Introduction to Microarray Data Analysis

BMI 7830

Kun Huang

Department of Biomedical Informatics
The Ohio State University
Introduction to Microarray

Data Analysis Workflow

Data Normalization
What is microarray?

• Question – how many copies of RNA (or DNA) segments are there?

• Ideal – sequence every sequence. Can it be achieved?

• A “middleman” approach – target what we want or what we know using “probes” (go back to Dr. Parvin’s lecture)

• If thousands of “probes” are put in a chip/slide – we have an array
Two major technologies in microarray

- Affymetrix-like arrays – **single channel** (background-green, foreground-red)
- cDNA arrays – **two channel** (red, green, yellow)
- Protein array
- Tissue microarray – don’t confuse with microarray
- ...
How does microarray work? – Affymetrix
How does microarray work?
How is microarray manufactured?

- **Affymetrix GeneChip**
  - silicon chip
  - oligonucleotide probes lithographically synthesized on the array
  - cRNA is used instead of cDNA
How does microarray work?
How does microarray work?

Shining a laser light at GeneChip® array causes tagged DNA fragments that hybridized to glow

Non-hybridized DNA

Hybridized DNA
How does microarray work? – Two-color
How does two-channel microarray work?

- Printed microarrays
- Long probe oligonucleotides (80-100) long are “printed” on the glass chip
How does two-channel microarray work?

• Printing process introduces errors and larger variance
• Comparative hybridization experiment
Probe selection

- Protocol for extracting mRNA
- 3’ bias – why? Think degradation.
- Multiple probes for one region
- G-C content

http://www.bcgsc.ca/people/malachig/htdocs/alexa_platform/alexa_arrays/intro.htm
Probe selection

- Protocol for extracting mRNA
- 3’ bias – why? Think degradation.
- Multiple probes for one region

http://www.bcgsc.ca/people/malachig/htdocs/alexa_platform/alexa_arrays/intro.htm
http://www.majordifferences.com/2013/05/difference-between-prokaryotic-mrna-and.html#.VAiBDGRdU01
Probe selection
Probe selection

- Protocol for extracting mRNA
- 3’ bias – why? Think degradation.
- Multiple probes for one region
- G-C content

http://www.bcgsc.ca/people/malachig/htdocs/alexa_platform/alexa_arrays/intro.htm
Probe selection

- Protocol for extracting mRNA

[Diagram of exon arrays and probe selection]

http://www.bcgsc.ca/people/malachig/htdocs/alexa_platform/alexa_arrays/intro.htm
Probe Design

• Fabrication expense and frequency of error increases with the length of probe, therefore 25 oligonucleotide probes are employed in Affymetrix GeneChips.

• Problem: cross hybridization

• Solution: introduce mismatched probe with one position (central) different with the matched probe. The difference gives a more accurate reading.
Probe Design

http://www.dkfz.de/gpcf/24.html
Introduction to Microarray

Data Analysis Workflow

Data Normalization
Microarray Data Analysis

Supplementary Figure 1: Clustering of laboratory/platform combinations using log ratio values of common genes
Microarray Data Analysis

Raw data (image)

 QC (artifact)

 Probe level data

 PM vs MM probes

 Reading for probes

 Normalization

 QC (MA plot, box plot, batch effect, PCA)

 Visualization

 Comparative analysis

 Clustering analysis

 Classification (feature selection)

 Enrichment analysis

 Pathway and network analysis
Take a look …

- Poorly defined borders
- Large holes
- Saturated spot
- Dust specs
- Fiber or scratch?
- Bubble
- Edge effect
- Background haze

(McShane, NCI)
Spatial Images of the Microarrays

- Data for the same brain voxel but for the untreated control mouse
- Background levels are much higher than those for the Parkinson’s disease model mouse
- There appears to be something non-random affecting the background of the green channel of this slide
Example – Affymetrix Data Files

- Image file (.dat file)
- **Probe results file (.cel file)**
- Library file (.cdf, .gin files)
- Results file (.chp file)
Example – Affymetrix Data Files

- Image file (.dat file)
- **Probe results file (.cel file)** – we start from here most of the time
Reading from microarray data

cDNA array – ratio, log ratio

\[ T_i = \frac{R_i}{G_i} \quad \text{OR} \quad \log \text{ratio} = \log_2 \frac{R_i}{G_i} \]

Affymetrix array

Difference_{probe pair} = PM - MM

Average Difference_{probe set} = \sum_{i=1}^{n} \frac{(PM_i - MM_i)}{n}
Microarray Softwares

- DChip
- Open source R
- Bioconductor
- BRBArray tools (NCI biometric research branch)
- Matlab
- GeneSpring
- Affymetrix
- …
Microarray Databases

- Gene Expression Ominbus (GEO) database – NCBI
- EMBL-EBI microarray database (ArrayExpress)
  - http://www.ebi.ac.uk/Databases/microarray.html
- Stanford Microarray Database (SMD)
  - http://genome-www5.stanford.edu/
- caARRAY sites
- Other specialized, regional and aggregated databases
  - http://psi081.ba.ars.usda.gov/SGMD/
  - http://www.oncomine.org/main/index.jsp
  - …
Introduction to Microarray

Data Analysis Workflow

Data Normalization
How do we process microarray data (normalization)?

Why normalization – microarray data is highly noisy

- Experimental design
- Replication
- Comparison
How do we process microarray data

- Normalization
  - Intensity imbalance between RNA samples
  - Affect all genes
  - Not due to biology of samples, but due to technical reasons
  - Reasons include difference in the settings of the photodetector voltage, imbalance in total amount of RNA in each sample, difference in uptaking of the dyes, etc.
- The objective is to adjust the gene expression values of all genes so that the ones that are not really differentially expressed have similar values across the array(s).
Box plot

Median

Upper quartile

Low quartile
Normalization

- Two major issues to consider
  - Which genes to use for normalization
  - Which normalization algorithm to use
Normalization

• Which genes to use for normalization
  • Housekeeping genes
    • Genes involved in essential activities of cell
      maintenance and survival, but not in cell
      function and proliferation
    • These genes will be similarly expressed in all
      samples.
    • Difficult to identify – need to be confirmed
  • Affymetrix GeneChip provides a set of house
    keeping genes based on a large set of tests on
    different tissues and were found to have low
    variability in these samples (but still no
    guarantee).
Normalization

• Which genes to use for normalization
  • Spiked controls
    • Genes that are not usually found in the samples (both control and test sample). E.g., yeast gene in human tissue samples.
Normalization

- Which genes to use for normalization
  - Using all genes
    - Simplest approach – use all adequately expressed genes for normalization
    - The assumption is that the majority of genes on the array are housekeeping genes and the proportion of over expressed genes is similar to that of the under expressed genes.
    - If the genes on the chip are specially selected, then this method will not work.
Normalization

- Which normalization algorithm to use
  - Intra-slide normalization
  - Inter-slide for cDNA arrays
Scatter plots of the Microarrays

- A measure of the actual expression levels, i.e., differences between the median foreground and the median background for the red channel and green channel:
  "F635 Median - B635",
  "F532 Median - B532"

Slope = 1
Normalization

- Line (global) normalization
  - Simplest but most consistent
  - Move the median to zero (slope 1 in scatter plot, this only changes the intersection)
  - No clear nonlinearity or slope in MA plot

\[
X_i^{\text{norm}} = k \times X_i \\
c = \log(k) \\
M_i^{\text{norm}} = \log(X_i^{\text{norm}}) = c + M_i
\]
Normalization

• Intensity-based (Lowess) normalization
  • Lowess fit
  • Overall magnitude of the spot intensity has an impact on the relative intensity between the channels.

(McShane, NCI)
Normalization

- Intensity-based (Lowess) normalization
  - “Straighten” the Lowess fit line in MA plot to horizontal line and move it to zero

\[ X_i^{\text{norm}} = k(A) \times X_i \]

\[ c(A) = \log(k(A)) \]

\[ M_i^{\text{norm}} = \log(X_i^{\text{norm}}) = c(A) + M_i \]
Normalization

- Intensity-based (Lowess) normalization
  - Nonlinear
  - Gene-by-gene, could introduce bias
  - Use only when there is a compelling reason

(McShane, NCI)
Normalization

- Location-based normalization
  - Background subtracted ratios on the array may vary in a predictable manner.
  - Sample uniformly across the chip
  - Nonlinear
  - Gene-by-gene, could introduce bias
  - Use only when there is a compelling reason
Normalization
• Quantile normalization
  • Nonlinear
  • Same intensity distribution
Normalization

- Which normalization algorithm to use
  - Inter-slide normalization
  - Not just for Affymetrix arrays
Normalization

- Linear (global) – the chips have equal median (or mean) intensity
- Intensity-based (Lowess) – the chips have equal medians (means) at all intensity values
- Quantile – the chips have identical intensity distribution
- Quantile is the “best” in term of normalizing the data to desired distribution, however it also changes the gene expression level individually
- Avoid overfitting
- Avoid bias
Affymetrix array normalization

- Inter-slide normalization only
- Probe-level normalization
- Affymetrix MicroArray Suite (MAS) 5.0
- Robust Multiarray Average (RMA)

- Quantile
- GC-RMA
Affymetrix array normalization

- Inter-slide normalization only
- Probe-level normalization
- Affymetrix MicroArray Suite (MAS) 4.0
  - Simple subtraction of MM from PM
  - Use only probes within 3 times of SD of PM-MM to exclude outliers
  - Not robust
- MAS 5.0
  - Use weight (Turkey Biweight Estimate) for each probe based on its intensity difference from the mean
  - Log transformed data for mean (geometric mean)
  - Robust
Affymetrix array normalization

- Robust Multiarray Average (RMA)
  - Background correction on each chip.
    - Assuming strictly positive distribution. No negative numbers
  - Do NOT use MM information
  - Normalization (inter-chip).
    - Quantile
  - Probe level intensity calculation.
    - Linear model for signal, affinity, and noise.
  - Probe set summarization.
    - Combine probes for one probeset into a single number
    - Median polishing (chip to its median, gene to its median, iterate and converge)
Affymetrix array normalization

- **GC-Robust Multiarray Average (GC-RMA)**
  - Correct background noise and non-specific binding
  - Affinity computed from position specific base effect
  - MM information is used (subtracted from PM after correction)
Affymetrix array normalization

- RMA/GCRMA pros and cons (comparing to MAS5.0)
  - Less variance at low expression values
  - Less false positives
  - Consistent fold change estimates
  - More false negatives, especially for low-expression level probes
  - Quality control after normalization is difficulty
  - Quantile normalization may overfit and hide real differences
Some notes regarding microarray data

- Multiple probes for single gene
  - Average
  - Sum
  - Pick the one with largest mean (personal favorite)
  - Check consistency and discard if probes are not in consistency (not use!)

- Microarray data are NOISY!!!
  - Cannot be used as true measurement
  - Need to be validated (e.g., qRT-PCR)