A New Probabilistic Framework for cDNA Microarray Data Analysis

Babatunde A. Ogunnaike, Claudio A. Gelmi and Jeremy S. Edwards*

University of Delaware, Department of Chemical Engineering

DELAWARE BIOTECHNOLOGY INSTITUTE

University of Delaware Chemical Engineering

1. Data representation and transformation

- Let I_{ti} (treated cell) and I_{ri} (reference cell) be the fluorescence signal intensities measured from each spot "i" on a microarray.
- Transform the data from a "cartesian" representation to a "polar" form, i.e.:

Cartesian form (I_{t_i}, I_{r_i}) → Polar form (R_i, x_i)

$$R_i = \sqrt{I_{t_i}^2 + I_{r_i}^2} \quad \text{(intensity magnitude)}$$

$$x_i = \frac{I_{t_i}}{I_{t_i} + I_{r_i}} \quad \text{(fractional intensity)}$$

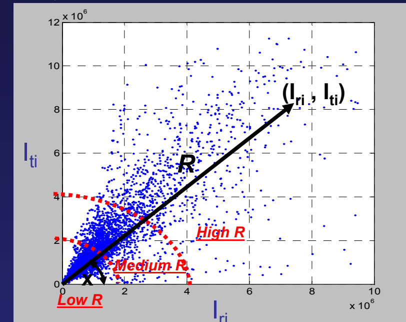

Figure 1. Intensity signal of control versus treated cell.

- **Advantages** of the polar transformation:
  - It is more *natural* (the variance of the microarray data changes with intensity measurement).
  - It normalizes the data *more efficiently* because none of the original information is lost
  - It enables statistical rigor and endows the problem with analytical tractability not possible with either raw intensity data or the popular intensity ratios.

The data can be partitioned into at least 3 groups:
- *Low R (see Fig. 1)*
- *Medium R*
- *High R*

If intensity signal ∝ (# of mRNA molecules)

Intensity measurements (I) ~ Gamma distribution

$$X_i = \frac{I_{ti}}{I_{ti} + I_{ri}} \sim \text{Beta distribution}$$

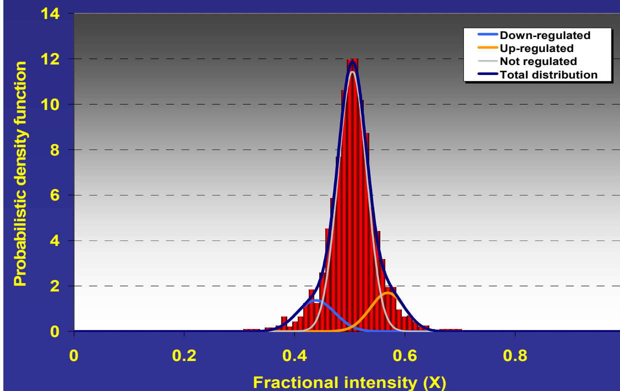

Figure 2. Histogram of fractional intensities and fitting of Beta distributions.

- The histogram of fractional intensity (X) data can be characterized as a mixture of at least three Beta distribution functions:

$$f(x) = \phi_L \cdot f_L(x) + \phi_0 \cdot f_0(x) + \phi_h \cdot f_h(x)$$

$$\sum \phi_i = 1$$

Where,

$f_l(x)$ = Beta distribution function for genes showing lower differential expression.

$f_0(x)$ = Beta distribution function for genes showing no differential expression.

$f_h(x)$ = Beta distribution function for genes showing higher differential expression.

2. Drawing inferences

(i) Bias detection and correction

Theoretically, if there is NO bias the mean of $f_0(x)$ should be $\mu_0 = 0.5$

Bias is indicated if $\mu_0 \neq 0.5$; remove via correction for x_i:

$$\tilde{X}_i = \frac{X_i}{X_i + w \cdot (1 - X_i)}$$

where $w = \mu_0/(1-\mu_0)$

(ii) Probability of expression status

Using the probability distribution functions it is possible to calculate the probability of expression status (down, up or not regulated) for every gene.

Example: The probability that gene g_i with a fractional intensity x_i is down-regulated (D) is given by:

$$P(g_i \in D) = \frac{\phi_l f_l(x_i)}{\phi_l f_l(x_i) + \phi_0 f_0(x_i) + \phi_h f_h(x_i)}$$

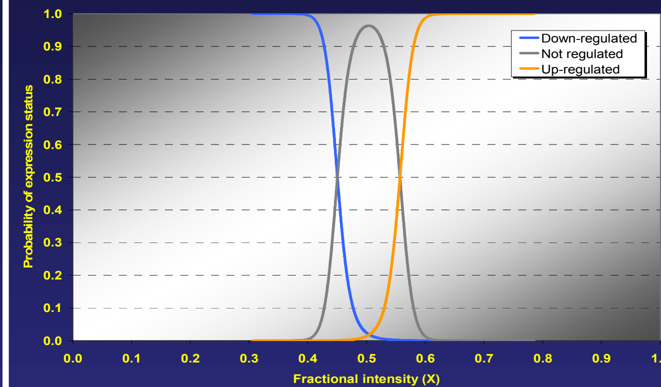

Figure 3. Probability of expression status.

- **Figure 2** shows the probability of expression status (down, up or not differentially regulated) for each gene as a function of the fractional intensity (X).

3. Application results

- Experimental data from gene expression studies in *Deinococcus radiodurans* following DNA damage (carried out by Dr. John Battista of LSU).

- **Table 1** shows the identification of relevant genes using classical fold-change criteria. These results are compared before and after bias correction using our probabilistic framework.

- **Table 2** shows the number of candidates genes identified per group (low, medium and high R). All of them with a very high probability of expression status.

Table 1. "Fold-change" criterion results.

	2 Fold-change	3 Fold-change
Total genes identified	**743**	**192**
Up-regulated	145	77
Down-regulated	598	115
With Bias correction		
Total genes identified	**429**	**131**
Up-regulated	233	104
Down-regulated	196	27
False positives	402	88
False negatives	88	27

Table 2. Probabilistic framework results.

	Count	x value Thresholds	Equivalent "Fold change"	Associated Probabilities
(High R group)				
Total genes identified	**68**			
Up-regulated	68	$x \geq 0.83$	4.74	prob ≈ 1.0
Down-regulated	N/A	N/A	N/A	N/A
(Med R group)				
Total genes identified	**7**			
Up-regulated	2	$x \geq 0.81$	4.18	prob = 1
Down-regulated	5	$x \leq 0.28$	0.38 (2.64)	prob = 1
(Low R group)				
Total genes identified	**17**			
Up-regulated	6	$x \geq 0.74$	2.85	prob = 1
Down-regulated	11	$x \leq 0.28$	0.38 (2.62)	prob ≈ 1

4. Future work

- Extend to GeneChip® technology.
- Compare different microarray technologies; develop appropriate metrics for assessing the "quality".
- Incorporate degree of confidence in the inference regarding the probability of expression status.

(*) : ogunnaik@che.udel.edu