Biochip informatics-(I) :

biochip normalization
& differential expression

Ju Han Kim, M.D., Ph.D.
SNUBI: SNUBiomedical Informatics
http://www.snubi.org/

Biochip Informatics - (I)

- Biochip basics
- Preprocessing
- Episodes 1 and 2
- Global normalization
- Intensity dependent normalization
- Controlling regional variation
- Alternatives
- Differential expression
- Multiple hypothesis testing
- Classification
**Biochip basics**

[Diagram showing the process of biochip basics]

**Terminology**

- **Sample (Target):** RNA (cDNA) hybridized to the array, aka target, mobile substrate.
- **Probe:** DNA spotted on the array, aka spot, immobile substrate.
- **Sector (Block):** rectangular matrix of spots printed using the same print-tip (or pin), aka print-tip-group
- **Slide, Array:** printed microarray
- **Batch:** collection of microarrays with the same probe layout.
- **Channel:** data from one color (Cy3 = cyanine 3 = green, Cy5 = cyanine 5 = red).
A Biochip Informatics Strategy

Interesting Patients → Interesting Animals → Interesting Cell Lines → Make Biochip

Appropriate Tissue → Appropriate Conditions → Extract RNA → Hybridize Biochip

Functional Clustering → Data Pre-processing → Scan Biochip

Access Significance → Biological Validation → Post-cluster Analysis & Integration → Informatical Validation?

Clinical relevance of Biochip informatics

Dx. → Discovery → Tx. → Px.
Biochip informatics: challenges

• **Pre-processing:**
  - technology variation
  - noise & data filtering
  - missing / negative values / P & A calls
  - data scaling
  - Can I assume normality?
  - chip quality, other artifacts

• **Functional Clusters:**
  - clustering quality, consistency, & robustness

• **Statistical Issues:**
  - study design / # of replicates / multiple testing

• **Integrative Biochip Informatics**
  - Can we get more out of it?

---

**Why integrative approach?**

Variables (10’s-100’s) vs. Variables (10000-100000)

High-dimensionality systems with insufficient data are extremely underdetermined
Largely unlabelled data
Not tractable by standard biostatistical techniques
Preprocessing

- technology variation
- noise & outlier detection
- missing / negative values / P & A calls
- data scaling
- Can I assume normality?
- chip quality, other artifacts

**Preprocessing: Noise**

Weighted average difference

\[ \hat{\theta} = \frac{\sum_{n=1}^{N} (PM_n - MM_n)\phi_n}{N} \]
Preprocessing: missing values

- put in zeros
- row average values
- weighted KNN
- SVD (requires iteration)

Preprocessing: Filtering

- intensity-based filter
  - floors or cut-offs: min. expression level
  - 2 s.d above (local) background
  - percentage-based cut-offs
  - also consider saturation
- low variance filter (across conditions)
- statistical filter (with replicates)
- low entropy filter / jackknife clustering

\[ H(x) = - \sum_i p(x_i) \log_2(p(x_i)) \]

- target ambiguity filter (databases)
Filtering: Low entropy filter

* Jackknife clustering
**Chip quality**

- Debris
- Scratch?
- Bubble
- Edge effect
- Background haze

**Data structure**

Gene expression level of gene 10 on slide 4

\[ \text{Gene expression level} = \log_2(\text{Red intensity} / \text{Green intensity}) \]
density plots of intensities

For the same, equally-treated samples, \( I(R_i) = I(G_i) \)

Intensity vs. intensity plot
Intensity vs. intensity plot

For different sample, what about R/G ratios? What possibly are the problems?

Log transformations

- **Pros**
  - constant c.v., i.e., mean ~ s.d.
  - useful way of handling ratio values
  - symmetry for up or down

- **Cons**
  - level of expressions & significance, low < high
  - non-positive values
  - only linear calibration transformations

- **Alternatives**
Episode - I

a frequency pattern
Is it typical or an artefact?

Episode - II

a frequency pattern
Is it typical or an artifact?
Episode - II  

a frequency pattern  
Is it typical or an artifact?

Pin tips
Normalization

Correcting systematic variation
- Simple additive and multiplicative
- Linear vs. non-linear
- Sources of error
- Kinds?
  - dye
  - spotting
  - experiment, slide
  - scale, scanning

Sources of Variability
(cDNA Array Example)
- Biological Heterogeneity in Population
- Specimen Collection/ Handling Effects
  - Tumor: surgical bx, FNA
  - Cell Line: culture condition, confluence level
- Biological Heterogeneity in Specimen
- RNA extraction
- RNA amplification
- Fluor labeling
- Hybridization
- Scanning
  - PMT voltage
  - laser power

(Geschwind, Nature Reviews Neuroscience, 2001)
Sources of error

- tissue contamination
- mRNA preparation and RNA degradation
- amplification efficiency
- reverse transcription efficiency
- hybridization efficiency and specificity
- clone identification and mapping
- PCR yield, contamination
- spotting efficiency
- Pin geometry
- DNA-support binding
- Dye labeling
- Slide variation
- other array manufacturing-related issues
- scanning
- image segmentation
- signal quantification
- ‘background’ correction

Why normalize?

To correct systematic variations such as

- To balance the fluorescence intensities of the two dyes (green Cy3 and red Cy5 dye)
- To allow the comparison of expression levels across experiments (slides)
- To adjust scale of the relative gene expression levels (as measured by log ratios) across replicate experiments
Normalization issues, cDNA chips

- Within-slide
  ✓ What genes to use
  ✓ Location
  ✓ Scale
- Paired-slides (dye swap)
  ✓ Self-normalization
- Between slides

Which genes to use?

- All genes
- Constantly expressed genes
- Controls
  ✓ Spiked controls
  ✓ Genomic DNA titration series
- Other ‘useful’ set of genes
Global normalization

Total amount of mRNA measured is constant

\[ \log_2 \frac{R}{G} \rightarrow \log_2 \frac{R}{G} - c = \log_2 \frac{R}{(kG)} \]

- Set mean or median to zero.
- Assumes \( R \) and \( G \) are related by a constant factor.
- Ignoring intensity-and-space-dependent variations.

Non-parametric smoothing: lowess

- Spread is a function of intensity.
- Regression without parametric assumption.
- on a X-Y plot.
- with a kernel function and bandwidth.
**Geometric transformations: translation**

\[
x' = x + T_x, \quad y' = y + T_y
\]

\[
P = \begin{bmatrix} x \\ y \end{bmatrix}, \quad P' = \begin{bmatrix} x' \\ y' \end{bmatrix}, \quad T = \begin{bmatrix} T_x \\ T_y \end{bmatrix}
\]

\[
\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 1 & 0 & T_x \\ 0 & 1 & T_y \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}
\]

**Geometric transformations: rotation**

\[
x = r \cos \phi, \quad y = r \sin \phi
\]

\[
x' = r \cos (\phi + \theta) = r \cos \phi \cos \theta - r \sin \phi \sin \theta
\]

\[
y' = r \sin (\phi + \theta) = r \cos \phi \sin \theta + r \sin \phi \cos \theta
\]

\[
\therefore x' = x \cos \theta - y \sin \theta, \quad y' = x \sin \theta + y \cos \theta
\]

\[
P' = R \cdot P\]

\[
\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}
\]
**Geometric transformations: scaling**

\[ x' = x \cdot s_x, \quad y' = y \cdot s_y \]

*Scaling factor*: \( s_x \) (x축으로 크기 조정), \( s_y \) (y축으로 크기 조정)

\[ P' = S \cdot P \]

*Uniform Scaling*: \( s_x = s_y \)

\[
\begin{bmatrix}
  x' \\
  y'
\end{bmatrix} = \begin{bmatrix}
  s_x & 0 \\
  0 & s_y
\end{bmatrix}\begin{bmatrix}
  x \\
  y
\end{bmatrix}
\]

\[ P' = T + P \]

\[ P' = S \cdot P \]

\[ P' = R \cdot P \]

\[ P' = (T \cdot R \cdot S \cdot T...) \cdot P \]

**Variance is a function of intensity**

![Graph showing variance as a function of intensity](image)
The M vs. A plot

\[
\begin{align*}
\begin{bmatrix} x' \\ y' \end{bmatrix} &= \begin{bmatrix} \cos(-\pi/2) & -\sin(-\pi/2) \\ \sin(-\pi/2) & \cos(-\pi/2) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\
M &= \log_2(R/G) : \log \text{intensity-ratio} \\
A &= \log_2(RG)^{1/2} : \text{mean log-intensity}
\end{align*}
\]

Intensity-dependent normalization

Dye bias is dependent on spot intensity!

\[
\log_2R/G \Rightarrow \log_2R/G - c(A) = \log_2R/(k(A)G)
\]

- Apply a robust scatter-plot smoother, lowess
- The lowess() function in R with f=20%
- Assumes roughly symmetric up- and downs at all intensity levels
Nonparametric smoothing

- **Smoothing**
  - Consider X Y plot
  - Draw a regression line which requires no parametric assumptions
  - The regression line is not linear
  - The regression line is totally dependent on the data

- **Two components of smoothing**
  - Kernel function, calculating weighted mean
  - Bandwidth, a window span determining smoothness of the regression line

Nonparametric smoothing

- **Types of kernel functions**
  - Uniform, Triangular, Normal, Others

- **Bandwidth**
  - The wider, the smoother
  - Bigger impact than the kernel function
Roughly equal number of genes are up- or down-regulated at all intensity levels (or only few genes are expected to change).

**Global vs. print-tip-group lowess**

Roughly equal number of genes are up- or down-regulated at all intensity levels (or only few genes are expected to change)
Global vs. print-tip-group lowess

For every print-tip group, changes roughly symmetric at all intensity levels

Within print-tip-group box plots for print-tip-group normalized M

Figure 2. Within-diff normalization. Boxplot displaying the log-ratio distribution after within-print-tip-group location normalization for each of the 15 print-tip groups. The array was printed using a 4 by 4 print-head and the print-tip groups are numbered first from left to right then from top to bottom starting from the top left corner (data from apo A1 knock-out mouse #8 in experiment (A)).
Taking scale into account

Assumptions: all print-tip-group have the same spread

True ratio is \( \mu_y \).
The observed ratio is \( M_y = a \mu_y \).
A robust alternatives to the latter estimator is

\[
\hat{a}_i = \frac{MAD}{\sqrt{\prod_i MAD_i}}
\]

where the MAD(Median Absolute Deviation) is defined by

\[MAD_i = \text{median}_j \{ |M_y - \text{median}_j(M_y)| \}\]

Location / scale calibration
**Location + scale normalization**

(a) (b)

Figure 2. Within-slide normalization. Spatial plot of the array highlighting the spots with the largest 5% absolute log-ratios. The different shades of red represent positive log-ratios and the different shades of green represent negative log-ratios. The plot is divided into 16 grids representing the 16 different print-tip groups. Each small rectangular cell represents the log-ratio of a spot on the array. (a) Extreme log-ratios after within-print-tip-group location normalization but before scale adjustment. (b) Extreme log-ratios after within-print-tip-group location and scale normalization (data from apo A1 knock-out mouse #8 in experiment (A)).

---

**Effect of normalization**

- X404G Signal, Median Unnormalized
- X404G Signal, Median Normalized
- X841G Signal, Median
- X841G Signal, Median Normalized
- X841G Signal Mean
- X841G Signal Mean Normalized
Figure 7. Within-slide normalization. Density plots of the log-ratios before and after different normalization procedures. The solid black curve represents the density of the log-ratios before normalization. The red, green, blue, and cyan curves represent the densities after global median normalization, intensity dependent location normalization, within-print-tip-group location normalization, and within-print-tip-group scale normalization, respectively (data from spo A1 knock-out mouse #8 in experiment (A)).
Non-linear normalization

Robust Non-linear Normalization

• A rank score for more robust identification of non-differentially expressed genes

\[ R_k = \sum_{i,j=1}^{p} (x_{ki} - x_{kj})^2 \]

Assumption: there are non-differentially expressed genes at all range of median expression level
\[
\text{arsinh}(x) = \log\left( x + \sqrt{x^2 + 1} \right)
\]

\[
\lim_{x \to \infty} (\text{arsinh } x - \log x - \log 2) = 0
\]

**Differential expression**

Form a statistic from the central value and spread.

deviation, sum of deviation, variance, standard deviation
Statistical testing

- Form a statistic (such as T) (for each gene) from the data
- Calculate the null distribution(s) for the statistic
- Choose the rejection region
- Compare the statistic to the null distribution(s) of the statistic
- Assigning a score

Form a statistic

$u_{jk}(i)$: the log$_2$R/G value of the $j$-th gene on the $k$-th array in the two groups($i=1,2$)

- **log fold-change**: $\bar{u}_j(2) - \bar{u}_j(1)$
- **$T$**: $(\bar{u}_j(2) - \bar{u}_j(1))/s_j$
- **Wilcoxon (rank sum)**: $r_j = \Sigma r_{jk}$
- **SAM (shrunken centroid)**: $(\bar{u}_j(2) - \bar{u}_j(1))/(s_j+s_0)$
- **Baldi’s (Bayesian)**: $(\bar{u}_j(2) - \bar{u}_j(1))/\sqrt{\tau((1-w)s_j^2+wS_0^2)}$

* Depends on the way how the pop. variability is accounted and how to borrow strength across genes.
**Diagnostic test characteristics**

<table>
<thead>
<tr>
<th></th>
<th>$D+$</th>
<th>$D-$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T+$</td>
<td>$a$</td>
<td>$b$</td>
</tr>
<tr>
<td>$T-$</td>
<td>$c$</td>
<td>$d$</td>
</tr>
<tr>
<td></td>
<td>$a+c$</td>
<td>$b+d$</td>
</tr>
</tbody>
</table>

- **Sensitivity** = $a / (a+c) = p(T+|D+)$
- **Specificity** = $d / (b+d) = p(T-|D-)$
- **Positive Predictive Value** = $a / (a+b) = p(D+|T+)$ = 1 - FDR
- **Negative Predictive Value** = $d / (c+d) = p(D-|T-)$ = power
- $(\alpha)$ rejected a true null, $(\beta)$ fail to reject a false null

**Example**

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>99.99%</td>
<td>99.9%</td>
</tr>
</tbody>
</table>

예제) **Sensitivity=99.99%, specificity=99.9%** 인 최신의 에이즈 검사가 개발되었다. 철수는 이 검사에 양성반응을 보였다. 철수가 에이즈에 감염되었을 확률은 얼마인가? (현재 한국인의 에이즈 유병률은 0.0001이라고 한다.)

1. 99%
2. 95%
3. 80%
4. 50%
5. 10%
Diagnostic test characteristics

- Sensitivity = 99.99%
- Specificity = 99.9%

<table>
<thead>
<tr>
<th>AIDS</th>
<th>no AIDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>T+</td>
<td>9,999</td>
</tr>
<tr>
<td>T-</td>
<td>1</td>
</tr>
</tbody>
</table>

Prevalence = 0.0001

- Positive predictive value = 9,999 / 109,999 < 10%
- Negative predictive value = 1.0
**ROC curve**

**Multiple testing problem**

- Thousands of hypotheses are tested simultaneously
- The gene is not differentially expressed vs. there is no gene that is differentially expressed.
- Increased chance of false positives ($\alpha$, Type-I)
- Should adjust your p-values

**FWER (Family-wise error rate):** prob. of at least one false positive

✓ $FWER = Pr(\alpha > 0)$

**FDR (False discovery rate):** expected proportion of false positives among the rejected hypotheses (’95)

✓ $FDR = E(\alpha / rejected) \cdot Pr(\text{rejected} > 0)$

✓ Positive FDR, $pFDR = E(\alpha / \text{rejected} | \text{rejected} > 0)$
Multiple hypothesis testing

Biochip Informatics - (I)

- Biochip basics
- Preprocessing
- Episodes 1 and 2
- Global normalization
- Intensity dependent normalization
- Controlling regional variation
- Alternatives
- Differential expression
- Multiple hypothesis testing
- Classification
Multiple testing in microarray

- Microarray experiments are large and exploratory
- 5% of FDR says that, among the 100 genes said significant about 95 may be truly significant.
- What dose 5% of FWER in microarray experiments?
- ... and compared to the top (arbitrary) 100 list?
- What about symmetric vs. asymmetric rejection regions?
- Robustness against the kind of dependence?
- ....

Control of the FWER

Bonferroni correction
- complete null hypothesis that there is no gene that is differentially expressed (i.e., a+c = a+b+c+d)
- weak control of type-I error
  ✓ FWER = \[ Pr(\alpha > 0) = Pr(\text{at least one adjusted } p_g < c|H_0) \]
- ? dependence structure

Westfall/Young’s minP adjusted p-values
- by re-sampling
- strong control of type-I error (i.e., regardless of a+c)
- step-down procedure
- maxT
Control of the FWER

Golub’s leukemia data

- Statistic: \( d_j = (\bar{u}_j(2) - \bar{u}_j(1))/(s_j + s_0) \)
- \( s_0 \): fudge factor
  - a small positive number chosen as the percentile of the \( s_j \) values that makes the C.V. of \( d_j \) approximately constant as a function of \( s_j \).
  - Also dampens large values of that arise from genes with low expression
- Estimate the number \( m_0 \) of invariant genes.
- Taking B permutations, compute the mean number of significant genes in B.
- \( \hat{E}(V) = \#\{d_j: \text{gene} \text{ unchanged and } d_j \geq t_2 \text{ or } d_j \leq t_1\} \cdot m^*/m \)
- Estimated FDR = \( \hat{E}(V)/R \)

FDR and SAM
The SAM procedure

- Compute the ordered statistics $d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(J)}$
- For $b=1,\ldots, B$ (randomly) permute the class labels, compute test statistic $d_{(b)}$ and the corresponding order statistics $d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(J)}$

From the set of $B$ permutations, estimate the expected order statistics by $\hat{d}_{(j)} = (1/B) \sum_b d_{(j)}^{(b)}$ for $j=1,\ldots,J$
- Plot the $d_{(j)}$ values versus the $\hat{d}_{(j)}$. For fixed threshold $\triangle$, determine $t_1(\triangle)$ and $t_2(\triangle)$.
- Estimate the FDR
Molecular Classification of Cancer

Golbe, et al., Science
1999;403:503-511
Machine Learning Approach in Bioinformatics

- **Supervised Machine Learning**
  - Linear Discriminant Analysis / Logistic Regression / PCA
  - Classification Tree
  - Artificial Neural Network
  - Support Vector Machine
  - Rough Sets
  - Reinforcement Learning
  - Hidden Markov Model

- **Unsupervised Machine Learning**
  - Hierarchical Tree Clustering
  - Partitional Clustering
  - Self-Organizing Feature Maps
  - Matrix Incision Algorithms

Supervised vs unsupervised classifications

- **Classification**
  - Conditional Densities
    - known
    - unknown

- **Supervised Learning**
  - Parametric
  - Nonparametric
  - “Optimal” Rules
  - Plug-in Rules
  - Density Estimation
  - Decision Boundary Building

- **Unsupervised Learning**
  - Parametric
  - Nonparametric
  - Mixture Resolving
  - Cluster Analysis

*Jain et al. 2000, IEEE Transactions*
Artificial Neural Network

A Universal Function Approximator

\[ y_i = f_i(x) \]

\[ x_i = \sum_{j \in N(i)} w_{ij} y_j + w_i \]

\[ y_i = f_i(x) = f_i\left( \sum_{j \in N(i)} w_{ij} y_j + w_i \right) \]
### Classification tree as an ex. of classification

#### Sunburn at the Beach

<table>
<thead>
<tr>
<th>NAME</th>
<th>HAIR</th>
<th>HEIGHT</th>
<th>WEIGHT</th>
<th>LOTION</th>
<th>RESULT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sarah</td>
<td>Blonde</td>
<td>Average</td>
<td>Light</td>
<td>No</td>
<td>Sunburn</td>
</tr>
<tr>
<td>Dana</td>
<td>Blonde</td>
<td>Tall</td>
<td>Average</td>
<td>Yes</td>
<td>None</td>
</tr>
<tr>
<td>Alex</td>
<td>Brown</td>
<td>Short</td>
<td>Average</td>
<td>Yes</td>
<td>None</td>
</tr>
<tr>
<td>Anna</td>
<td>Blonde</td>
<td>Short</td>
<td>Average</td>
<td>No</td>
<td>Sunburn</td>
</tr>
<tr>
<td>Emily</td>
<td>Red</td>
<td>Average</td>
<td>Heavy</td>
<td>No</td>
<td>Sunburn</td>
</tr>
<tr>
<td>Pete</td>
<td>Brown</td>
<td>Tall</td>
<td>Heavy</td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>John</td>
<td>Brown</td>
<td>Average</td>
<td>Heavy</td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>Kate</td>
<td>Blonde</td>
<td>Short</td>
<td>Light</td>
<td>Yes</td>
<td>None</td>
</tr>
</tbody>
</table>

Example from Winston (1980)

---

#### Classification tree as an ex. of classification

[Diagram]

- HAIR COLOR
  - BLONDE
  - BROWN
- LOTION USED
  - NO
  - YES + EMILY
    - ALEX
    - PETE
    - JOHN
  - + SARA
    + ANNE
  - DANA
  - KATE

Example from Winston (1993)
Classification tree as an ex. of classification

Example from Winston (1996)

Classification tree as an ex. of classification

Example from Winston (1995)
Classification tree as an ex. of classification

How about?

\[ \sum_{c} -\left( \frac{n_{bc}}{n_b} \right) \log \left( \frac{n_{bc}}{n_b} \right) \]

\( C \) represents classes in group b
\( n_{bc} \) number of cases of class c in group b
\( n_b \) number of cases in group b

Classification by the shrunken centroids

For the K class problem, i genes and j slides
- \( n_k \): number of samples in class k
- \( C_k \): indices of the \( n_k \) samples

The mean expression value of gene i in class k

Let

\[ d_{ik} = \frac{x_{ik} - \bar{x}_j}{m_k(s_j + s_\theta)} \]  \hspace{1cm} [1]

where \( s_j \) is the pooled within-class standard deviation for gene i:

\[ s_i^2 = \frac{1}{n - K} \sum_{k} \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2 \]  \hspace{1cm} [2]
shrunken centroids

classification exercise
Feature selection

K-fold cross-validation
Cross-validation probabilities

Thank you!