.stanford.edu/pbrown/mguide/)
  [Link](http://www.microarrays.org)
  - Parts and assembly instructions for printer and scanner;
  - Protocols for sample prep;
  - Software;
  - Forum, etc.

• cDNA microarray animation
  [Link](http://www.bio.davidson.edu/courses/genomics/chip/chip.html)

• Affymetrix
  [Link](http://www.affymetrix.com)