Image Processing and Analysis for Gene Microarrays

A. O. Hero

University of Michigan - Ann Arbor

http://www.eecs.umich.edu/~hero

Collaborators: G. Fleury, A. Swaroop, D. Ghosh

Outline

1. Image formation for microarrays
2. Filtered Poisson model for spotted cDNA arrays
3. Pareto filtering for gene pattern extraction
4. Application: development and aging in mouse retina

Scientific Objectives

Establish genetic basis for development, aging, and disease on the basis of genetic probes

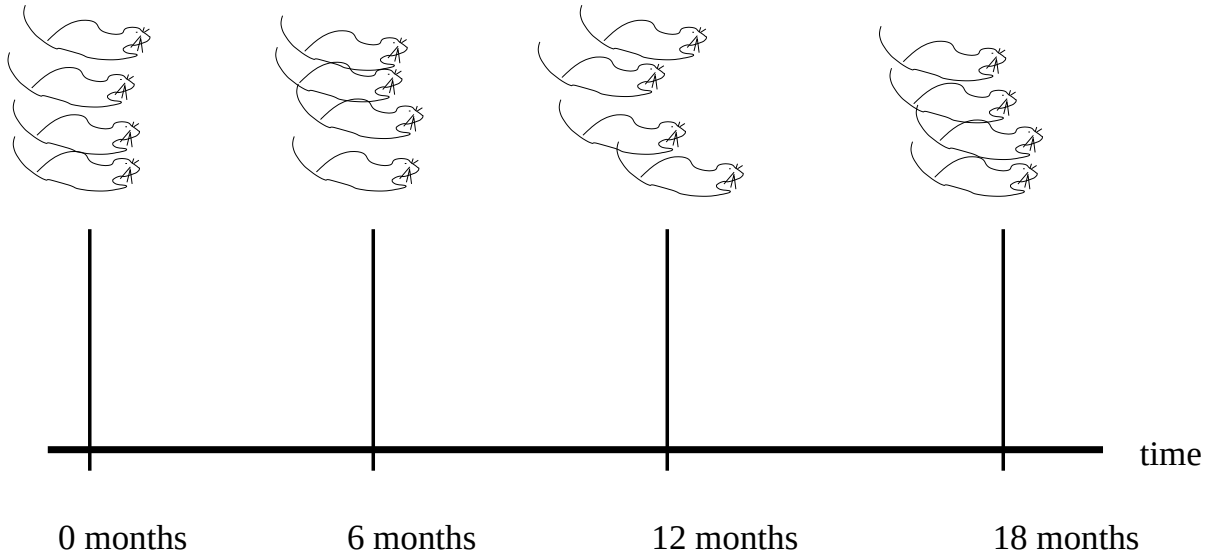

Figure 1: *Sample gene trajectories over time.*

Microarray Experiments

Figure 2: *Microarray experiment cycle.*

Image Formation

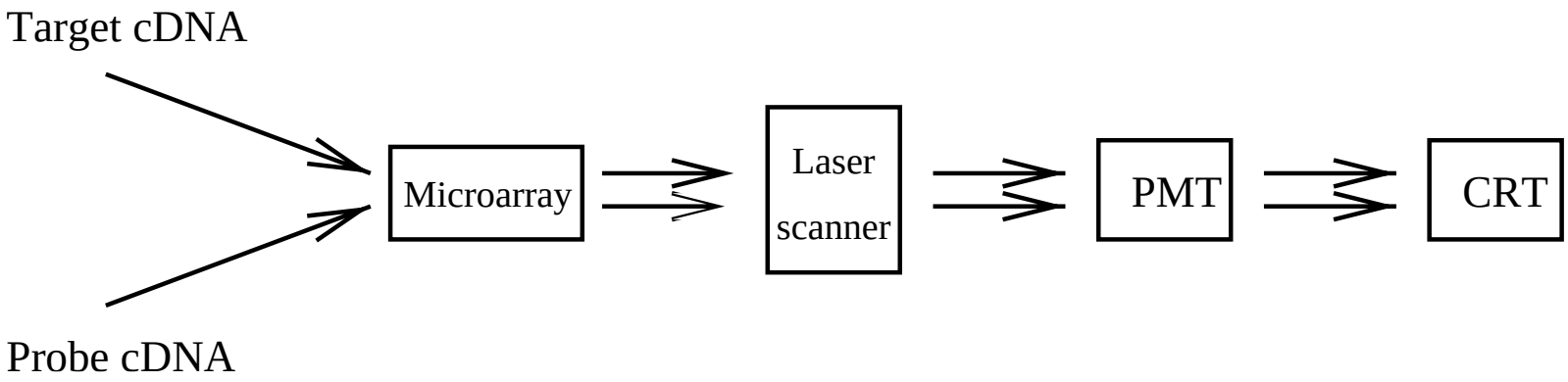

Figure 3: *Microarray image formation.*

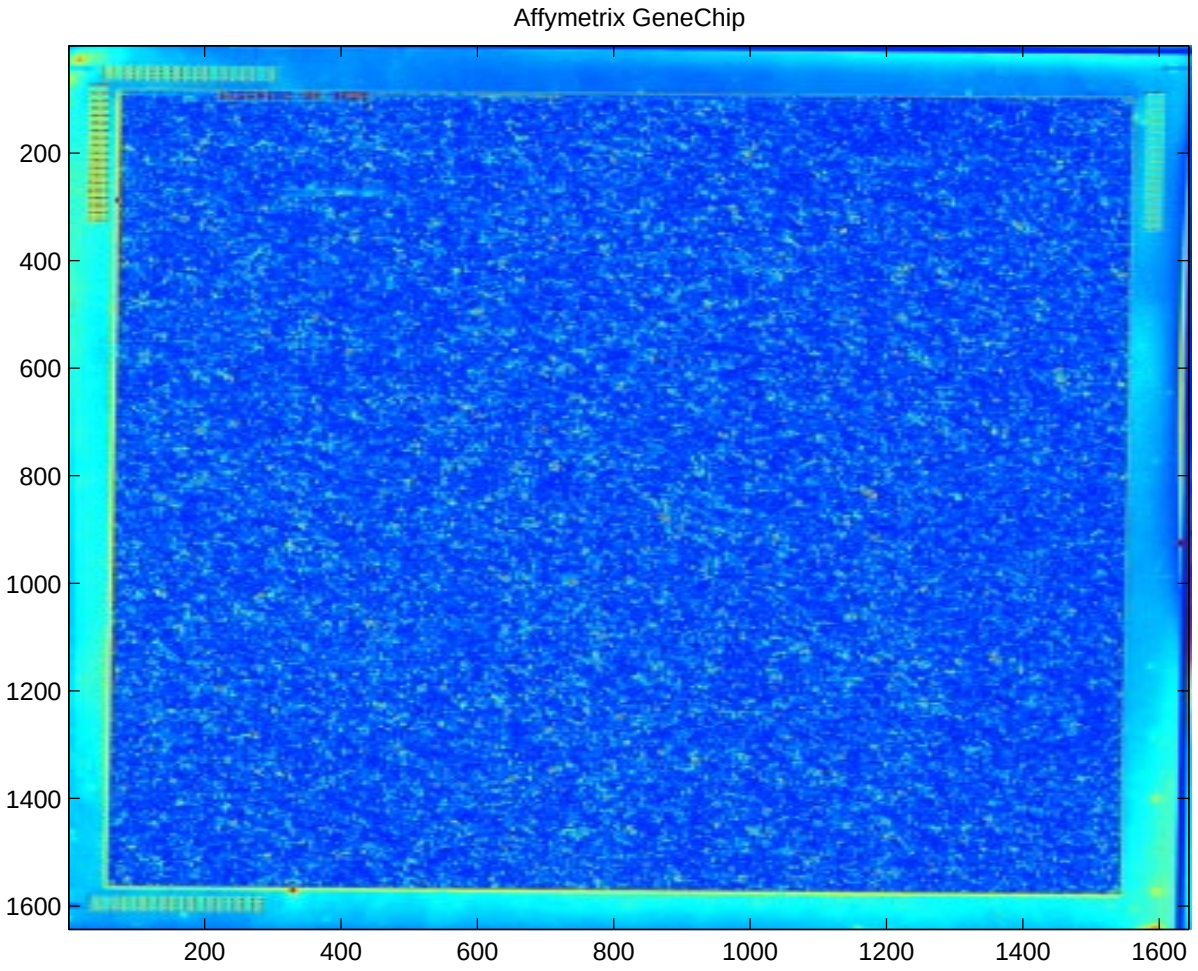

Figure 4: *Affymetrix GeneChip microarray.*

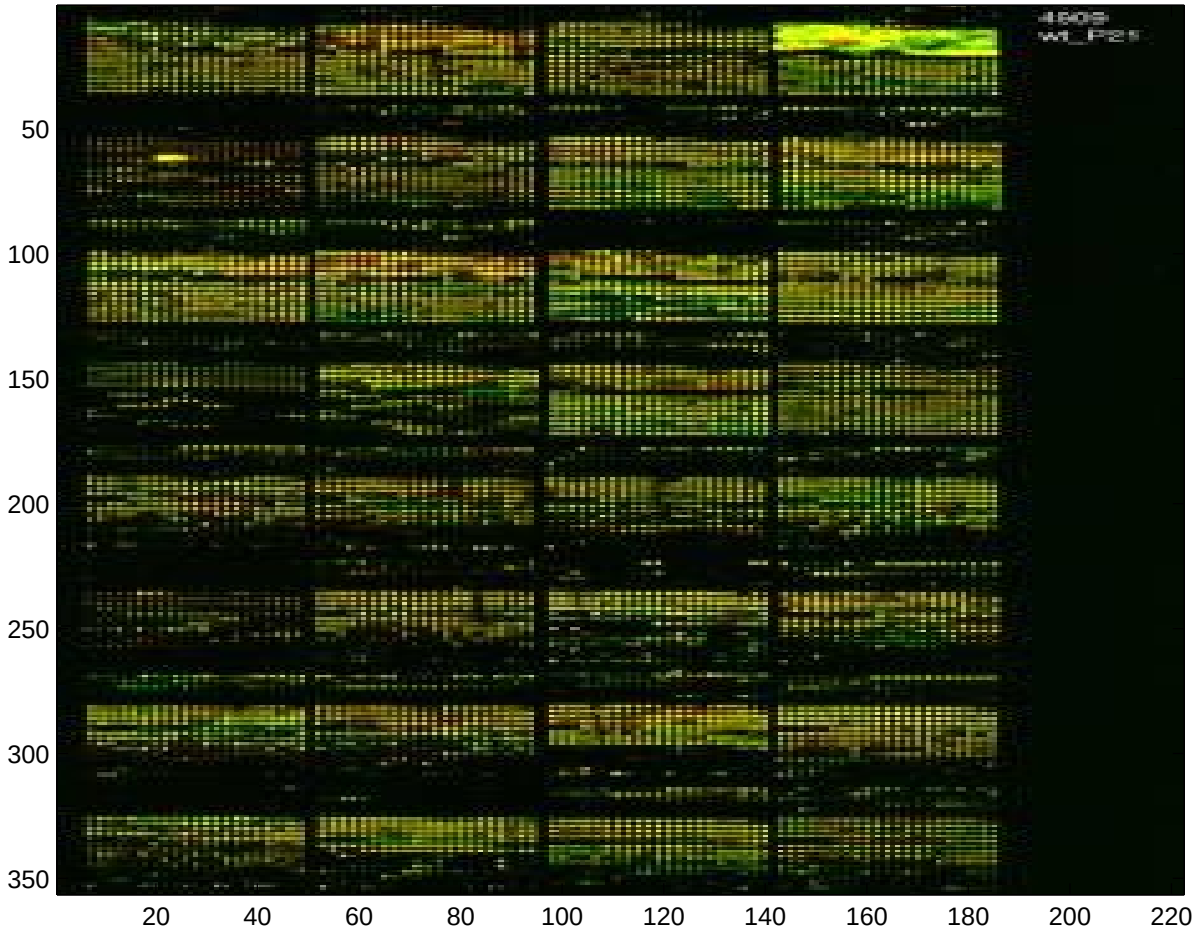

Figure 5: *cDNA spotted array.*

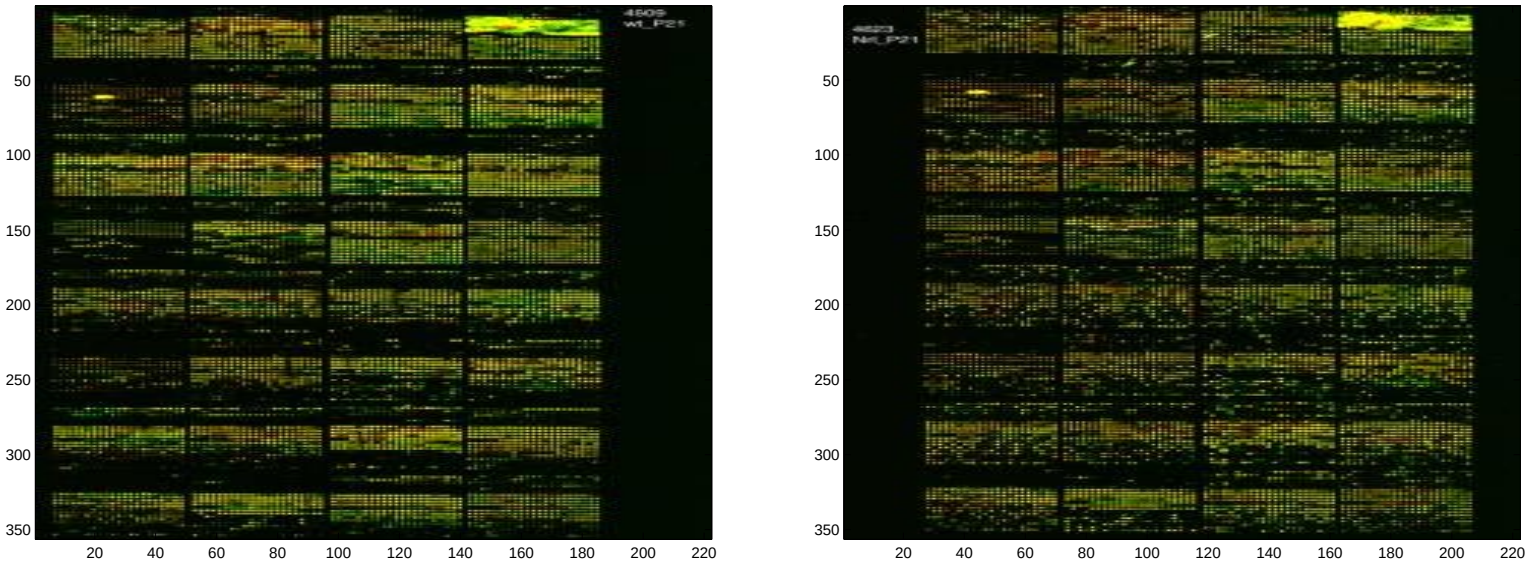

Figure 6: *cDNA spotted array (left: cy3/cy5 wildtype, right: cy3/cy5 knockout).*

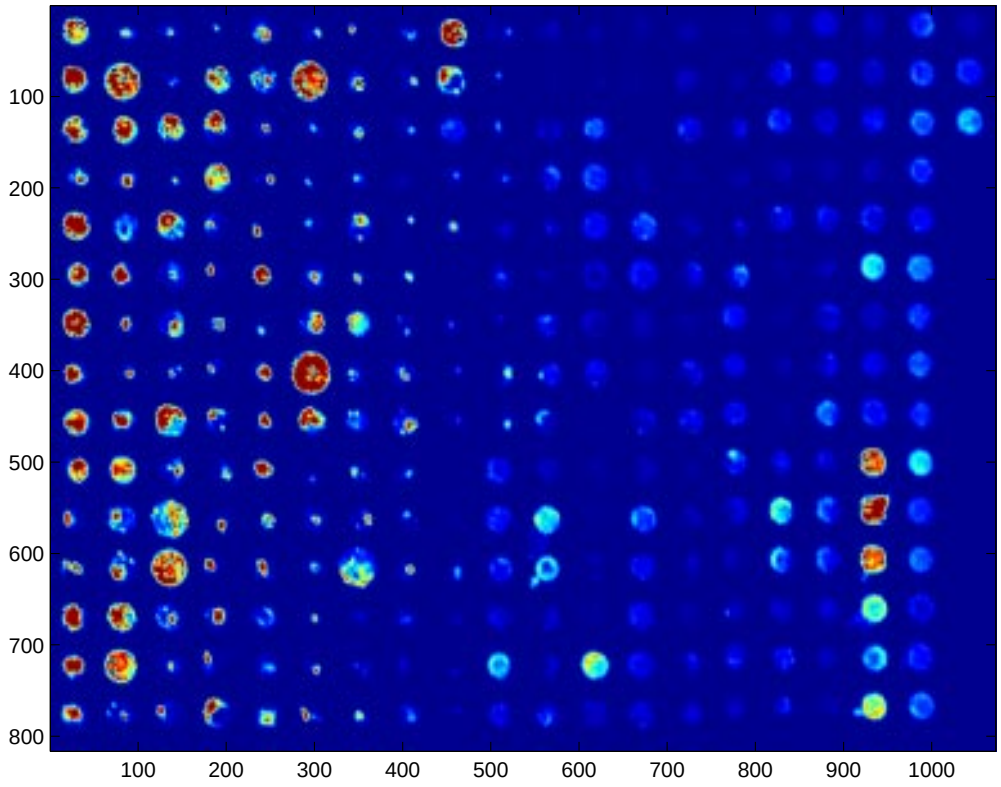

Figure 7: *Blowup of cDNA spotted array.*

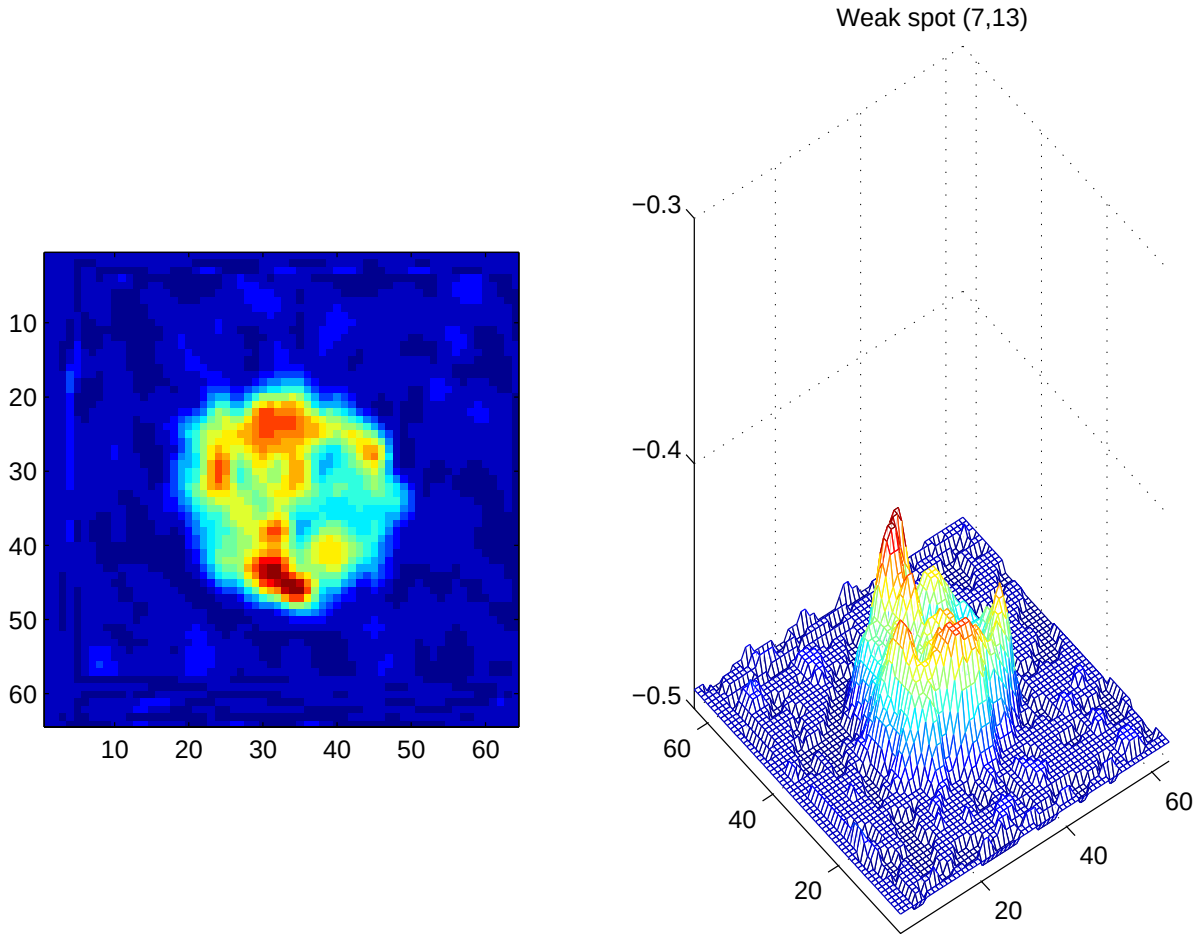

Figure 8: *Weak Spot.*

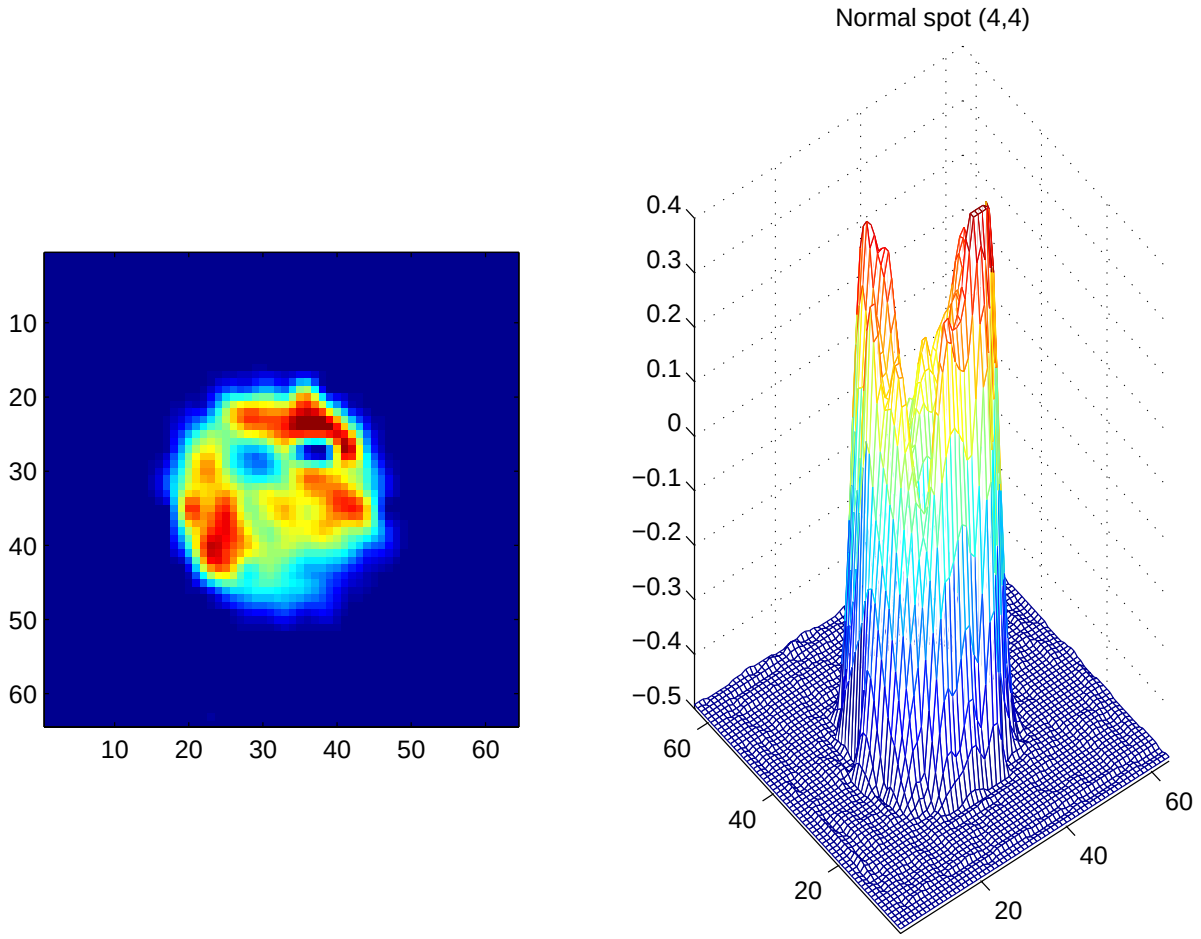

Figure 9: *Normal spot.*

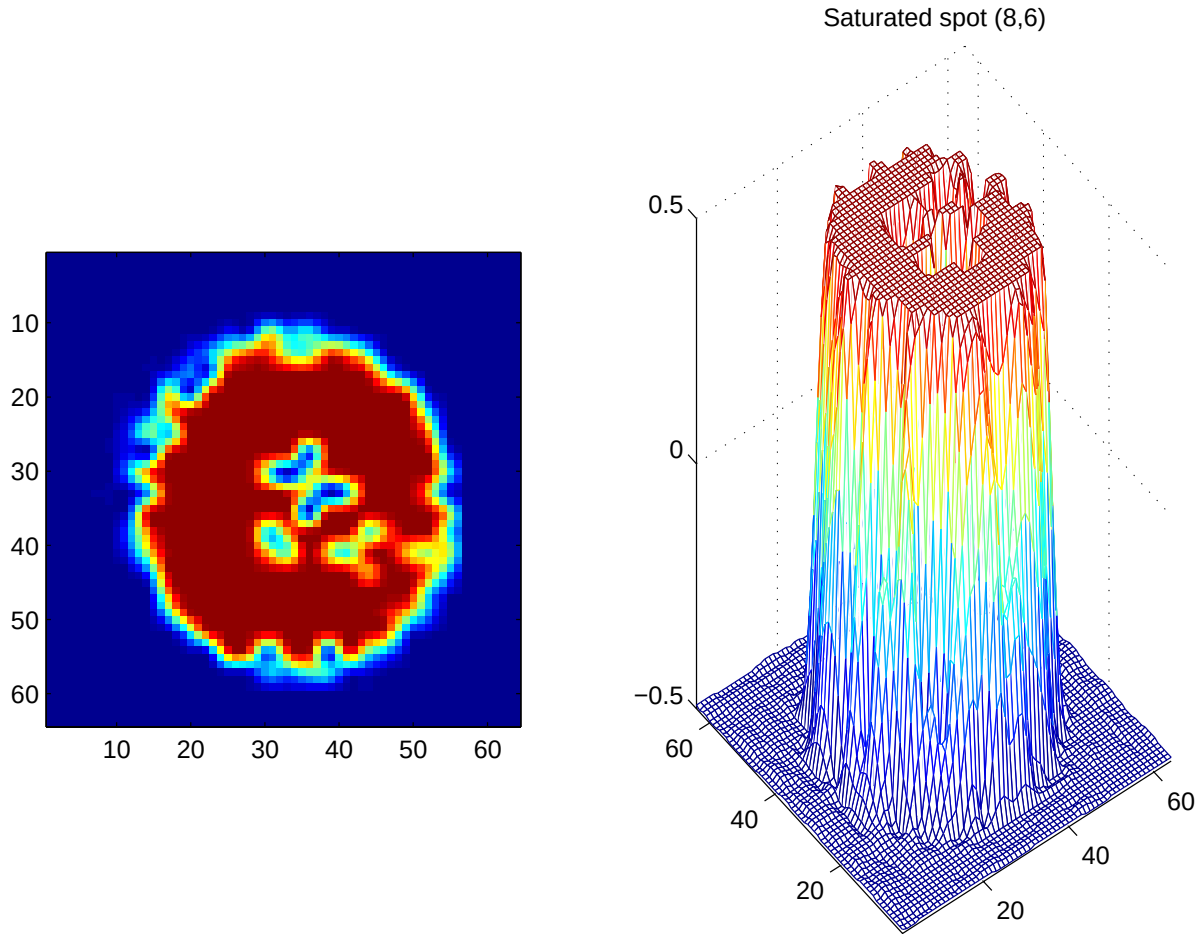

Figure 10: *Saturated spot.*

Filtered Poisson Measurement Model

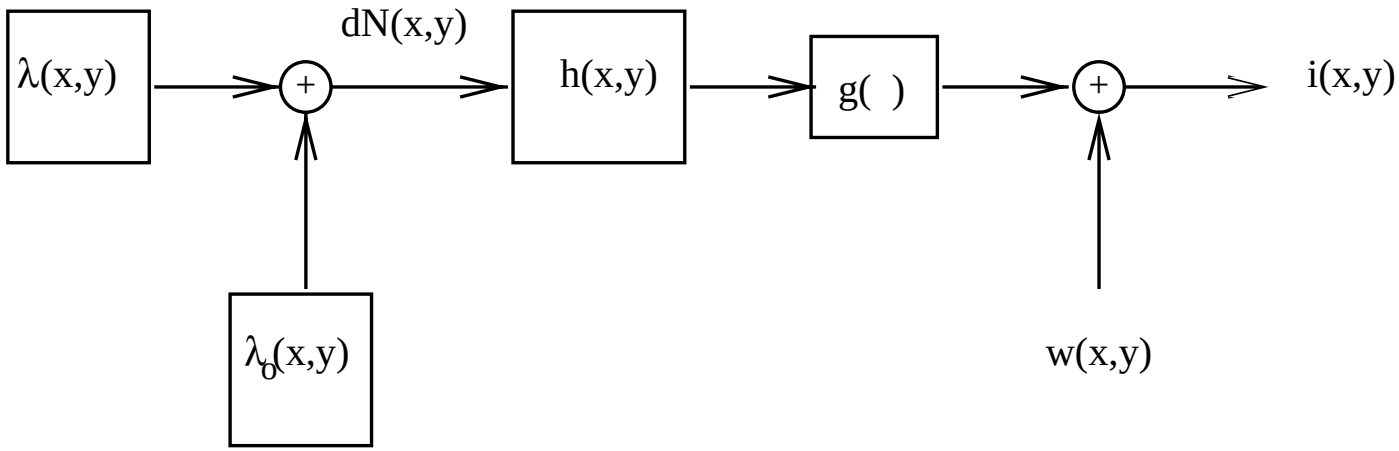

Figure 11: *Filtered Poisson model for microarray image.*

Mathematical Model

$$I(x,y) = g\left(r\int\int h(x-u,y-v)dN(u,v)\right) + w(x,y)$$

- $I(x,y)$: measured intensity
- $dN(u,v)$: inhomogeneous spatial Poisson process with intensity $\lambda_d + \lambda_o$
- $h(u,v)$: point spread function of image scanner
- g: spatially homogeneous non-linear response function
- $w(u,v)$: thermal electronic noise

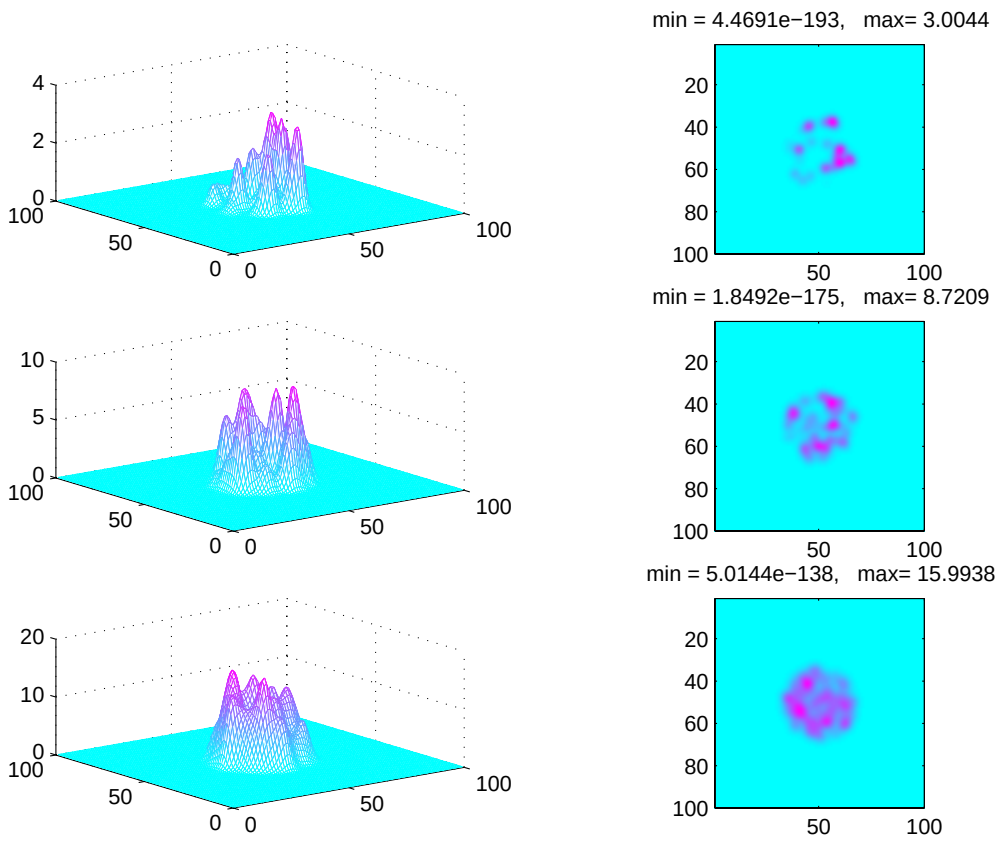

Figure 12: *Spots simulated from filtered Poisson model with Gabor components.*

Extraction of Gene Hybridization Levels

Objective: Estimate $\theta_j,\ j = 1,\ldots,\#_{probes}$

$$\lambda(x,y) = \sum_{j=1}^{\#_{probes}} \theta_j \Phi_j(x-u_j, y-v_j)$$

where

- $\Phi_j(u,v)$: (normalized) intensity of j-th spot

  Multi-component model for Φ_j

$$\Phi_j(u,v) = \sum_{k=1}^{\#basis} \alpha_{j,k}\phi_k(u,v)$$

- u_j, v_j: position of j-th spot

Compound Channel Representation

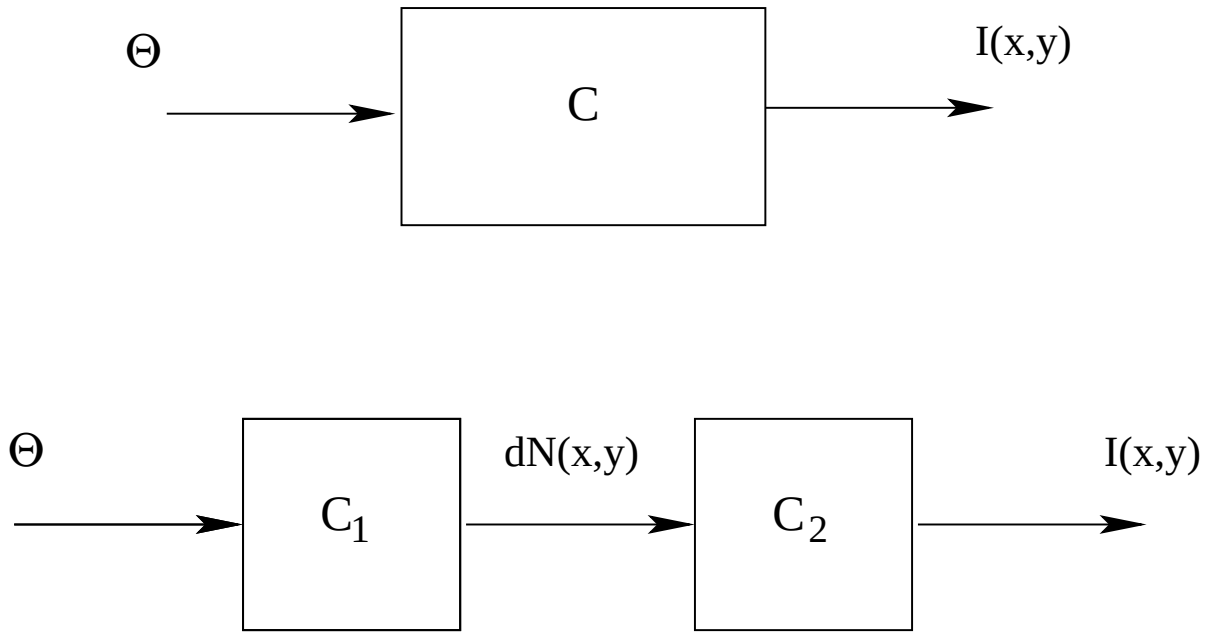

Figure 13: *Top: statistical representation of I as the output of channel C with input Θ. Bottom: decomposition of C into Poisson and Gaussian channels C_1 and C_2, respectively.*

Performance Predictions

Let Θ be a random vector of parameters:

$$E[(\Theta-\hat{\Theta}(I))(\Theta-\hat{\Theta}(I))^T] \quad \geq \quad E[(\Theta-E[\Theta|I])(\Theta-E[\Theta|I])^T]$$

$$= \underbrace{E[(\Theta-E[\Theta|dN]])(\Theta-E[\Theta|dN]])^T]}_{\text{quantum-limited cov}}$$

$$+ \underbrace{E[(E[\Theta|dN]-E[\Theta|I])(E[\Theta|dN]-E[\Theta|I])^T]}_{\text{Gauss-limited cov}}$$

Distortion-Rate Bounds

Define mutual information

$$\mathrm{MI}(I;\Theta) \quad = \quad E\left[\ln \frac{f_{I|\Theta}}{f_I}\right] = H(I) - H(I|\Theta)$$

Define rate-distortion function (monotone decreasing):

$$R(d) = \inf_{P_{I|\Theta}: MSE \le d} \mathrm{MI}(I;\Theta)$$

Define Shannon channel capacity:

$$C = \sup_{P_\Theta} \mathrm{MI}(I;\Theta)$$

Shannon's inequality

$$R(d) \le C \quad \text{or} \quad d \ge R^{-1}(C)$$

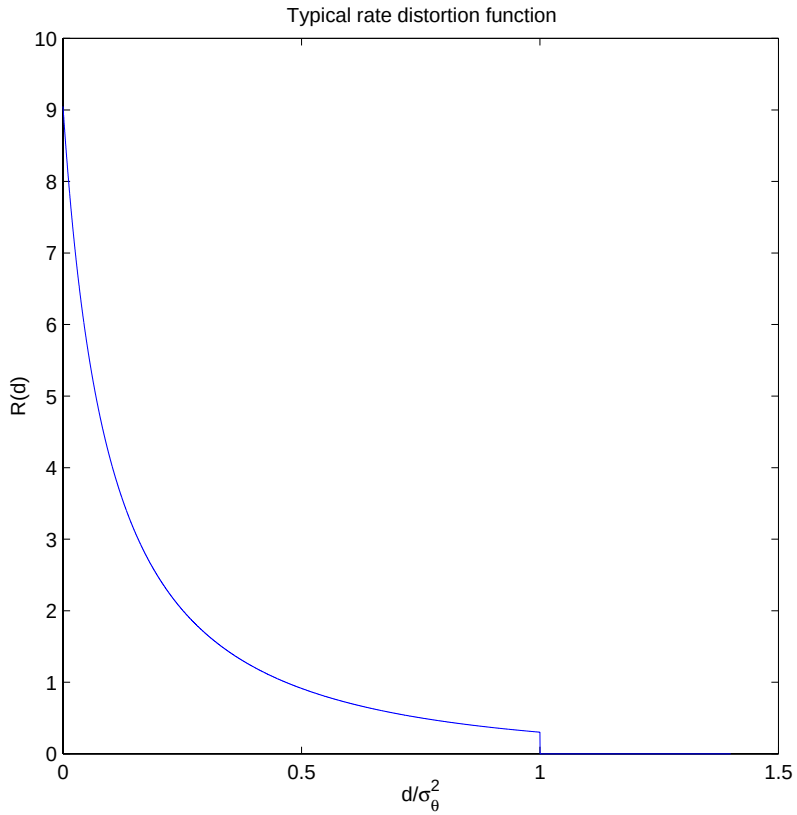

Figure 14: *Typical Rate-distortion function.*

Shannon's "Data Processing Theorem"

$$C \leq \min\{C_1, C_2\}.$$

Point process channel

$$C_1 = \sup_{P_\Theta} \mathrm{MI}(\Theta, dN)$$

Continuous process channel

$$C_2 = \sup_{P_{dN}} \mathrm{MI}(dN; I)$$

Gabor Superposition - Spot Position MSE

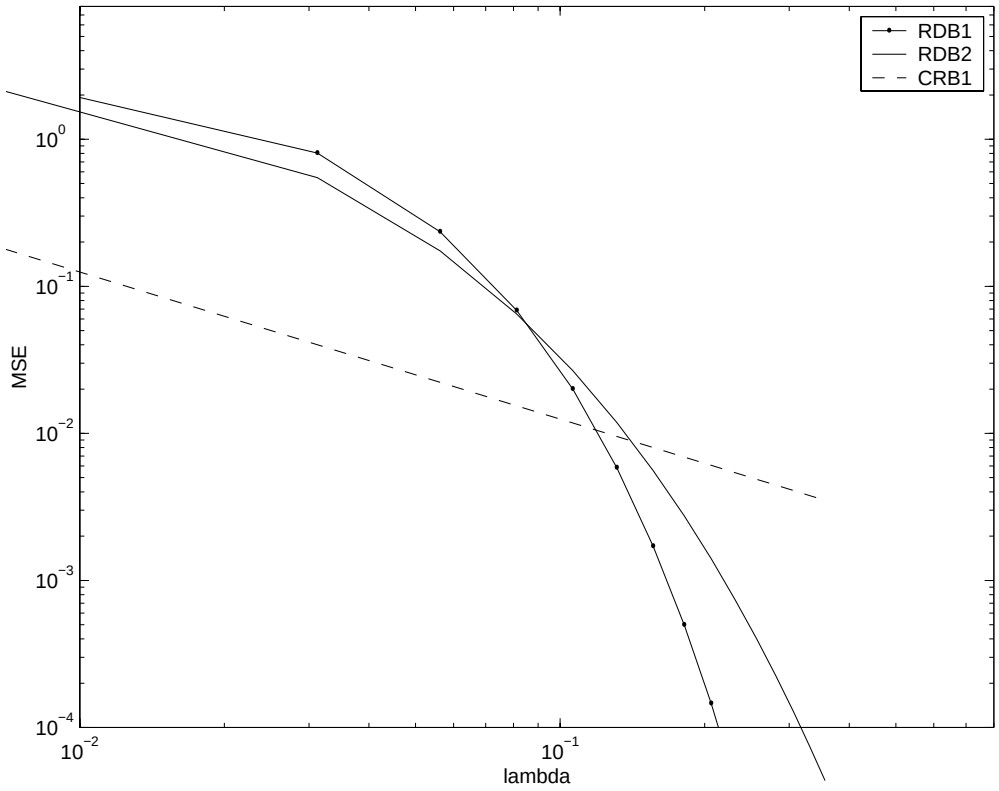

Figure 15: *MSE lower bounds on spot position.*

Gabor Superposition - Width MSE

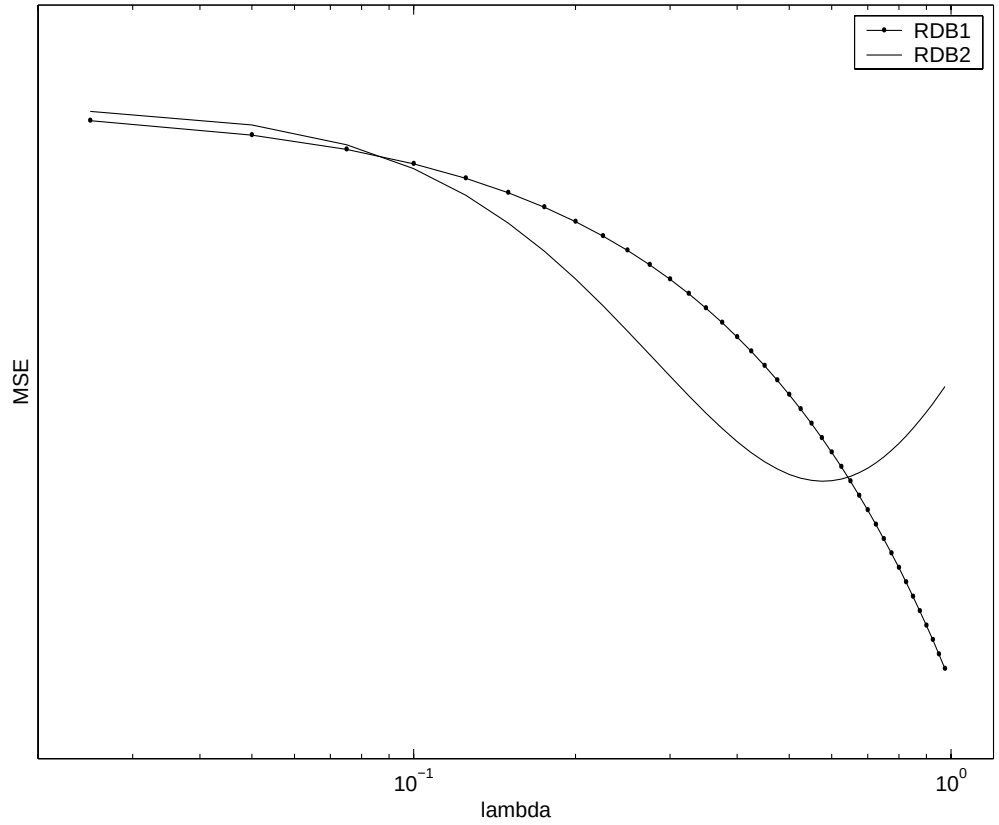

Figure 16: *Distortion-rate MSE lower bounds on Gabor widths of* $h(x,y)$.

Optimal Gene Extraction

Imputed Log-likelihood function (Antoniadis&Hero:SP92):

$$l(\theta,\alpha,u,v) = \int\int \widehat{dN}(x,y)\ln(\lambda(x,y)+\lambda_o) - \int\int \lambda(x,y)dxdy$$

where

$$\begin{aligned}\widehat{dN}(u,v) &= E[dN(u,v)|I;\ \overline{\theta},\overline{\alpha},\overline{u},\overline{v}] \\ \lambda(x,y) &= \sum_{j=1}^{\#_{probes}} \theta_j \Phi_j(x,y)\end{aligned}$$

Assuming

- $g(u) = u$
- spot intensities Φ_j don't overlap
- $\lambda_o = 0$

$$\hat{\theta}_j = \int\int_{\text{cell}_j} \widehat{dN}(x,y)$$

Extraction of Differential Hybridization Levels

For $d = 1, 2$:

$$I_d(x,y) = r_d \int\int h_d(x-u, y-v)\, dN_d(u,v) + w_d(x,y)$$

Estimate of $\Delta\theta_j = \theta_{1j}/\theta_{2j}$ is

$$\hat{\theta}_j = \frac{\int\int_{\text{cell}_j} \widehat{dN}_1(x,y)}{\int\int_{\text{cell}_j} \widehat{dN}_2(x,y)}\, \rho$$

$$\rho = \hat{r}_2/\hat{r}_1 = \frac{\sum_{j=\text{hskpg}} \int\int_{\text{cell}_j} \widehat{dN}_2(x,y)}{\sum_{j=\text{hskpg}} \int\int_{\text{cell}_j} \widehat{dN}_1(x,y)}$$

Gene Clustering and Filtering

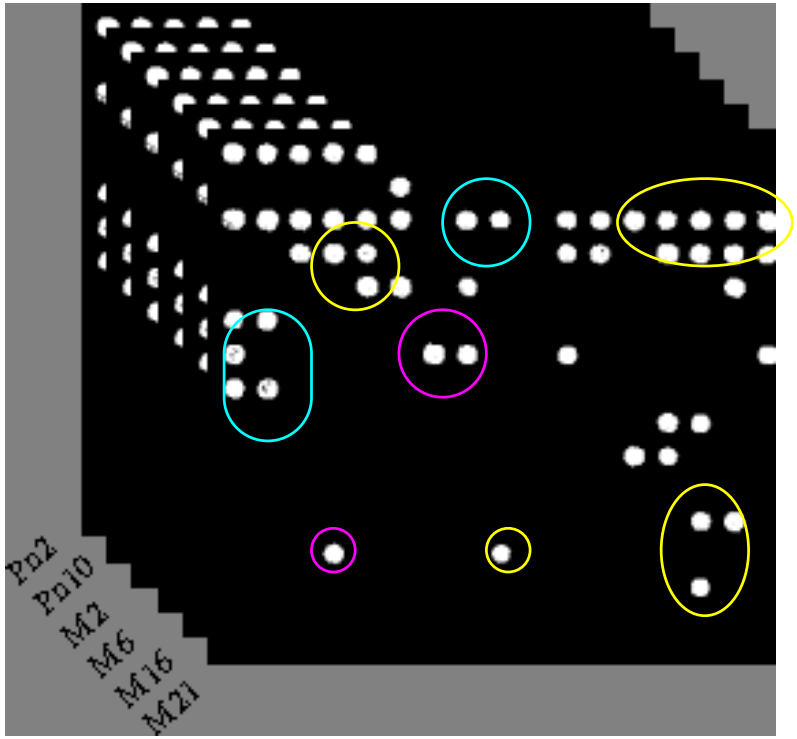

Figure 17: *Clustering on the Data Cube.*

Objective: Classify time trajectory of gene *i* into one of *K* classes

Gene Trajectory Classification

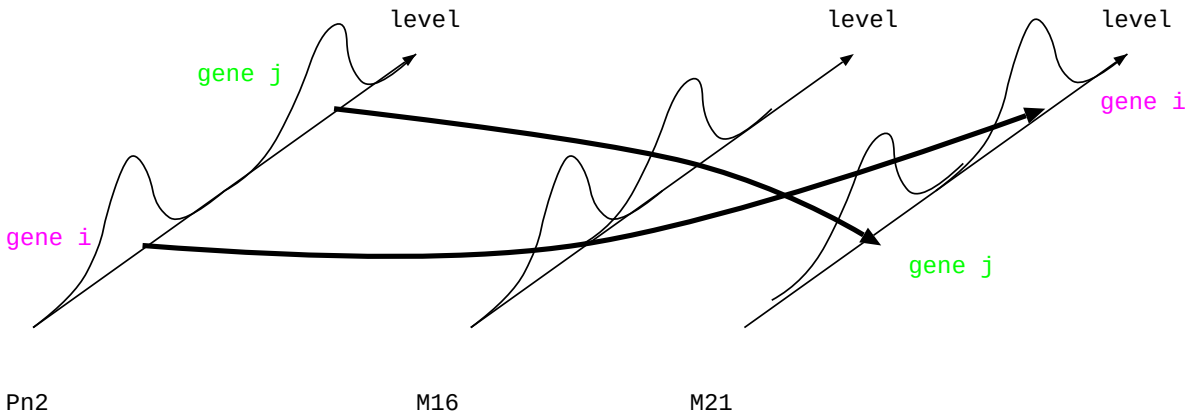

Figure 18: *Gene i is old dominant while gene j is young dominant*

Objective: classify gene trajectories from sequence of microarray experiments over time (t) and population (m)

$$\theta_i(m,t), \quad m = 1,\ldots,M,\ t = 1,\ldots,T$$

Trajectory Estimation

- K classes $C_1, \ldots, C_K$ of genes with priors $\pi_1, \ldots \pi_K$

$$f(\theta_i | i \in C) = \phi_C(\theta_i) = \prod_{m=1}^{M} \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma_C^2(t)}} \exp\left(-\tfrac{1}{2} \left[\theta_i(m,t) - \mu_C(t)\right]^2 / \sigma_C^2(t) \right)$$

- $\mu_C(t), \sigma_C(t)$ piecewise linear to be estimated via EM:

$$\mu_C(t) = \begin{cases} a_{C1}t + b_{C1}, & t = 1, \ldots, \tau_C \\ a_{C2}t + b_{C2}, & t = \tau_C + 1, \ldots, T \end{cases}$$

- Trajectory pdf is Gaussian mixture:

$$f(\theta_i) = \sum_{k=1}^{\#\text{classes}} \phi_k(\theta_i)\pi_k$$

Gene Filtering via Multiobjective Optimization

Gene selection criteria for i-th gene $\xi_1(\theta_i),\ ,\ldots,\ \xi_P(\theta_i)$

Examples of $\xi_p(\theta_i)$:

- Mean change from $t=1$ to $t=T$:

$$\xi_1(\theta_i) = \left|\overline{\theta}_i(*,1) - \overline{\theta}_i(*,T)\right|^2$$

- Standard deviation at $t=1$:

$$\xi_2(\theta_i) = \overline{\left(\theta_i(*,1) - \overline{\theta}_i(*,1)\right)^2}$$

- Standard deviation at $t=T$:

$$\xi_3(\theta_i) = \overline{\left(\theta_i(*,T) - \overline{\theta}_i(*,T)\right)^2}$$

- Mean slope magnitude:

$$\xi_4(\theta_i) = \overline{|\Delta\theta_i(*,*)|}$$

- Mean slope dispersion:

$$\xi_5(\theta_i) = \overline{\left(|\Delta\theta_i(*,*)| - \overline{|\Delta\theta_i(\bullet,\bullet)|}\right)^2}$$

Objective: find genes which maximize or minimize the selection criteria

Aggregated Criteria

Let $\{W_p\}_{p=1}^{P}$ be experimenter's "preference pattern"

$$\sum_{p=1}^{P} W_p = 1, \;\; W_i \geq 0$$

Find optimal gene via:

$$\max_i \sum_{p=1}^{P} W_p \xi_p(\theta_i), \quad \textit{or} \quad \max_i \prod_{p=1}^{P} (\xi_p(\theta_i))^{W_p}$$

Q. What are the set of optimal genes for all preference patterns?

A. These are *non-dominated* genes (Pareto optimal)

Defn: Gene i is dominated if there is a $j \neq i$ s.t.

$$\xi_p(\theta_i) \leq \xi_p(\theta_j), \; p = 1, \ldots, P$$

Example: pairwise comparisons

i-th treatment generates two classes of responses X_i and Y_i:

$\{X_i(m)\}_{m=1}^{n_1}$ and $\{Y_i(m)\}_{m=1}^{n_2}$

- Pooled within-class dispersion

$$\xi_1(X_i, Y_i) = n_1 \overline{\left(X_i(*) - \overline{X_i(*)}\right)^2} + n_2 \overline{\left(Y_i(*) - \overline{Y_i(*)}\right)^2}$$

- Between class distance

$$\xi_2(X_i) = |\overline{X_i(*)} - \overline{Y_i(*)}|^2$$

Objective: Find i which achieves minimum ξ_1 and maximum ξ_2.

Pareto Optimal Fronts

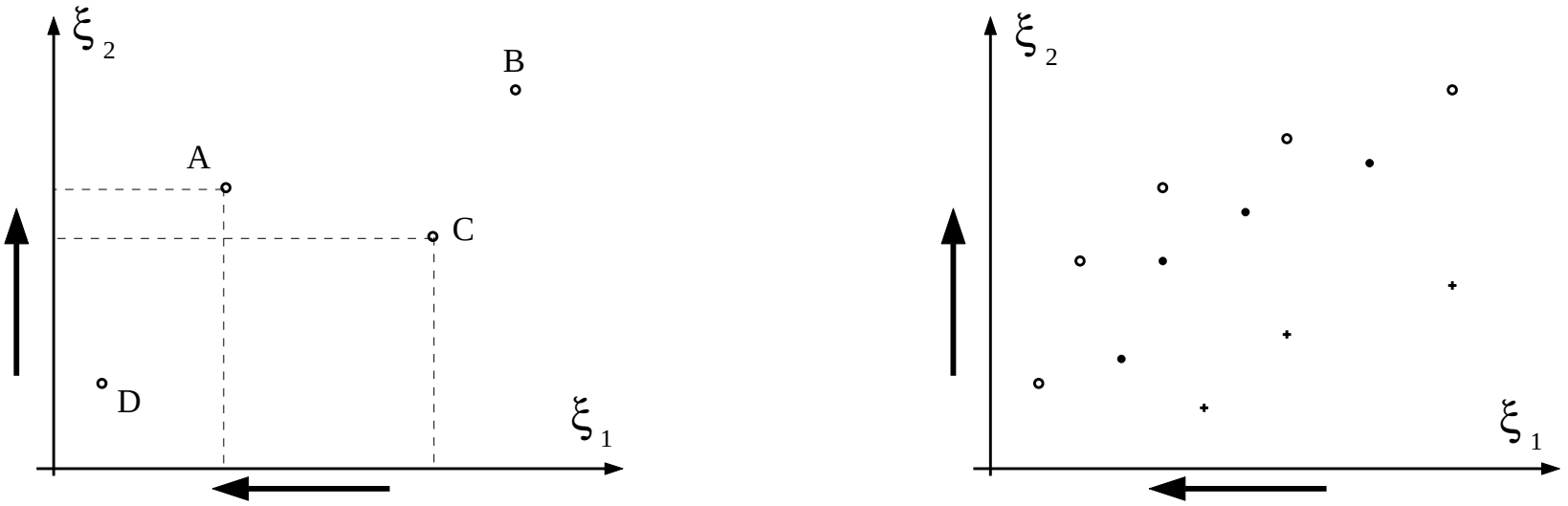

Figure 19: *a). Non-dominated property, and b). Pareto optimal fronts, in dual criteria plane.*

Pareto Gene Filtering vs. Paired T-test

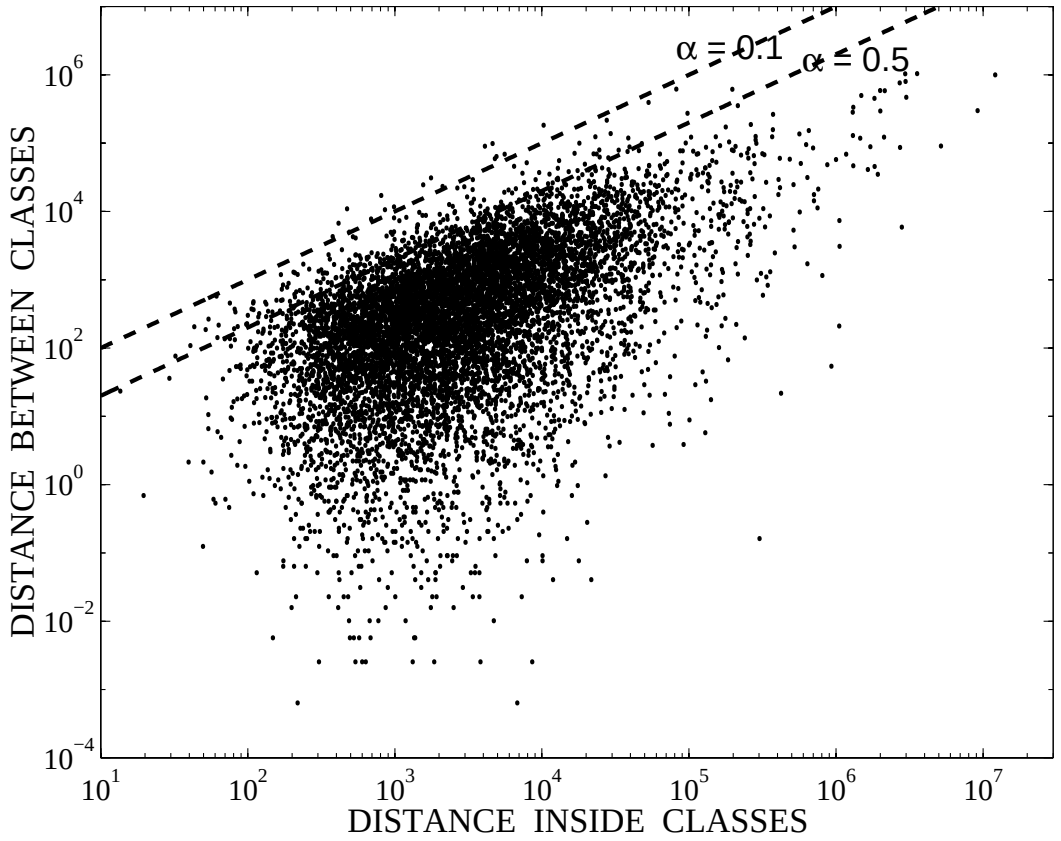

Figure 20: ξ_1 = *mean change vs* ξ_2 = *pooled standard deviation for 8826 mouse retina genes. Superimposed are T-test boundaries*

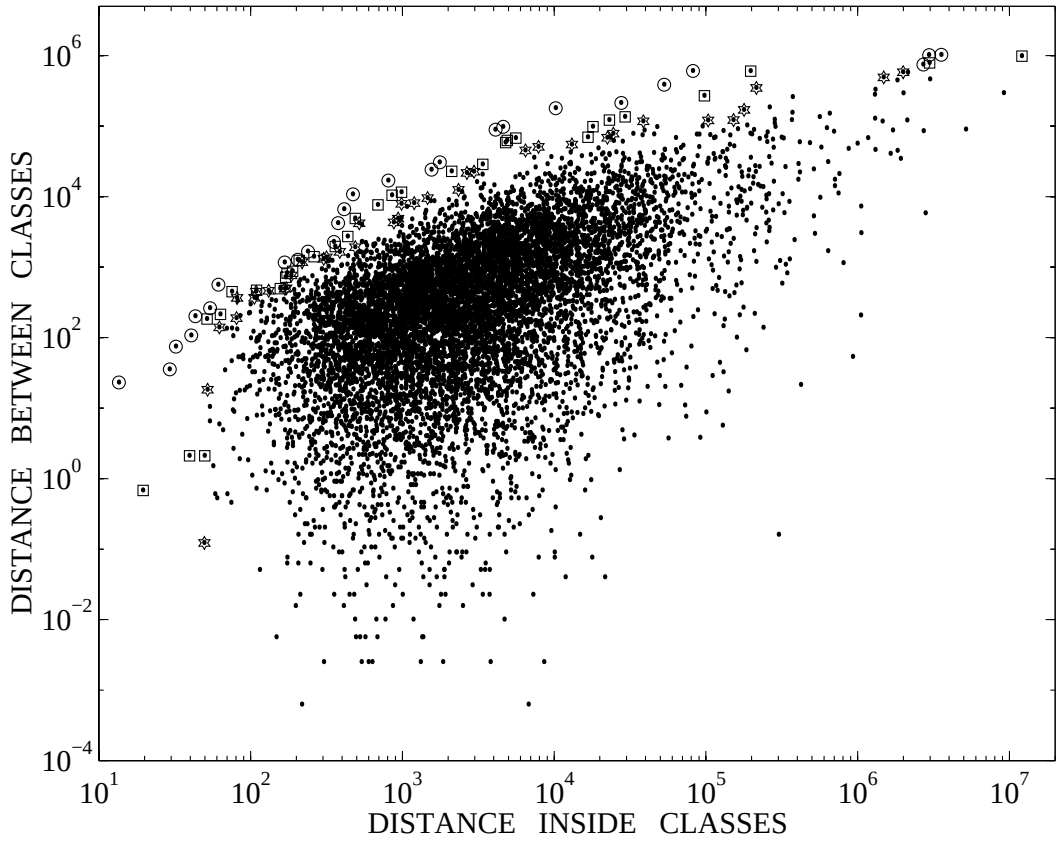

Figure 21: *First (circle) second (square) and third (hexagon) Pareto optimal fronts.*

Application: Development and Aging in Mouse Retina

Mouse Retina Experiment:

- Retinas of 24 transgenic mice sampled and hybridized
- 6 time points: Pn2, Pn10, M2, M6, M16, M21
- 4 mice per time sample
- Affymetrix GeneChip layout with 12422 poly-nucleotides
- Affymetrix attribute analyzed: “AvgDiff”
- Used Affymetrix filter to eliminate all genes labeled “A”

Objective: Find interesting gene trajectories within the set of remaining 8826 genes

Some Gene Trajectories

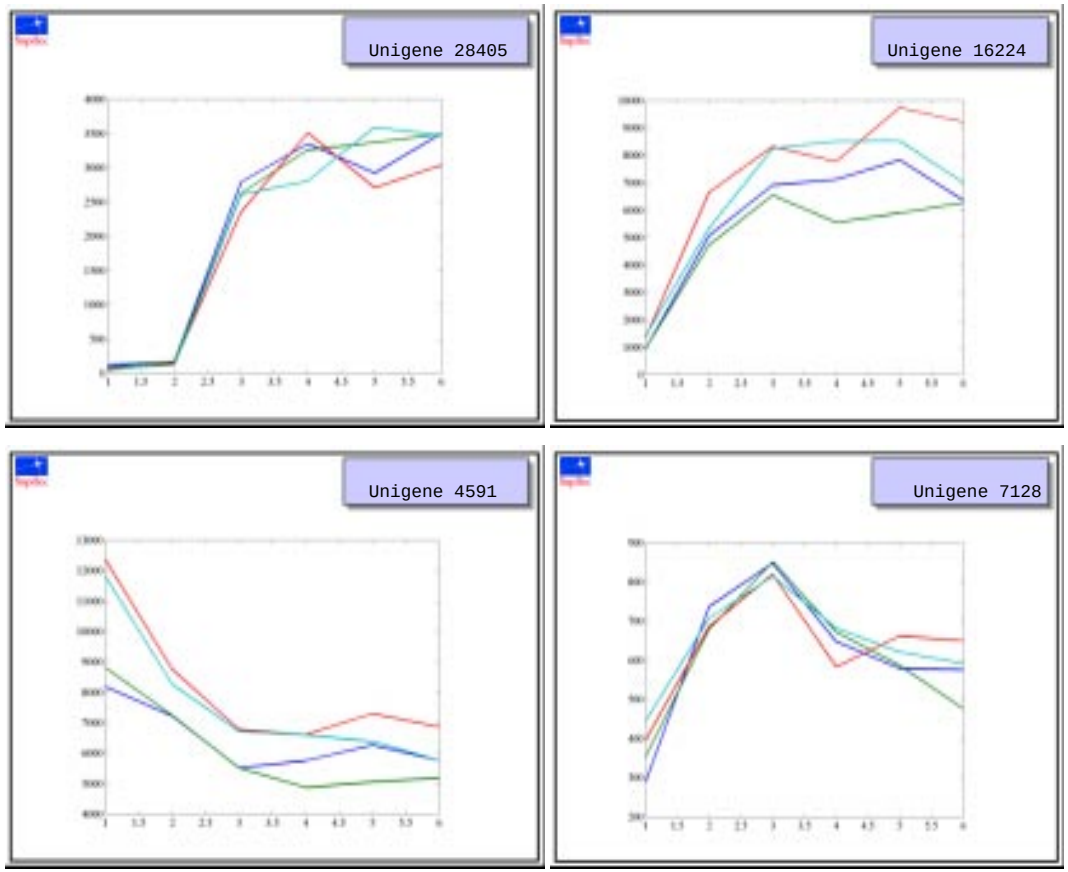

Figure 22: *Trajectories.*

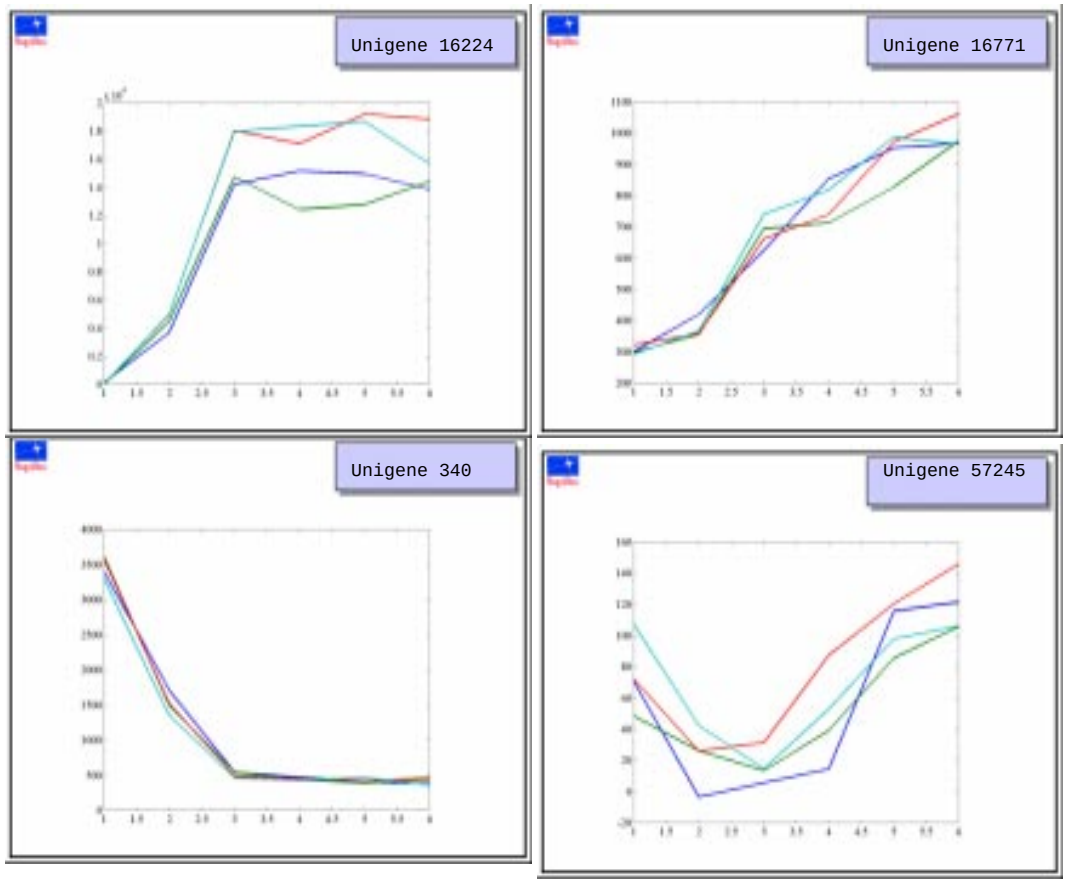

Figure 23: *Trajectories.*

Pairs of Trajectories for Replicated Segments

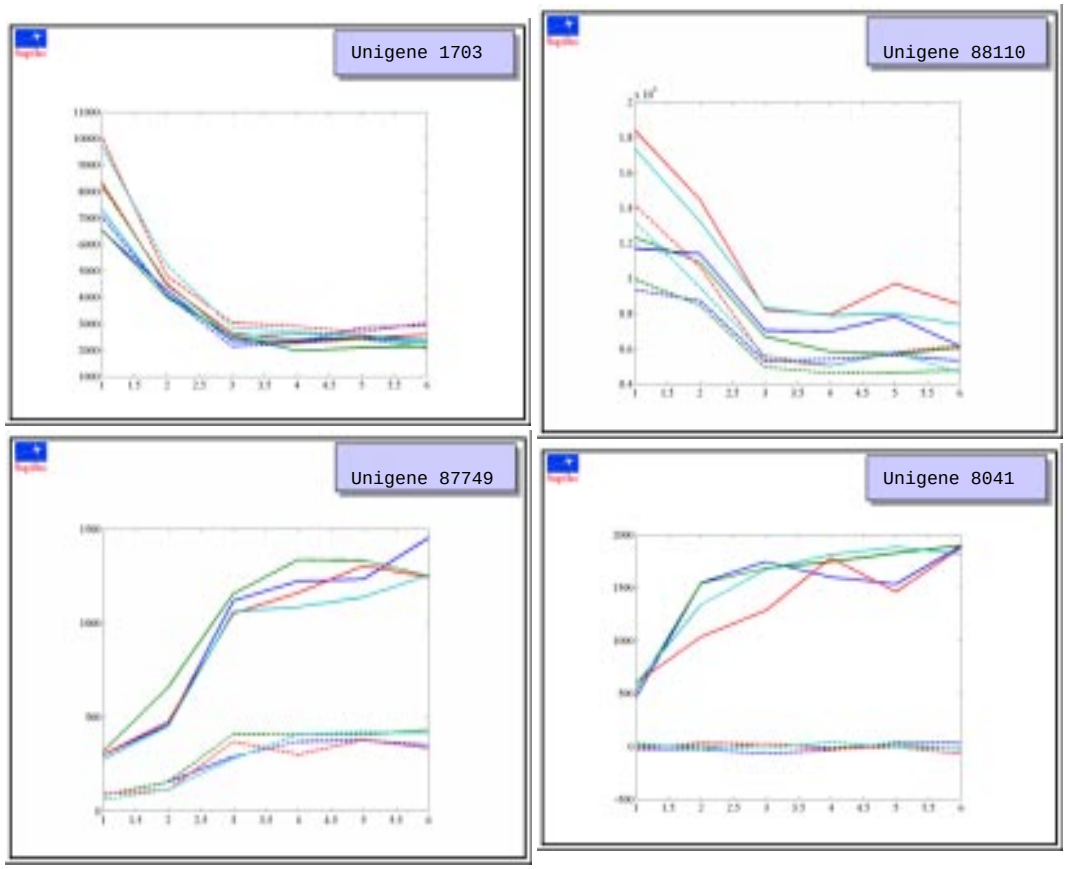

Figure 24: *Pairs of trajectories for replicated gene polynucleotide sequence.*

Non-Parametric Pareto Filter

Define *trend vector*: $\psi_i = [b_1, \ldots, b_6]$, $b_i \in \{0, 1\}$

- Old dominant filtering criteria:
  - high mean slope from $t = Pn1$ to $t = M21$

$$\xi_1(\psi_i) = \overline{b_i(*,*)}$$

  - high consistency over $6^4 = 4096$ possible combinations of trajectories

$$\xi_2(\psi_i) = \frac{\text{\# trajectories having } \psi_i = [1, \ldots, 1]}{4096}$$

Old Dominant Pareto Fronts

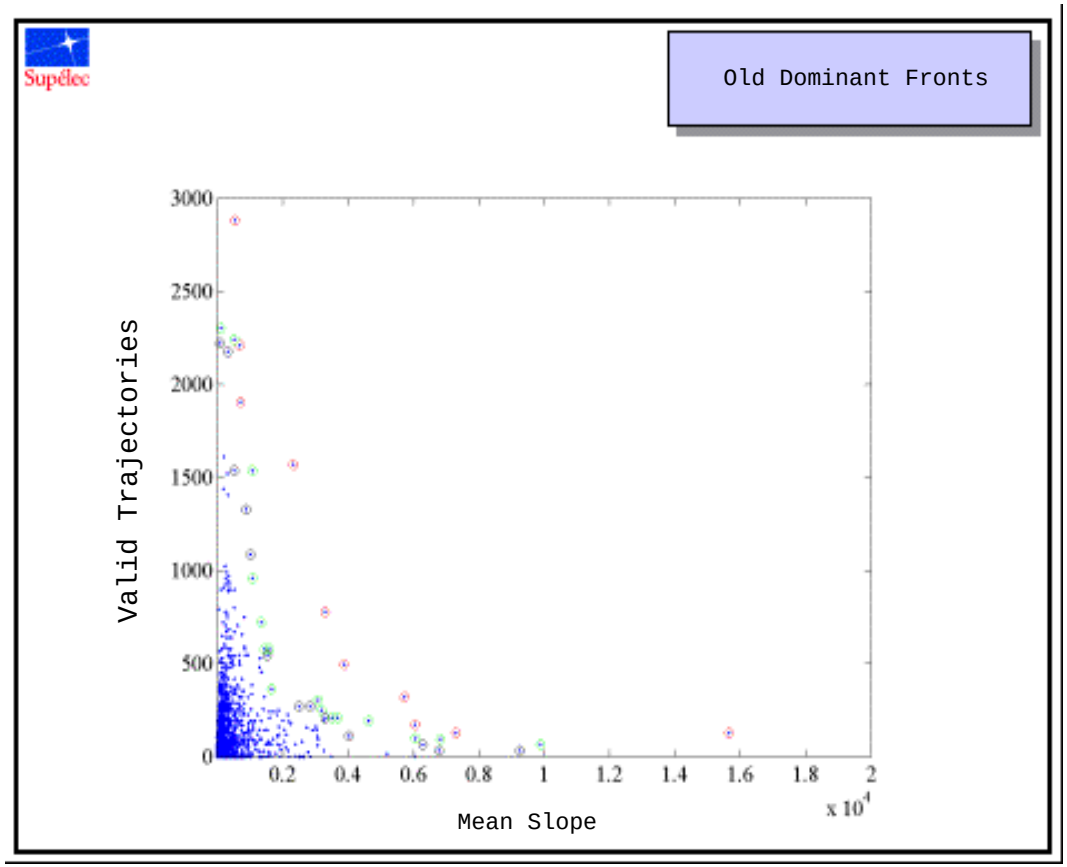

Figure 25: *Pareto fronts for old dominant genes.*

Old Dominant Genes in First Pareto Front

Unigene #	Affymetrix description
1186	Mouse Carbonic Anhydrase II cDNA
4263	Cystatin 3
16224	Guanylate cyclase activator 1a (retina)
16763	Mouse mRNA for aldolase A
16771	Mus musculus H-2K
18625	Aquaporin 1
28405	Mus musculus cDNA 3'end
42102	Mus musculus tubby like protein 1 mRNA
69061	Guanine binding protein α transducing 1
86632	Mus musculus 5'end cDNA

Table 1: *First Pareto Front gene description.*

Resistant Old Dominant Genes in first Three Fronts

- Leave-one-out cross validation

Let ψ_i^{-m} denote one possible set of $T \times (M-1) = 6 \times 3$ samples

Cross-validation Algorithm:

Do $m = 1, \ldots, 4^6$:

Compute $\left(\xi_1(\psi_i^{-m}), \xi_2(\psi_i^{-m})\right)$

Find Genes in First 3 Pareto fronts: G^{-m}

End

Resistant Genes = $\cap_{m=1}^{4^6} G^{-m}$

Unigene #	Affymetrix description
1186	*Mouse Carbonic Anhydrase II cDNA*
1276	Retinal S-antigen
2965	Mouse opsin gene
3918	ATP-binding casette 10
16224	Guanylate cyclase activator 1a (retina)
16763	Mouse mRNA for aldolase A
16771	*Mus musculus H-2K*
39200	CGMP phosphodiesterase gamma
42102	Mus musculus tubby like protein 1 mRNA
69061	Guanine binding protein α transducing 1
86632	*Mus musculus 5'end cDNA*

Table 2: *Resistant genes remaining in first three Pareto fronts*

Young Dominant Filtering Criteria

- low mean slope from $t = Pn1$ to $t = M21$

$$\xi_1(\psi_i) = \overline{b_i(*,*)}$$

- high consistency over $6^4 = 4096$ possible combinations of trajectories

$$\xi_2(\psi_i) = \frac{\#\text{ trajectories having } \psi_i = [0,\ldots,0]}{4096}$$

Young Dominant Pareto Fronts

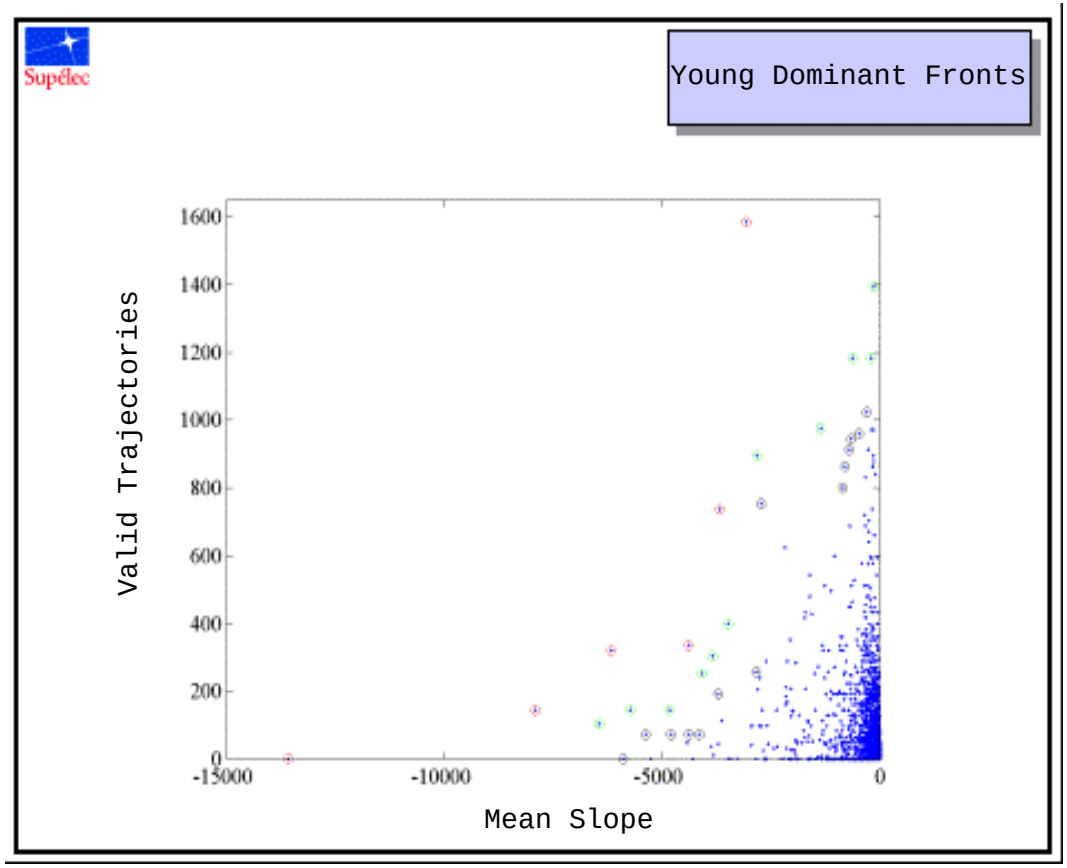

Figure 26: *Pareto fronts for young dominant genes.*

Three-objective Pareto Filtering

Objective Extract "aging genes"

- Strictly increasing filtering criteria:
  - persistent positive trend

$$\xi_1(\psi_i) = \overline{\min_t b_i(*,t)} = \max$$

  - high consistency over $6^4 = 4096$ possible combinations of trajectories

$$\xi_2(\psi_i) == \frac{\#\text{ trajectories having } \psi_i = [1,\ldots,1]}{4096} = \max$$

  - no plateau

$$\xi_3(\theta_i) = \overline{[\theta_i(*,t+1) - 2\theta_i(*,t) + \theta_i(*,t-1)]^2} = \min$$

Pareto Optimal Aging Gene Trajectories

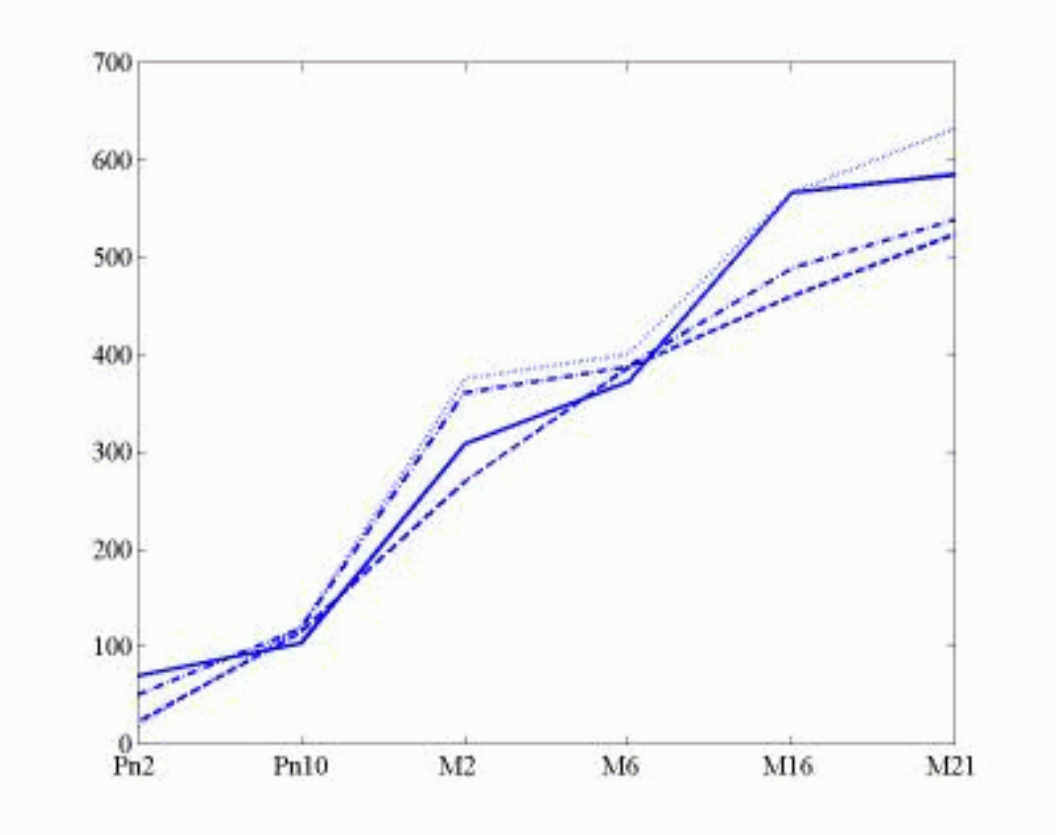

Figure 27: *Mus musculus* $5'$ *end cDNA (Unigene 86632) is sole-survivor resistant aging gene*

Conclusions

1. Filtered Poisson modeling of spotted cDNA microarrays
2. Lower bounds can be used to set image formation parameters
3. Iterative estimation in Poisson model is applicable
4. Pareto filtering performs robust and flexible gene data mining
5. Joint intensity extraction and gene filtering?
6. Evolutionary optimization algorithms for large data sets?
7. Large sample theory of Pareto fronts?