Statistics for Microarray Data

IBC Short Course, 11 July 2004

Gordon Smyth, Yee Hwa Yang,
James Wettenhall, Matt Ritchie

Walter and Eliza Hall Institute
University of California, San Francisco

Central Dogma of Molecular Biology

DNA activity in eukaryotic cell

Each gene is transcribed (at the appropriate time) from DNA into RNA, which then leaves the nucleus and is translated into the required protein.

Any gene which is active in this way at any particular time is said to be expressed.

Microarrays Measure Gene Expression

- Basic idea: measure the activity level of each gene (its expression level), in a particular cell at a particular time, by measuring the concentration of each gene in the cell's mRNA.
- Immobilize fragments (spots) of complementary DNA on a glass slide.
- Wash RNA over the slide, see how much binds (hybridizes) to each spot.

Types of Microarrays

**Two-colour spotted arrays:**
- cDNA microarrays: fragments of DNA, 300-3000 base pairs, on a glass slide.
- long-oligonucleotide spotted arrays: shorter but uniform length nucleotide sequences, 60-90 base pairs say.
- commercial spotted arrays (e.g., Agilent using inkjet technology).

**Single channel arrays:**
- high-density oligonucleotide arrays (e.g., Affymetrix): 20-25 base pairs, single-channel.
- Other commercial arrays (e.g., Applied Biosystems AB1700).

Array Printing

Printing custom library on glass slide.
Two colour arrays compare two RNA samples, e.g., normal vs mutant

Channel 1 emission proportion to gene expression in mutant RNA, Channel 2 to normal RNA

Microarray Image
Scanner output consists of two TIFF images, one for each of red and green channels.
Shown here is false-coloured image with two channels overlaid:
- more highly expressed in mutant
- equally expressed
- more highly expressed in normals

AGRF NIA 15k mouse array:
32k spots, 12 x 4 pin groups

Image Analysis

Image Analysis Software
- TIFF images are processed by image analysis program to acquire intensity values for each spot
- Programs commonly used in Australia include SPOT, GenePix, Imagene and Quantarray

Steps in Image Processing
1. Addressing: locate centers
2. Segmentation: divide pixels into foreground (part of spot) or background
3. Information extraction: calculate signal intensity pairs, background and quality measures for each spot
**Lecture 1: Spotted Microarrays**

**Segmentation**

- Seeded region growing SPOT
- Fixed circle method
  - Genepix: adaptive circles

*Foreground* measured from pixels inside the circle
*Background* measured from ambient intensity around spot.

**Local Backgrounds Measures**

- One channel grey scale
- Genepix: region including pink diamonds

**Morphological Background**

- SPOT estimates background using a nonlinear morphological filter – a lower, less variable measure.

**Quantification of Expression**

- Foreground red, green \( R_f, \ G_f \)
- Background red, green \( R_b, \ G_b \)
- Background corrected \( R = R_f - R_b, \ G = G_f - G_b \)
- Log-ratio (“Minus”) \( M = \log_2 R - \log_2 G \)
- Average intensity (“Add”) \( A = (\log_2 R + \log_2 G) / 2 \)

*Lots of issues: which bg is best, to subtract bg or not, quality, filtering*

**Spot Quality Weights**

*Example of a quality weighting:
  - Downweight spots smaller or larger than nominal size
  - Many other possibilities*

**Exploratory Data Display**
Spatial bias in cDNA arrays

Boxplots of $M = \log_{2} \frac{R}{G}$ by print-tip group (1-16)

Locations of spots with extreme 5% $M$: high, low

Spatial Plots: Background Intensities

MA Plot

Background Affects Variability

From Spot: note lack of fanning (not same data)
From GenePix: note fanning out at low intensities

Normalization

Smyth and Speed “Normalization of cDNA microarray data”, in: METHODS: Selecting Candidate Genes from DNA Array Screens, Dec 2003

No Absolute Measurement Scale

- The spotted DNA sequences differ in binding efficiency, so absolute intensities are not directly comparable between genes
- The response of the dyes is subject to arbitrary rescaling for each dye on each array – need to normalize the red/green balance for each array (to make M-values unbiased)
Statistics for Microarray Data:
Lecture 1: Spotted Microarrays

MA – Plot

M = \log_2 R - \log_2 G

A = \frac{1}{2} (\log_2 R + \log_2 G)

Print-tip-loess Normalization

M = \log_2 R - \log_2 G

A = \frac{1}{2} (\log_2 R + \log_2 G)

Before normalization

After normalization

A f t e r  N o r m a l i z a t i o n

A f t e r  n o r m a l i z a t i o n
Designing Gene Expression Experiments

For a review, see Yang and Speed, *Nature Reviews Genetics*, August 2002

Technical vs Biological Replication

- Technical vs biological replicates: the differences are in the extent to which the replicates are independent. There are many aspects of the process that contribute to common cause dependence.
- Biological replicates involve separate mRNA extractions, from separate organisms or cell lines, followed by separate labelings and hybridizations.

Technical Replication - Labeling

- 3 sets of self-self comparisons: (cerebellum vs cerebellum)
- Data 1 and Data 2 were labeled together and hybridized on two slides separately
- Data 3 was labeled separately

Technical Replication - Amplification

- 3 sets of A vs D performed on different days
- #10 and #12 are from same RNA isolation and amplification
- #12 and #18 are from different dissections and amplifications
- All data sets were labeled separately

Direct vs Indirect Comparisons

Two samples e.g. T vs C, KO vs. WT or transgenic vs. WT

Direct

\[
\text{average (log (T/C))}
\]

Indirect

\[
\text{log (T / Ref)} - \text{log (C / Ref )}
\]

Acknowledgements

WEHI Bioinformatics
- Terry Speed
- Matt Ritchie
- Natalie Thorne
- James Wettenhall
- everyone else

WEHI Scott Lab
- Joelle Michaud
- Catherine Carmichael
- Robert Escher
- Hamish Scott

AGRI
- Steve Wilcox
- Cathy Jensen
- Melanie O’Keefe

WEHI Immunology
- Steve Nutt
- Lynn Corcoran
- Melissa Holmes
- Mireille Lahoud

Berkeley
- Jean Yang
- Speed Lab
- Ngai Lab