

Mixture Model Analysis of DNA Microarray Images

Blekas, Galatsanos, Likas and Lagaris

Presentation by *Warren Cheung*

Overview

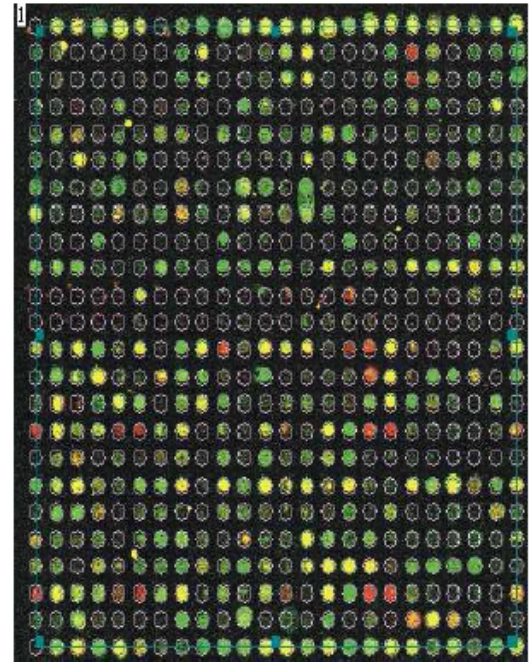
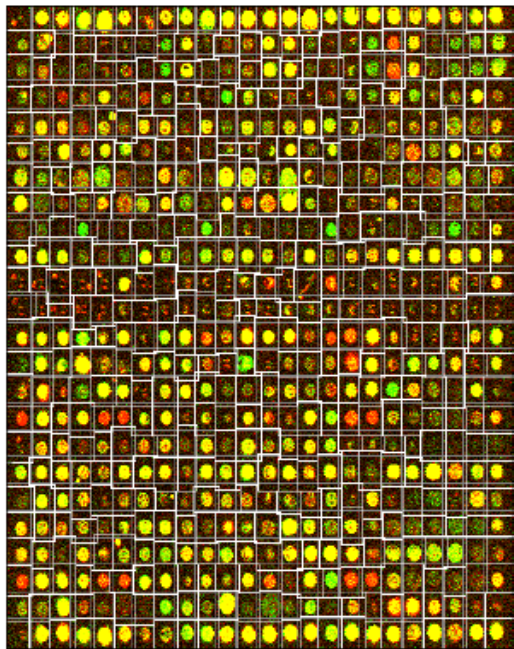
- Data — DNA Microarrays
- Segmentation Tasks
 - gridding
 - spot analysis
- Experimental Evaluation

DNA Microarrays

- DNA “chip” experiments
- samples of DNA (RNA, protein, etc...) placed on the chip
- “wash” with reporter molecules
 - fluorescent in two colors (red and green)
 - each reporter binds with a unique target
- take image

Result

- intensity of color indicates amount of target present
- “high-throughput” — can do many spots on a tiny chip
- need software to analyse these images



Two Segmentation Problems

- *gridding*: separate the spots from each other (global and local segmentation)
- *spot analysis*: separate the spots from the background (Gaussian Mixture Model)

Gridding

- generates a rectangular, axis-aligned boundary for each spot
- First pass: Global boundary
 - for each row, sum the intensities of every pixel in the row
 - find the minimum value between peaks, these are boundaries
 - repeat for the columns

Gridding (2)

- Second pass: refine the boundaries
 - consider two side-by-side spots
 - within the global row boundaries for these two spots, find sum of pixel intensities for each column
 - move the column boundary to the minimum
 - repeat for rows

Spot Analysis

- wish to label each pixel
- (B) Background
- (F) Foreground
- (A) Artifact

Gaussian Mixture Model

- treat this as a clustering problem
- use a mixture of K density functions
- $K = 2$: (B) and (F)
- $K = 3$: (B) and (F) and (A)
- use EM and MAP to optimise
- use cross-validated likelihood to choose K

EM: Expectation Maximisation

- maximization of the likelihood
- start from an initial guess
- iteratively improve the model
- E-step: compute “likelihood”
- M-step: find “best” modification of model parameters

EM on the basic model

- Basic model: probability that pixels with a certain intensity has a particular label is given by a Gaussian
- E-step: given current model of Gaussians $(\mu_j^{(t)}, \Sigma_j^{(t)})$, compute posterior
- M-step: update Gaussians (compute new mean, covariance matrices) and mixing weights

Cross-Validated Likelihood

- partition the data into u sets $X_s = X_1, \dots, X_u$
- for each X_s compute the model using the data $X - X_s$
- check by computing the likelihood of X_s
- get an average score over the u sets
- repeat for $K = 3$

MAP (Maximum a posteriori) estimation

- wish to take advantage of spatial information
- add an additional probability distribution — probability that a position has a particular label
- use a Markov random field model
- maximise the (log) density function using EM
- M-step involves solving quadratic programming (QP) problem

Evaluation of Experiments

- artificially created images
- publicly available microarray databases

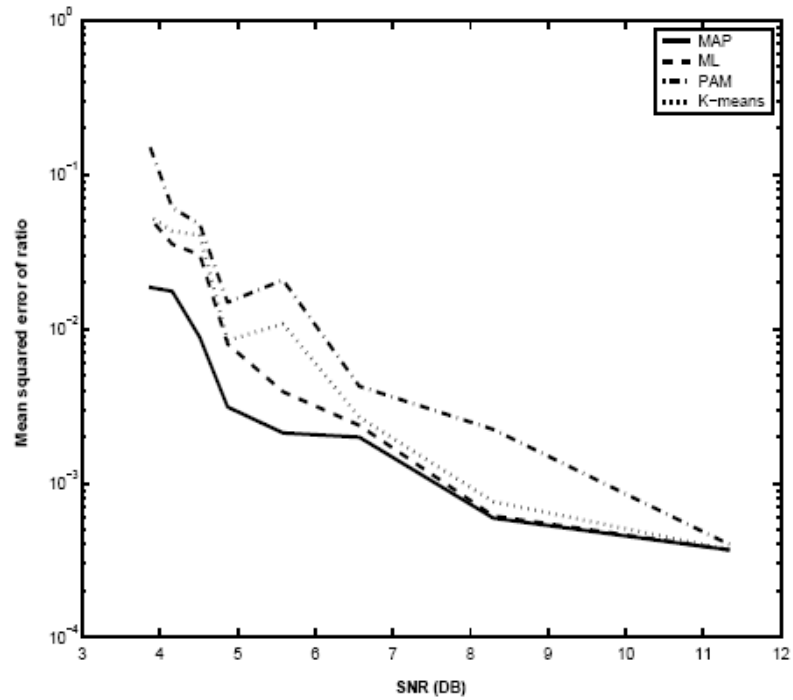
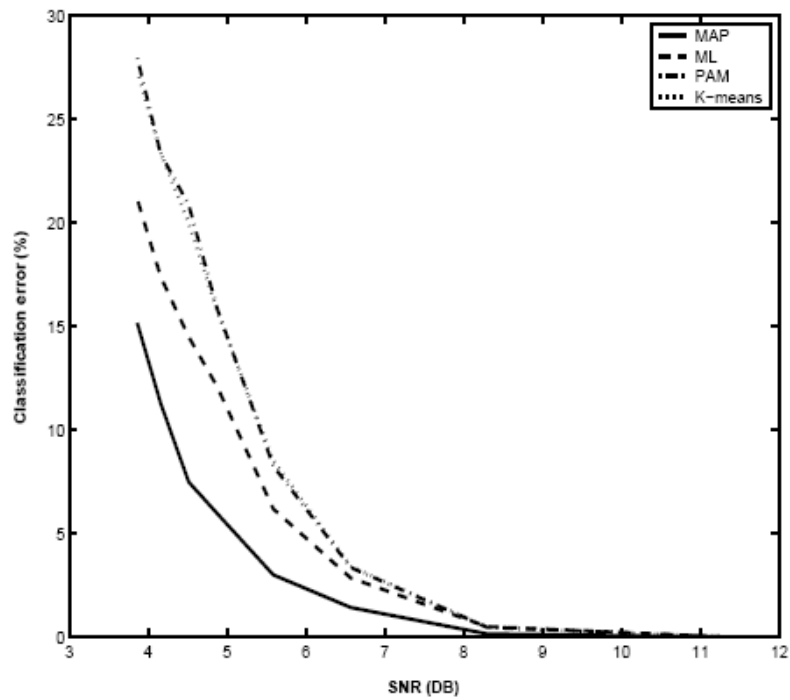
Gridding

- evaluated by visual inspection (“gold standard”)
- 3 grades: Perfect (entire spot), marginal (> 80%) or incorrect
- compare against 2 other tools

	Proposed	Spotfinder	Scanalyze
Perfect	89.6	72.8	48.7
Marginal (> 80%)	9.2	14.3	22.6
Incorrect	1.2	12.9	28.7

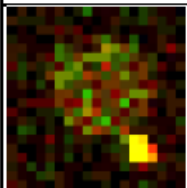


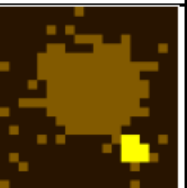



Spot Analysis - Artificial Data

- compare as noise is increased
- evaluate percentage of misclassified pixels and error in computed spot intensity
- as true result is known, they can calculate error

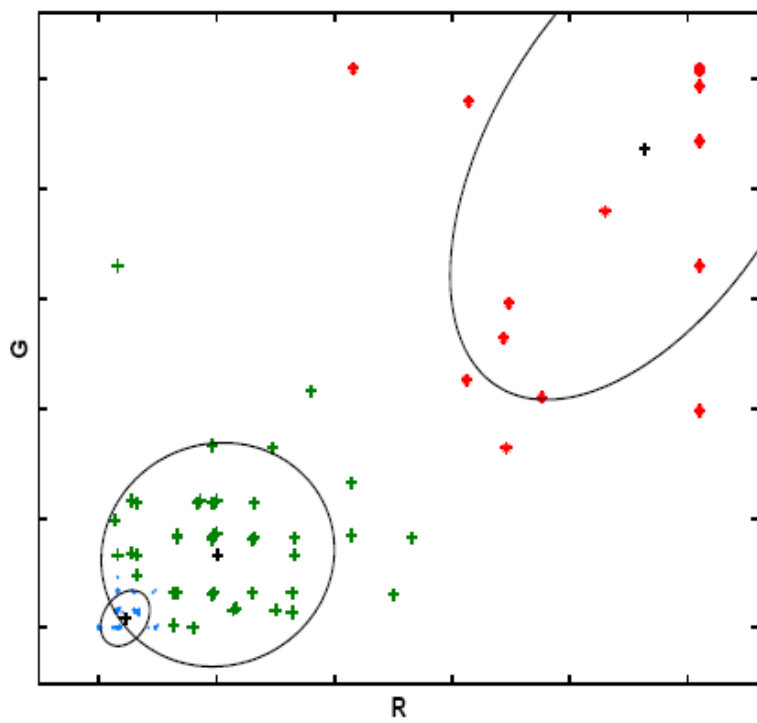


Spot Analysis - Real Data

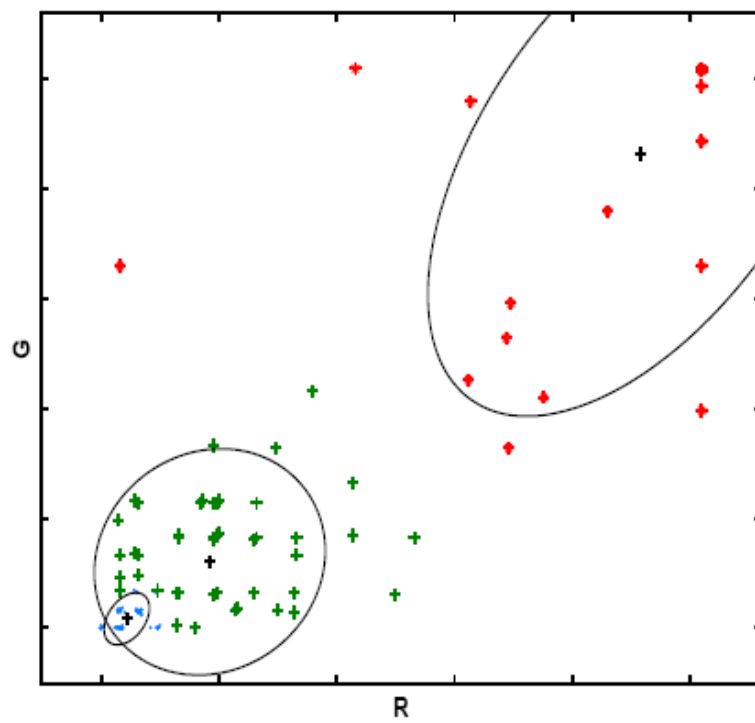
- compare spot intensity to other methods and existing tools
- argue that values are more representative of spots

<i>Original image</i>	<i>MAP-GMM</i>			<i>ML-GMM</i>	<i>K-means</i>	<i>PAM</i>	<i>Existing tools</i>
	$\beta = 0.01$	$\beta = 0.1$	$\beta = 1.0$				
							GenPix: 0.498 Spotfinder: 0.992
S_1	$r = 0.173$	$r = 0.207$	$r = 0.297$	$r = 0.175$	$r = 0.874$	$r = 0.644$	

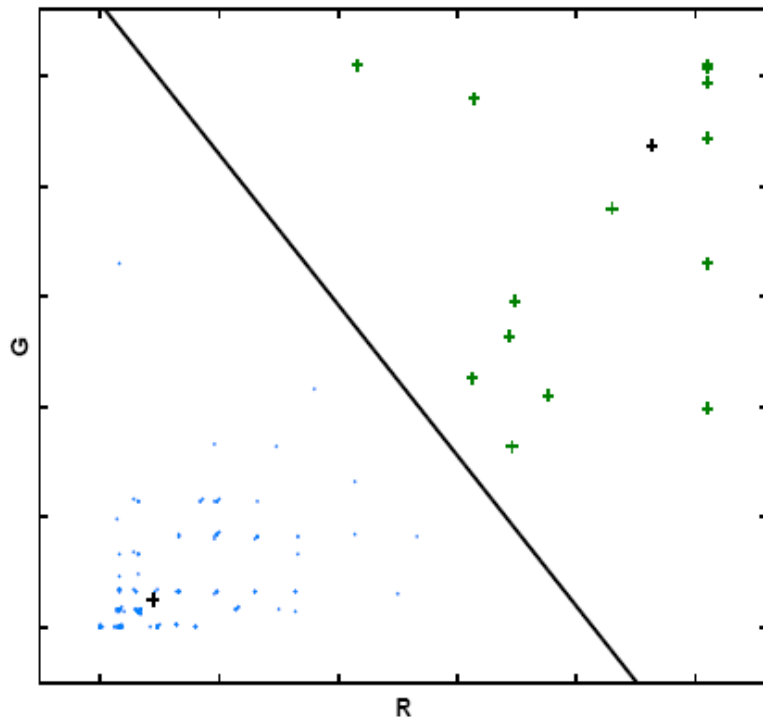
MAP-GMM



ML-GMM



K-means



Conclusion

- gridding: Exploit feature of experimental layout and knowledge of results to do rapid segmentation
- spot analysis: Iteratively improve a mixture model to fit the data