Gene filtering and data mining for gene microarray experiments

A. O. Hero

Dept. EECS*, Dept BME†, Dept. Statistics#

University of Michigan - Ann Arbor

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

Collaborators: G. Fleury, ESE - Paris

S. Yoshida, A. Swaroop UM - Ann Arbor

T. Carter, C. Barlow Salk - San Diego

Outline

1. Gene filtering problem
2. Multi-objective analysis
3. Applications

Kellog Sensory Gene Microarray Node: Objectives

Establish genetic basis for development, aging, and disease in the retina

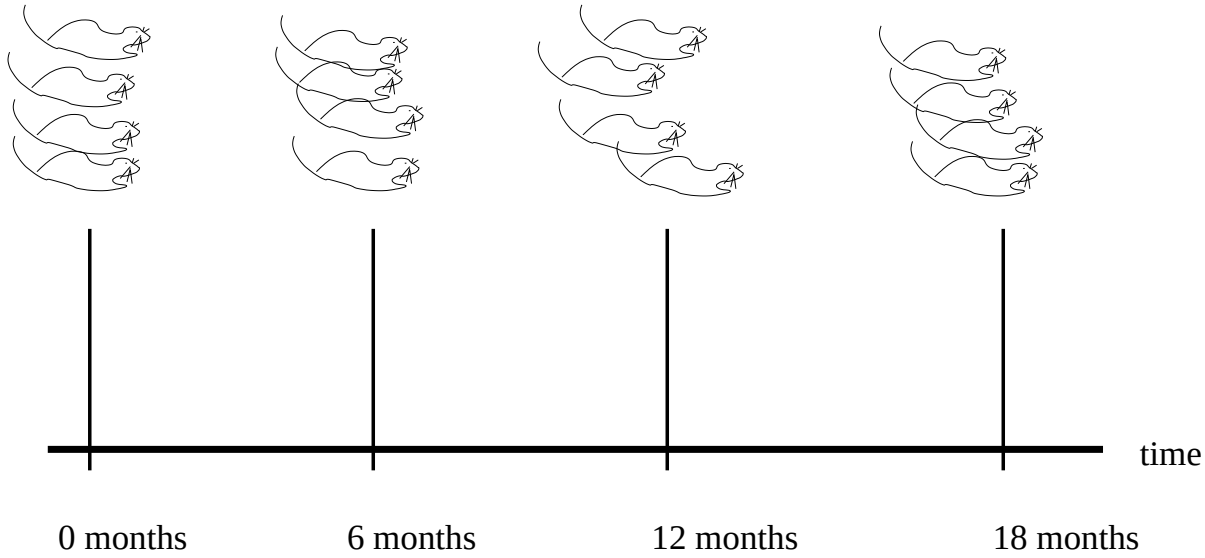

Figure 1: *Sample gene trajectories over time.*

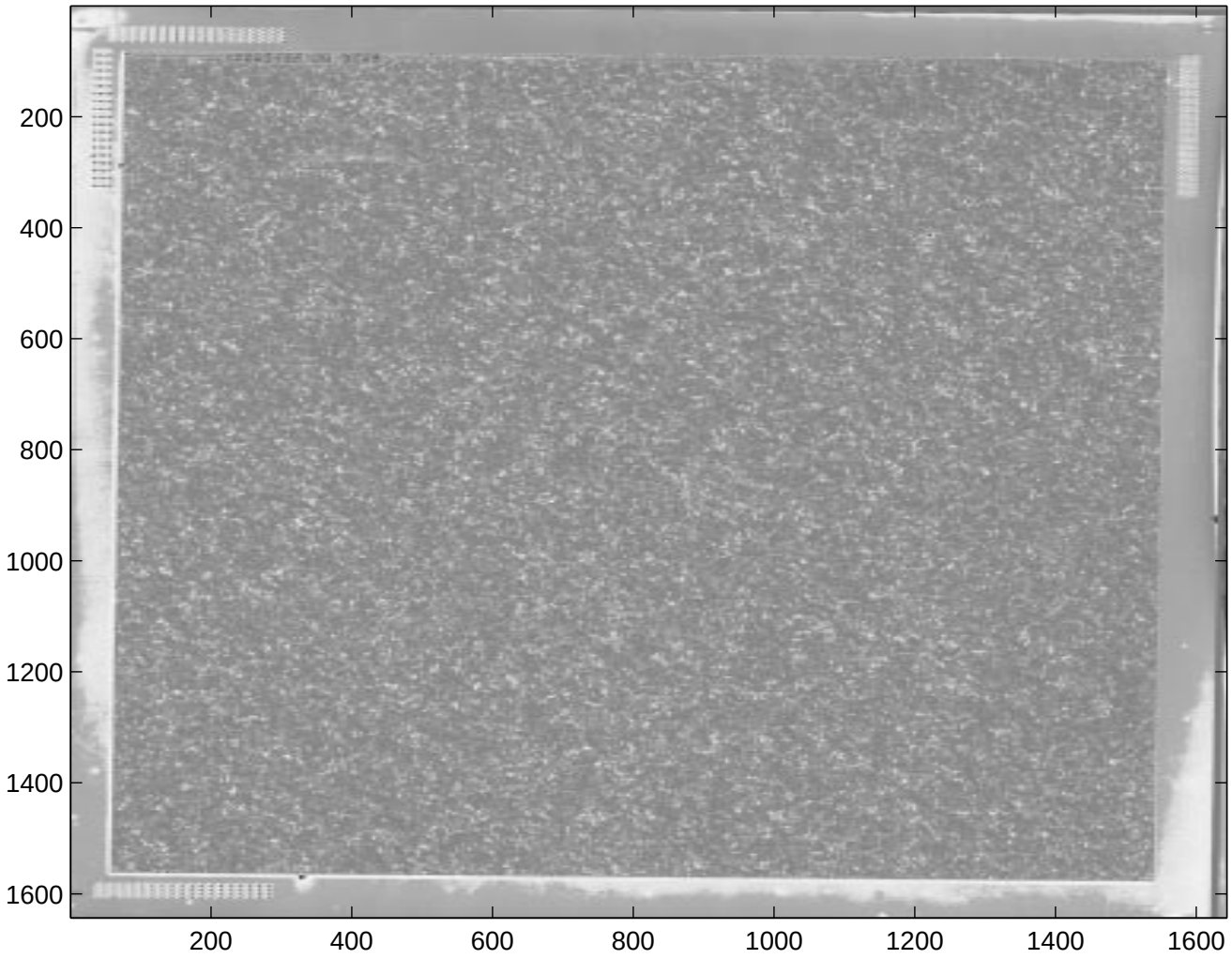

Figure 2: *Affymetrix GeneChip microarray.*

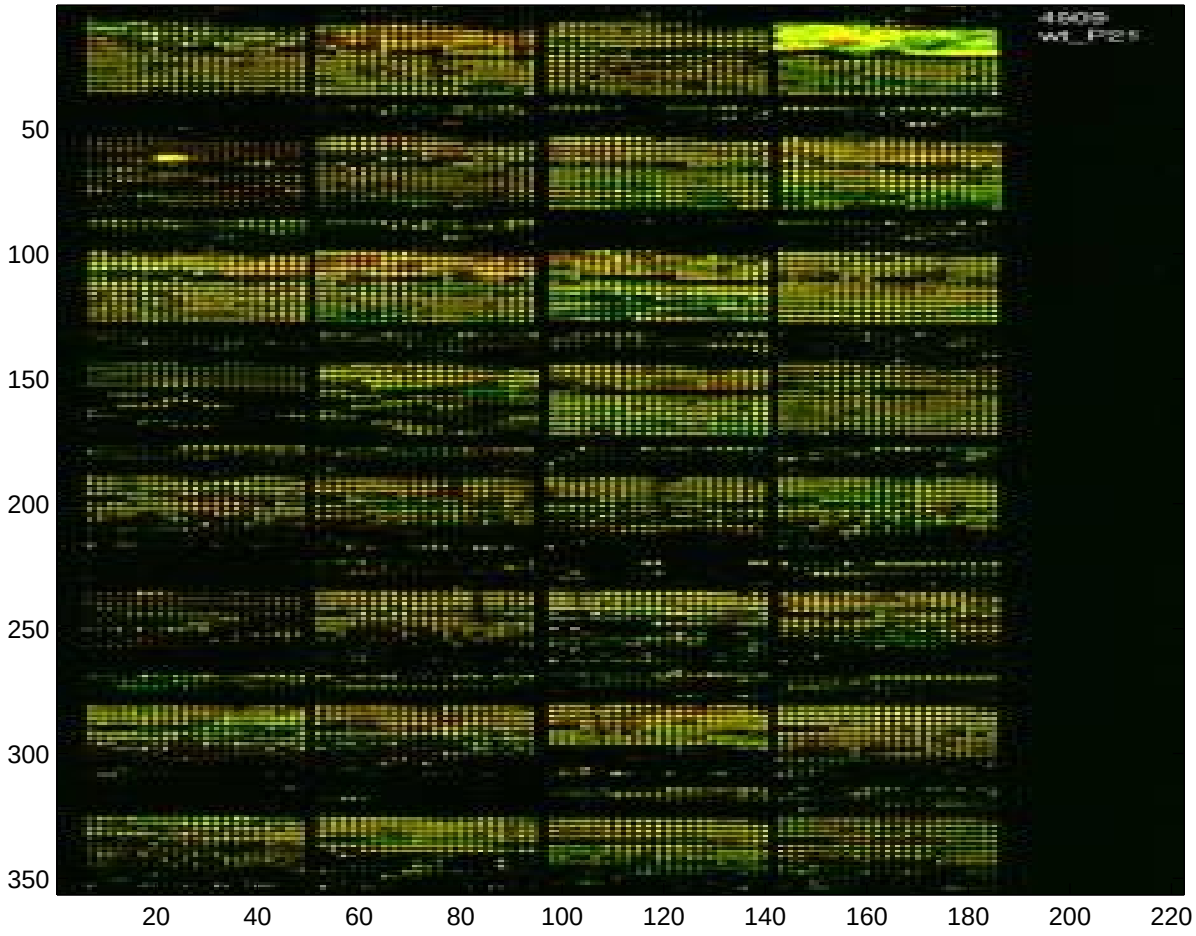

Figure 3: *cDNA spotted array.*

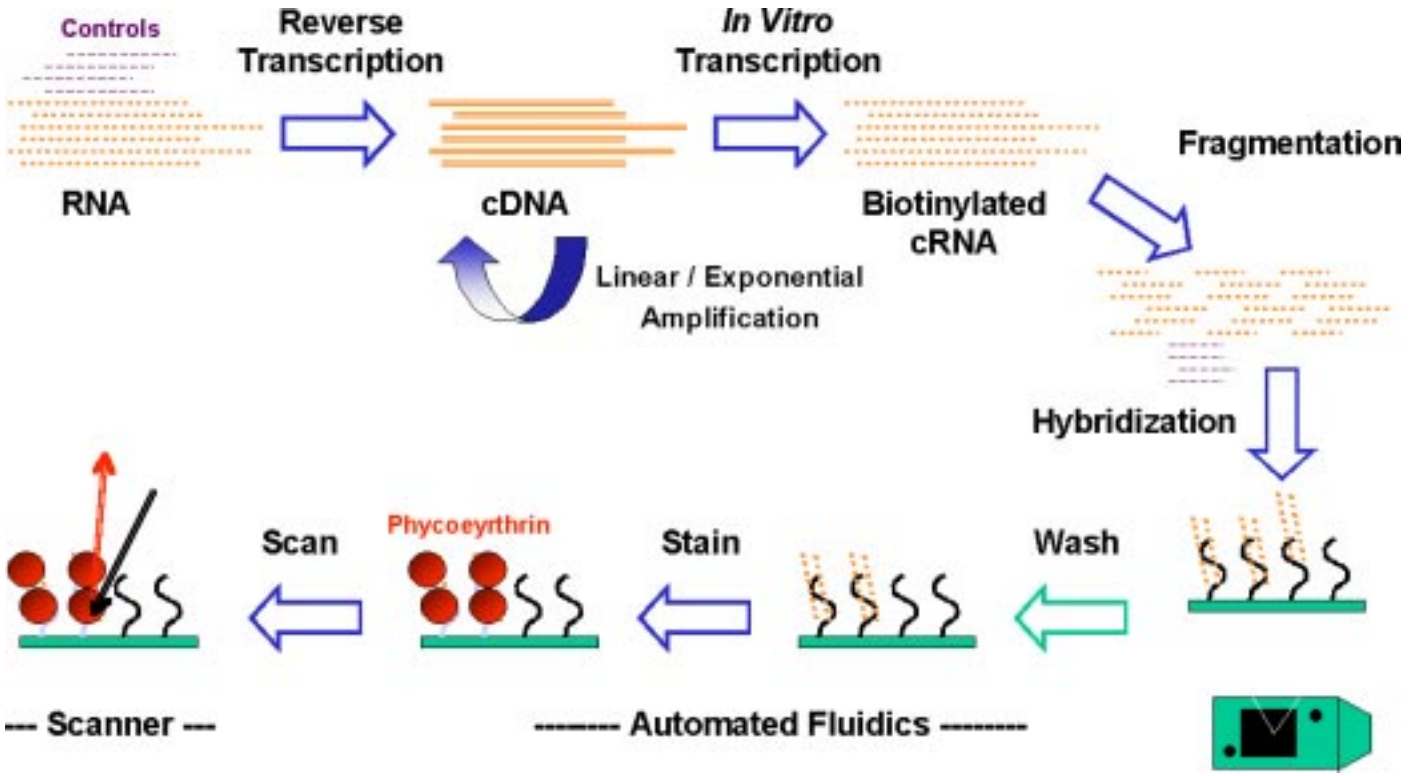

Figure 4: *Oligonucleotide (GeneChip) system (*`pathbox.wustl.edu`*).*

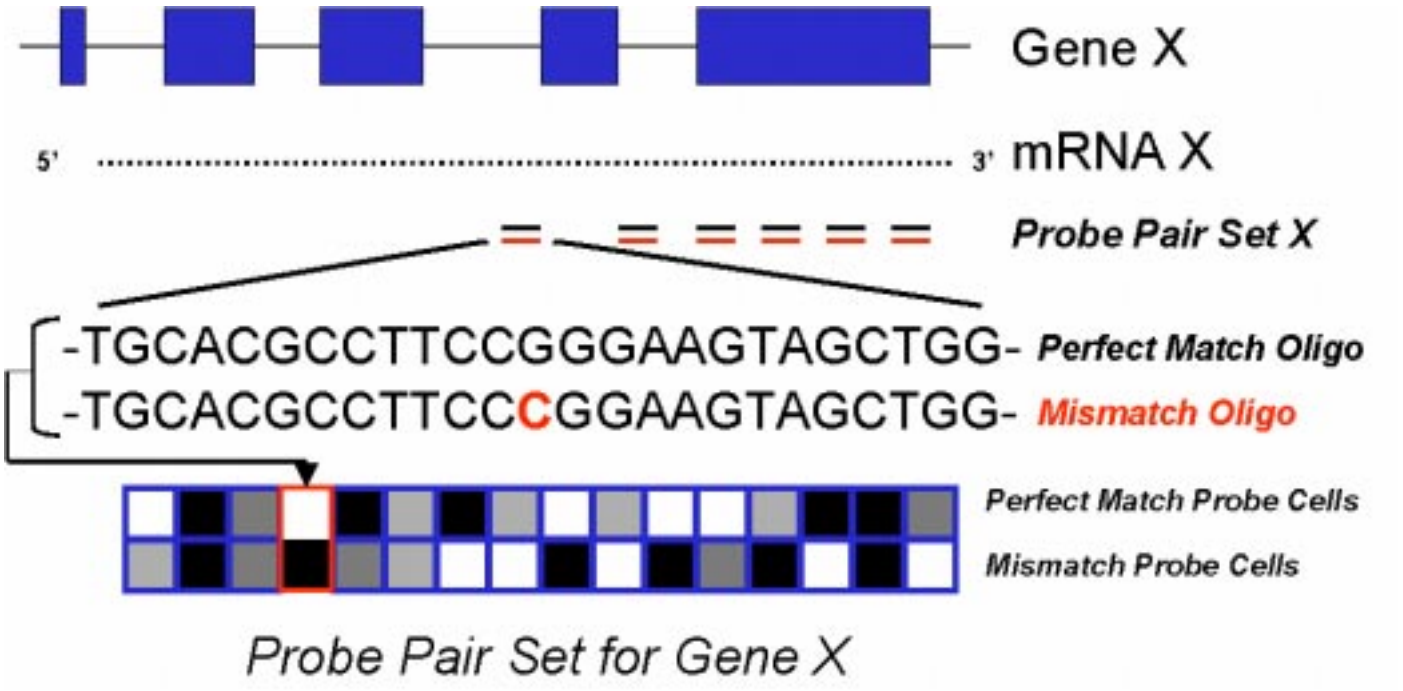

Figure 5: *Oligonucleotide PM/MM layout (*`pathbox.wustl.edu`*).*

(Affymetrix) Output for Each Gene Probe

- **Avg-diff**: avg differences between 20 PM and MM pairs
- **Log-avg** : log of ratios between 20 PM and MM pairs
- **Positive probe pairs**: number of matches to PM
- **Negative probe pairs**: number of matches to MM
- **Absolute Call**: P,A,M

Reference Datasets

1 (2001H) Affy human retinal aging study (Yosida, Swaroop)

- Y group: 8 individuals in age range 16-19 yrs
- O group: 8 individuals in age range 72-80 yrs

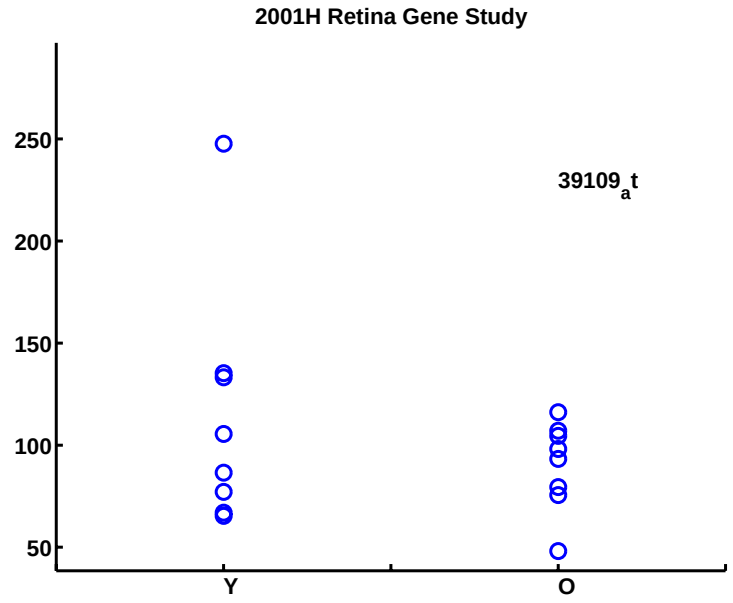

Figure 6: *Responses for a gene in human retinal aging study.*

2 (2001FW) Fred Wright's human fibroblast mixing experiment (`http://thinker.med.ohio-state.edu/projects/fbss/index.html`)

- 18 individuals in 3 groups of 6 subjects

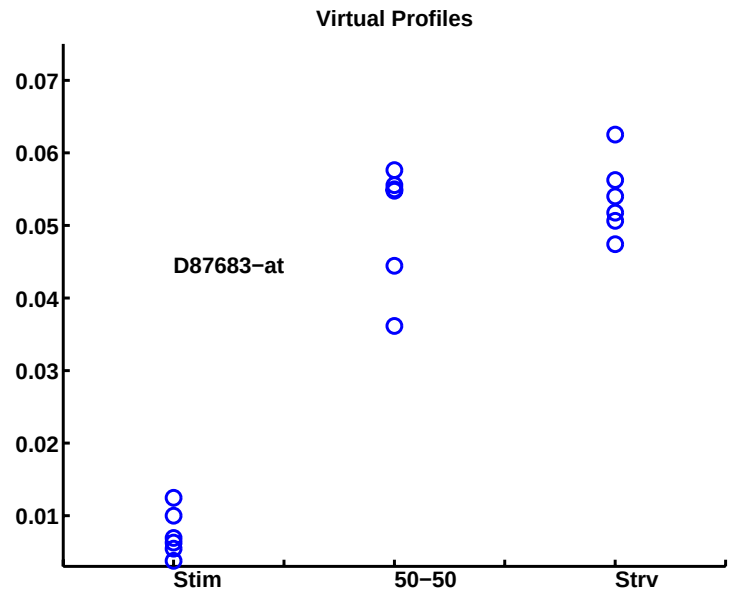

Figure 7: *Responses for a gene in FW human fibroblast mixture study.*

3 (2001M) Affy mouse retinal aging study (Yosida, Barlow, Lockhart, Swaroop)

- 24 mice in 6 groups of 4 subjects

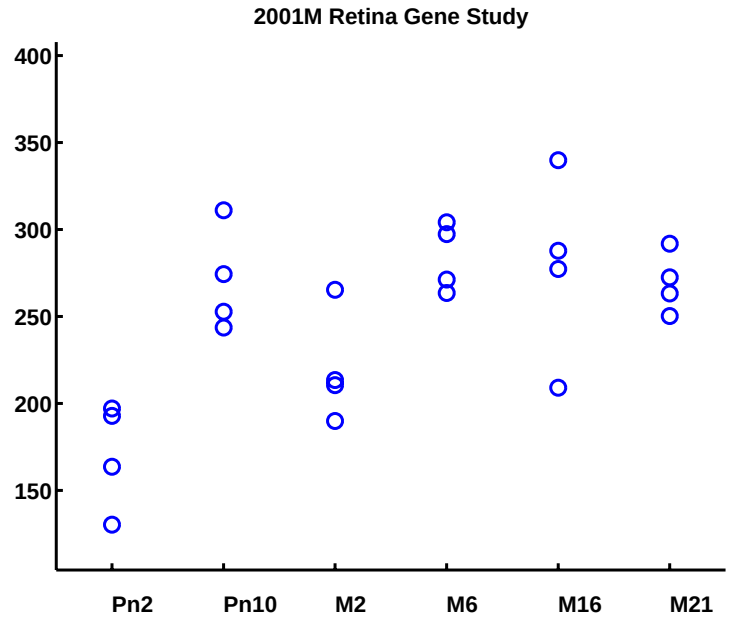

Figure 8: *Responses for a gene in mouse aging study.*

4 (2002M) Affy mouse differential study (Yosida Swaroop)

- 12 knockout mice in 3 groups of 4 subjects
- 12 wildtype mice in 3 groups of 4 subjects

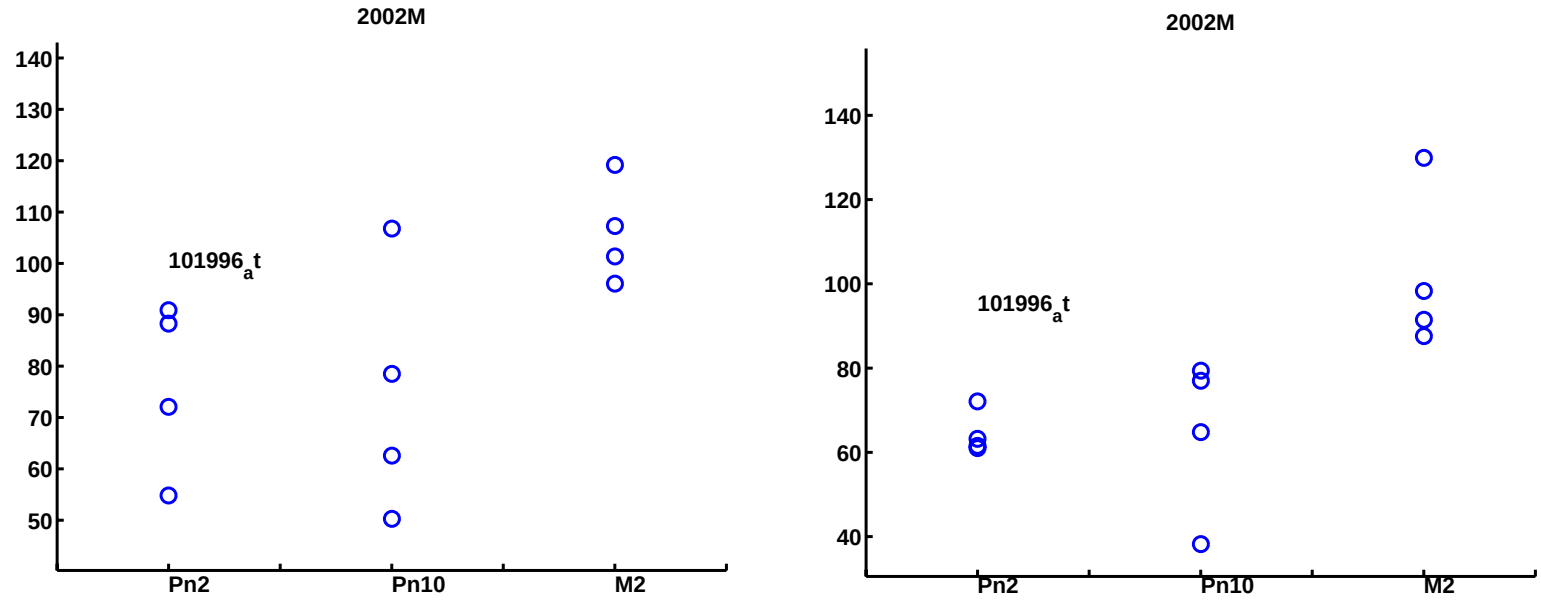

Figure 9: *Differential responses for a gene in mouse k (left) vs w (right) study.*

Pareto Gene Filtering (Fleury&etal:ICASSP02, Hero&etal:GENSIPS02)

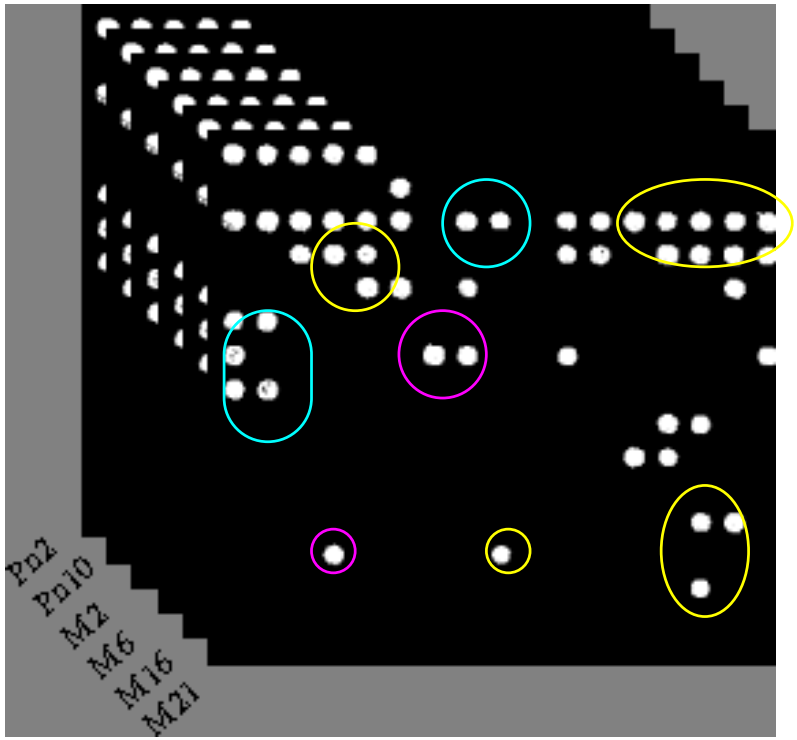

Figure 10: *Clustering on the Data Cube.*

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

Gene Trajectory Classification

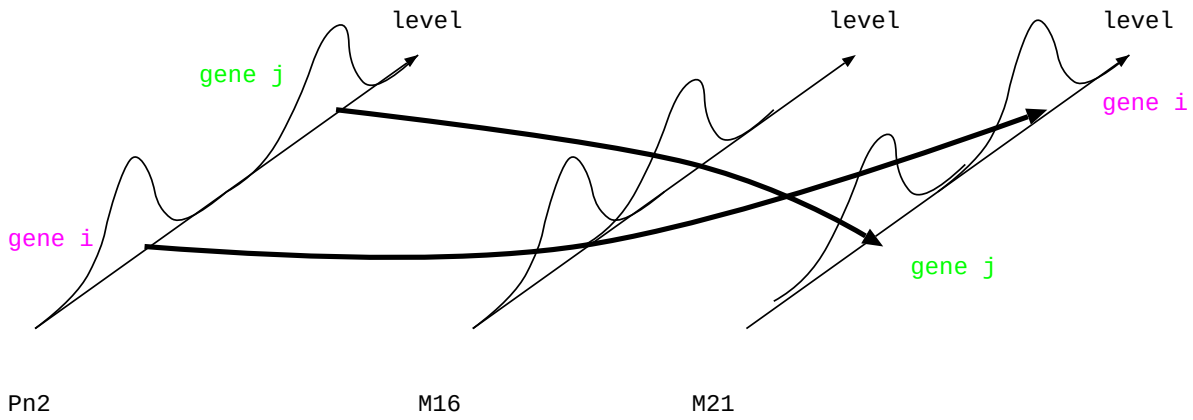

Figure 11: *Gene i is old dominant while gene j is young dominant*

Objective: extract gene trajectories (n) from sequence of repeated (m) microarray experiments over time samples (t)

$$y_{tm}(n), \quad n = 1, \ldots, N,\ t = 1, \ldots, T,\ m = 1, \ldots, M.$$

Clustering and filtering Methods

Principal approaches:

- Hierarchical clustering (kdb trees, CART, gene shaving)
- K-means clustering
- Self organizing (Kohonen) maps
- Vector support machines

Validation approaches:

- Significance analysis of microarrays (SAM)
- Bootstrapping cluster analysis
- Leave-one-out cross-validation
- Replication (additional gene chip experiments, quantitative PCR)

Gene Filtering via Multiobjective Optimization

Gene selection criteria: for n-th gene $\xi_1(Y(n)),\ ,\ldots,\ \xi_P(Y(n))$

Possible $\xi_p(Y(n))$'s for finding uncommon genes

- Squared mean change from $t = 1$ to $t = T$:

$$\xi_1(Y(n)) = |\overline{y}_{T*}(n) - \overline{y}_{1*}(n)|^2$$

- Standard deviation at $t = 1$:

$$\xi_2(Y(n)) = \overline{(y_{1m}(n) - \overline{y}_{1*}(n))^2}$$

- Standard deviation at $t = T$:

$$\xi_3(Y(n)) = \overline{(y_{Tm}(n) - \overline{y}_{T*}(n))^2}$$

Some possible scalar functions:

- t-test statistic (Goss etal 2000): $T(n) = \frac{\xi_1(Y(n))}{\frac{1}{2}\xi_2(Y(n)) + \frac{1}{2}\xi_3(Y(n))}$
- R^2 statistic (Hastie etal 2000): $R^2(n) = \frac{T_n}{1+T_n}$
- H statistic (Sinha etal 1998): $H(n) = \frac{\xi_1(Y(n))}{\sqrt{\xi_2(Y(n))\xi_3(Y(n))}}$

Objective: find genes which maximize or minimize the selection criteria

Aggregated Criteria

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

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

Find optimal gene via:

$$\max_n \sum_{p=1}^{P} W_p \xi_p(Y(n)), \quad \textit{or} \quad \max_n \prod_{p=1}^{P} (\xi_p(Y(n)))^{W_p}$$

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.

$$\xi_p(Y(i)) \leq \xi_p(Y(j)), \; p = 1, \ldots, P$$

Pareto Optimality: increasing criteria

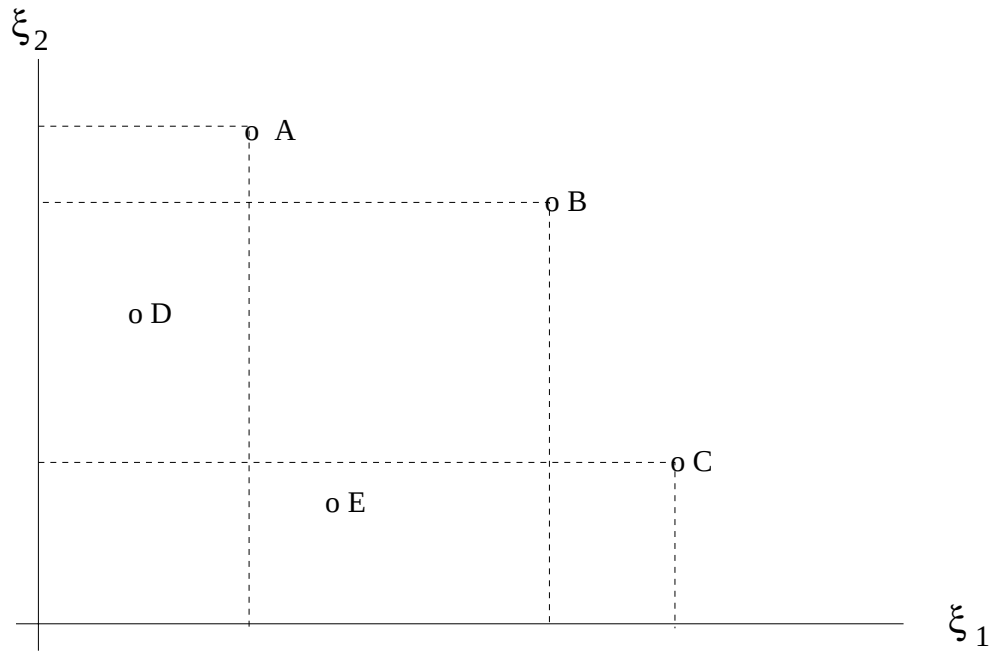

Figure 12: *Foir increasing criteria A, B, C are non-dominated genes and form the (first) Pareto front. A second Pareto front is formed by genes D,E.*

Pareto Optimality: inc/dec criteria

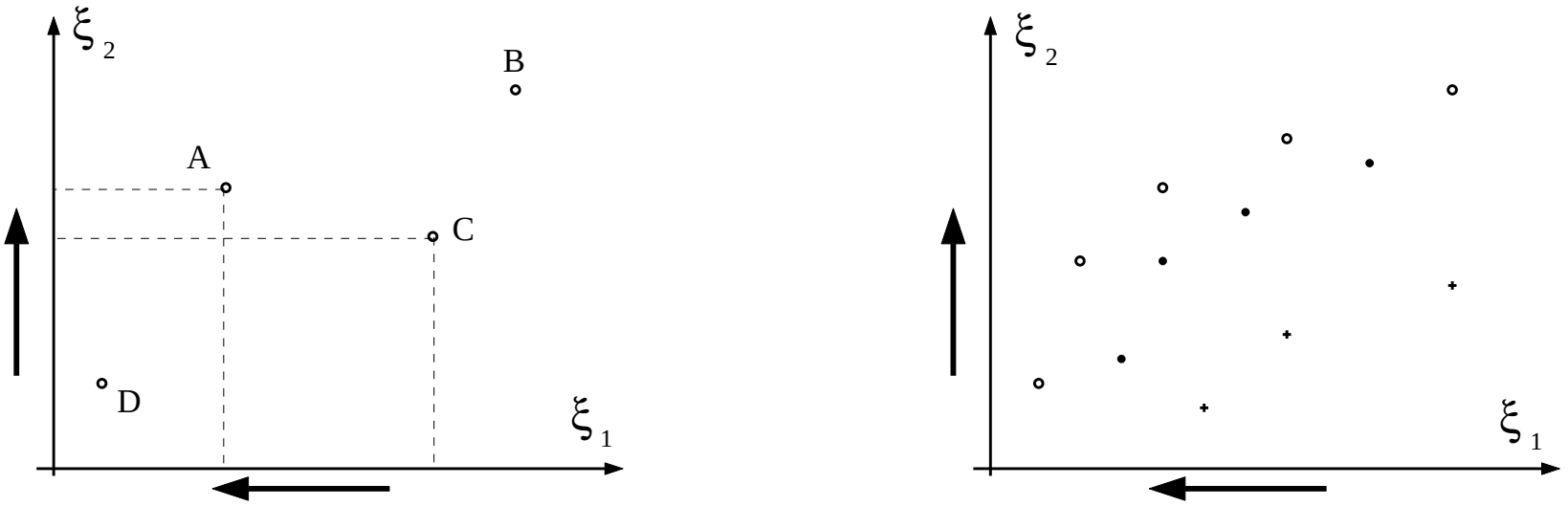

Figure 13: *a). Non-dominated property, and b). Pareto optimal fronts, in dual criteria plane.*

Pareto Gene Filtering vs. Paired T-test

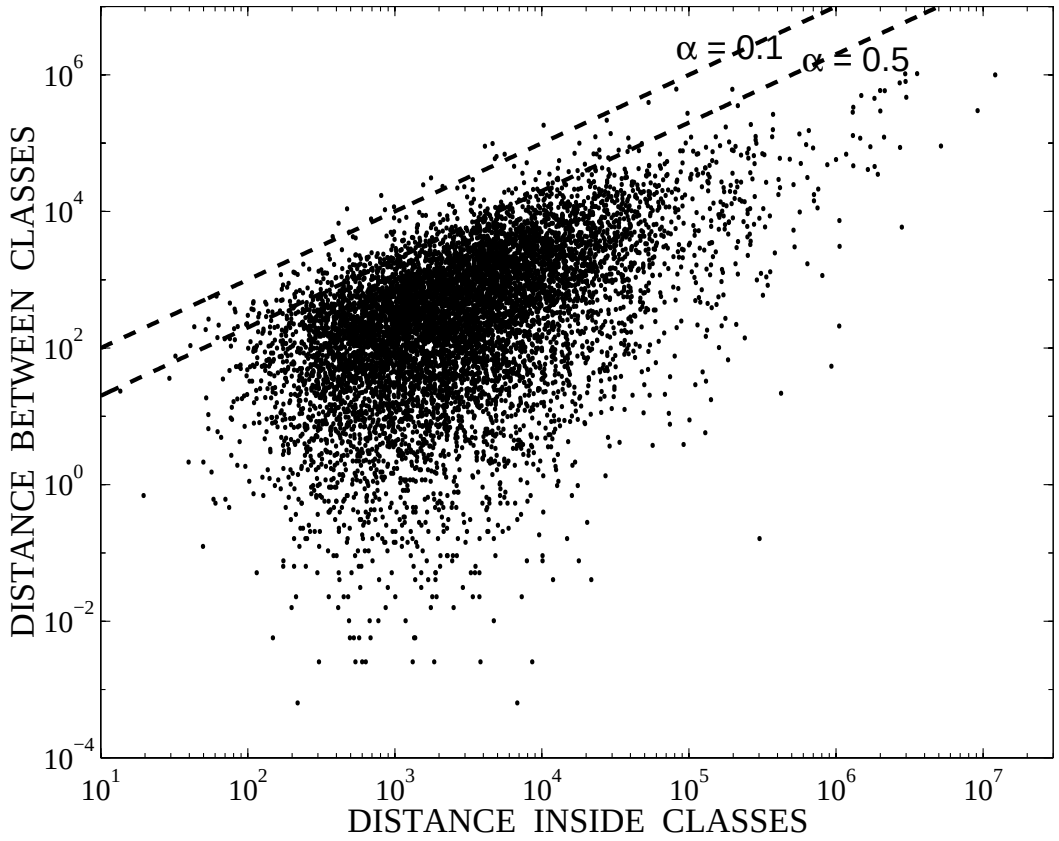

Figure 14: ξ_1 = *mean change vs* ξ_2 = *pooled standard deviation for 8826 human retina genes (2001H). Superimposed are T-test boundaries*

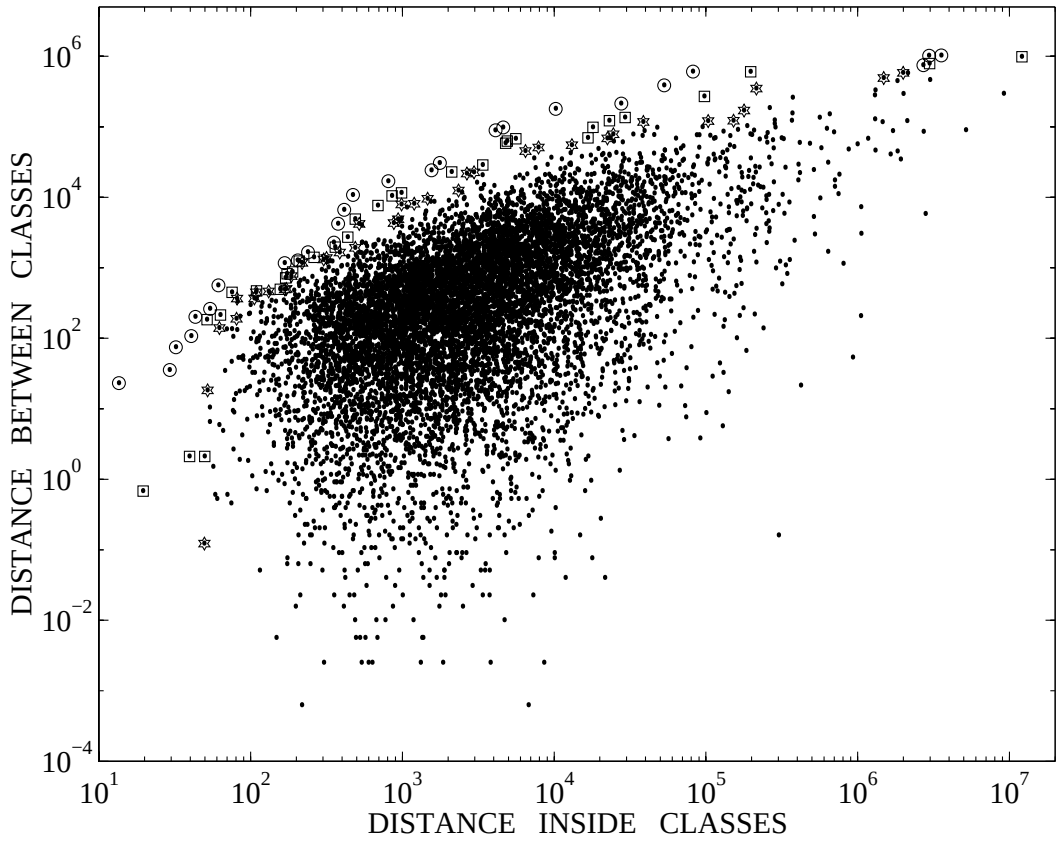

Figure 15: *First (circle) second (square) and third (hexagon) Pareto optimal fronts on (2001H) data.*

Profile Selection Criteria

1. Profile contrasts for trajectory $\{y_{mt}(n)\}_t$

$$\begin{bmatrix} \xi_1(n) \\ \vdots \\ \xi_P(n) \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1T} \\ \vdots & \ddots & \vdots \\ a_{P1} & \cdots & a_{PT} \end{bmatrix} \begin{bmatrix} \bar{y}_{1*}(n) \\ \vdots \\ \bar{y}_{T*}(n) \end{bmatrix}$$

$$A_2 = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}, \; A_2' = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} -1 & 0 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \; A_3' = \begin{bmatrix} -1 & 1 & 0 \\ -1 & -1 & 2 \\ 1 & 1 & 1 \end{bmatrix},$$

2. Profile monotonicity for trajectory $\{y_{mt}(n)\}_t$

$$\xi_2(n) = \prod_{t=2}^{T} I(\overline{y}_{*t}(n) - \overline{y}_{*(t-1)}(n))$$

3. Profile divergence of trajectories $\{w_{mt}(n)\}_t$**,** $\{k_{mt}(n)\}_t$

$$\xi_1(n) = \sum_{t=1}^{T} \overline{k}_{*t}(n) \log \frac{\overline{k}_{*t}(n)}{\overline{w}_{*t}(n)}$$

4. Combinations of above

Accounting for Sampling Errors: Cross-validation

- Leave-one-out cross validation

Let $Y^{-m}(n)$ denote one possible set of $T \times (M-1)$ samples

Cross-validation Algorithm:

`Do` $m = 1, \ldots, M^T$:

`Compute` $\left(\xi_1(Y^{-m}(n)),\, \xi_2(Y^{-m}(n))\right)$

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

`End`

Resistant `Genes =` $\cap_{m=1}^{M^T} G^{-m}$

Accounting for Sampling Errors: Posterior Pareto Analysis

Given prior on mean expression levels $\overline{\xi}_p(n) = E[\xi_p(Y(n))]$ find

$$\begin{aligned} & p(i|Y) \\ = \quad & P\left(\cap_{j\neq i}\left\{\underline{\xi}(i) \leq \underline{\xi}(j)\right\}^c |Y\right) \\ = \quad & \int dP(\underline{\xi}(i)|Y) \prod_{j\neq i} P\left(\left\{\underline{\xi}(i) \leq \underline{\xi}(j)\right\}^c |Y, \underline{\xi}(i)\right) \end{aligned}$$

Case of two criteria ($P = 2$)

$$\begin{aligned} p(i|Y) \quad = \quad & \int\int du_1 du_2 \, f_{\xi_1(i),\xi_2(i)|Y}(u_1, u_2) \\ & \prod_{j\neq i}\left[F_{\xi_1(j)|Y}(u_1) + F_{\xi_2(j)|Y}(u_2) - F_{\xi_1(j),\xi_2(j)|Y}(u_1, u_2)\right] \end{aligned}$$

PPA under Gaussian distributed criteria

1. Assume conditionally independent Gaussian model

$\varepsilon_{mt}(n) \sim N(0, \sigma_t^2(n))$

$$y_{mt}(n) = \mu_t(n) + \varepsilon_{mt}(n)$$

2. Assume non-informative prior

$$f_{\mu_t(n),\sigma_t^2(n)}(u,s) = \frac{c}{s^{a/2}}, \; u \in \mathbf{R}, \; s \in \mathbf{R}^+$$

then (large M):

$$F_{\mu_t(i)|Y}(u) \approx \left(1 + \frac{(\hat{\mu}_t(i) - u)_+^2}{\hat{\sigma}_t^2(i)}\right)^{-(M-a+2)/2}.$$

Application to Fred Wright's Mixture Study

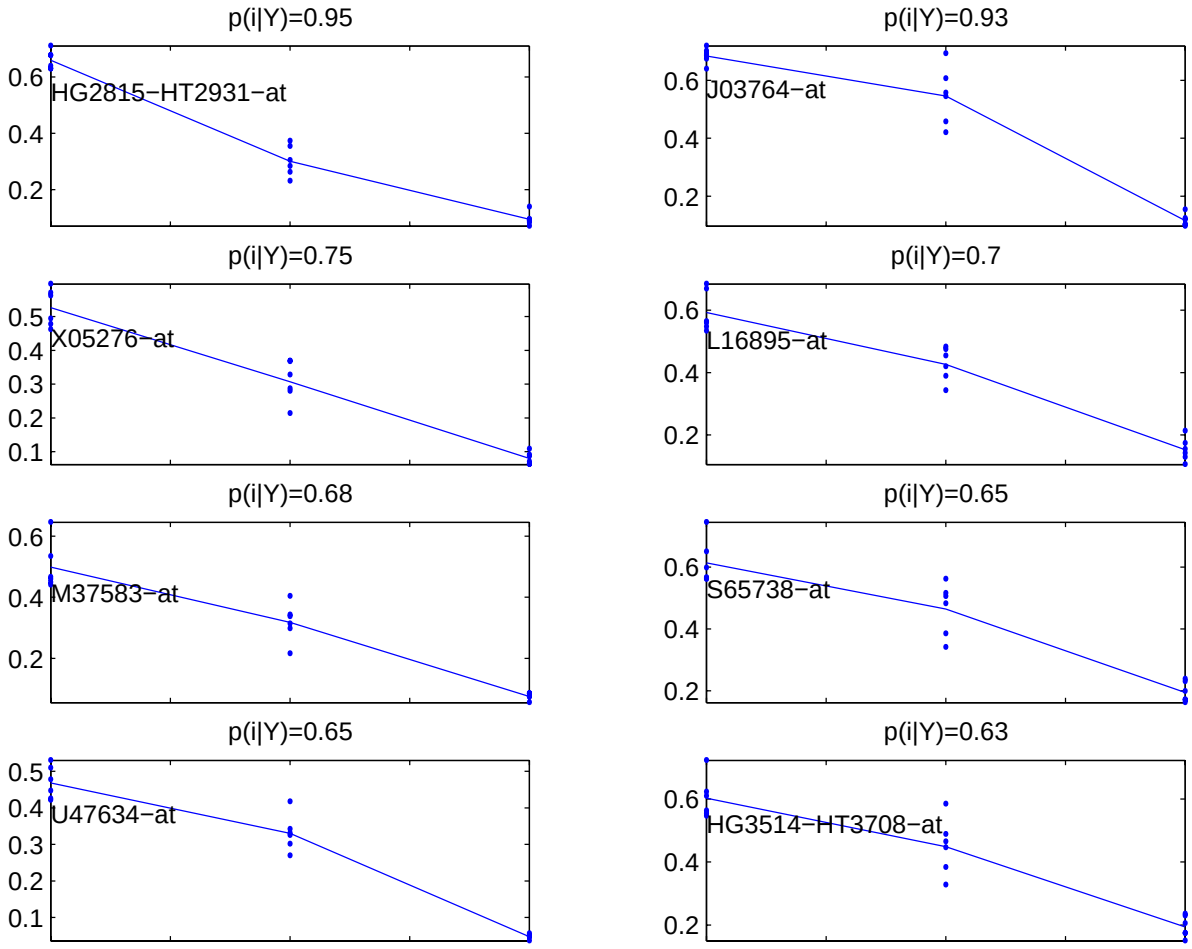

Figure 16: *8 ranked monotone decreasing gene profiles.*

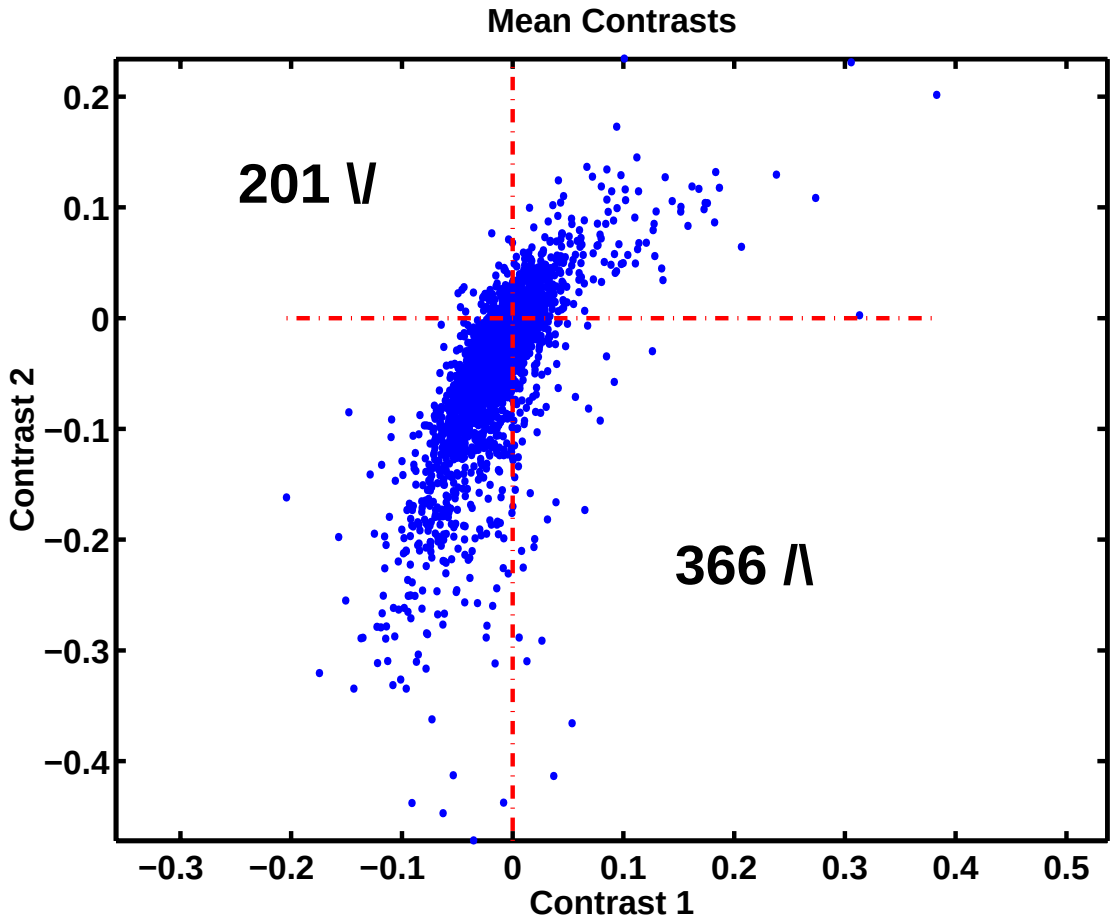

Figure 17: *Multicriterion scattergram for first two rows of* A_3'.

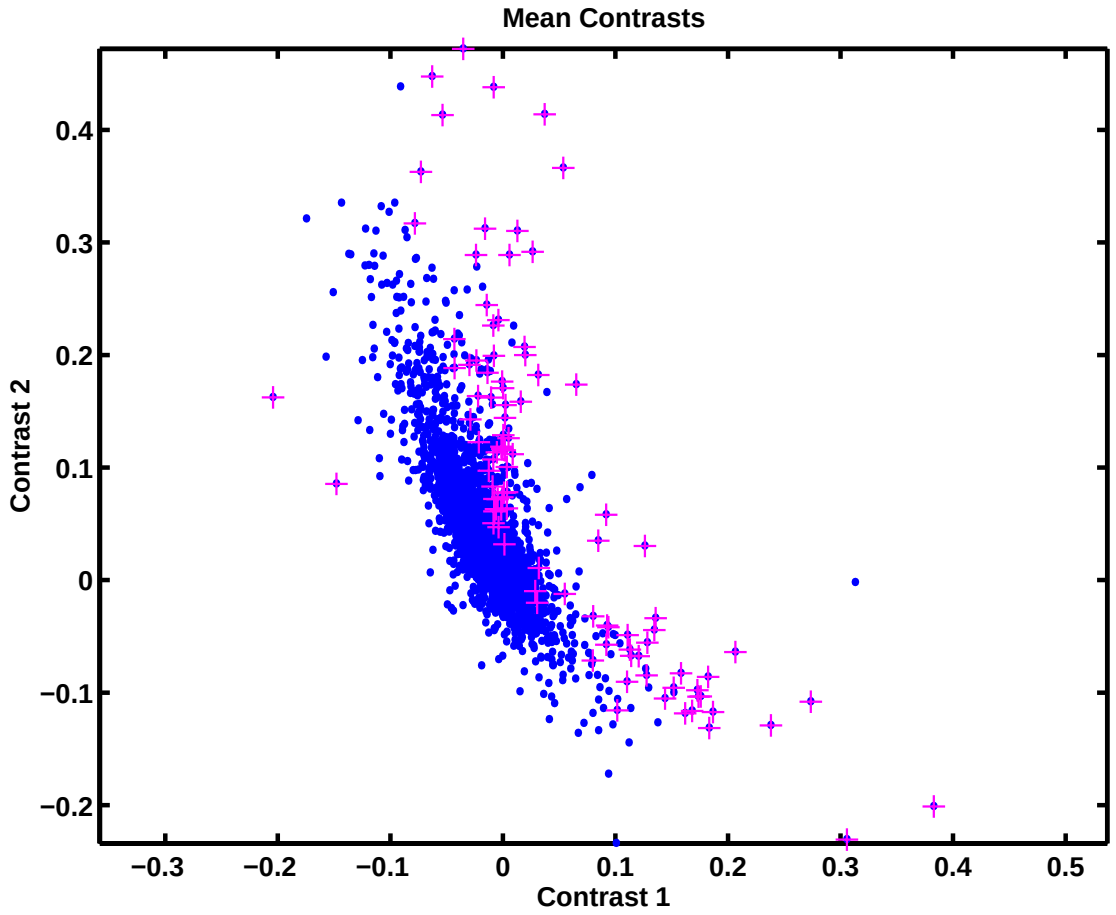

Figure 18: *Multicriterion scattergram for* $A = [-1, 1, 0; -1, -1, 2]$. *98 genes are non-linear profiles (p-value of 0.1).*

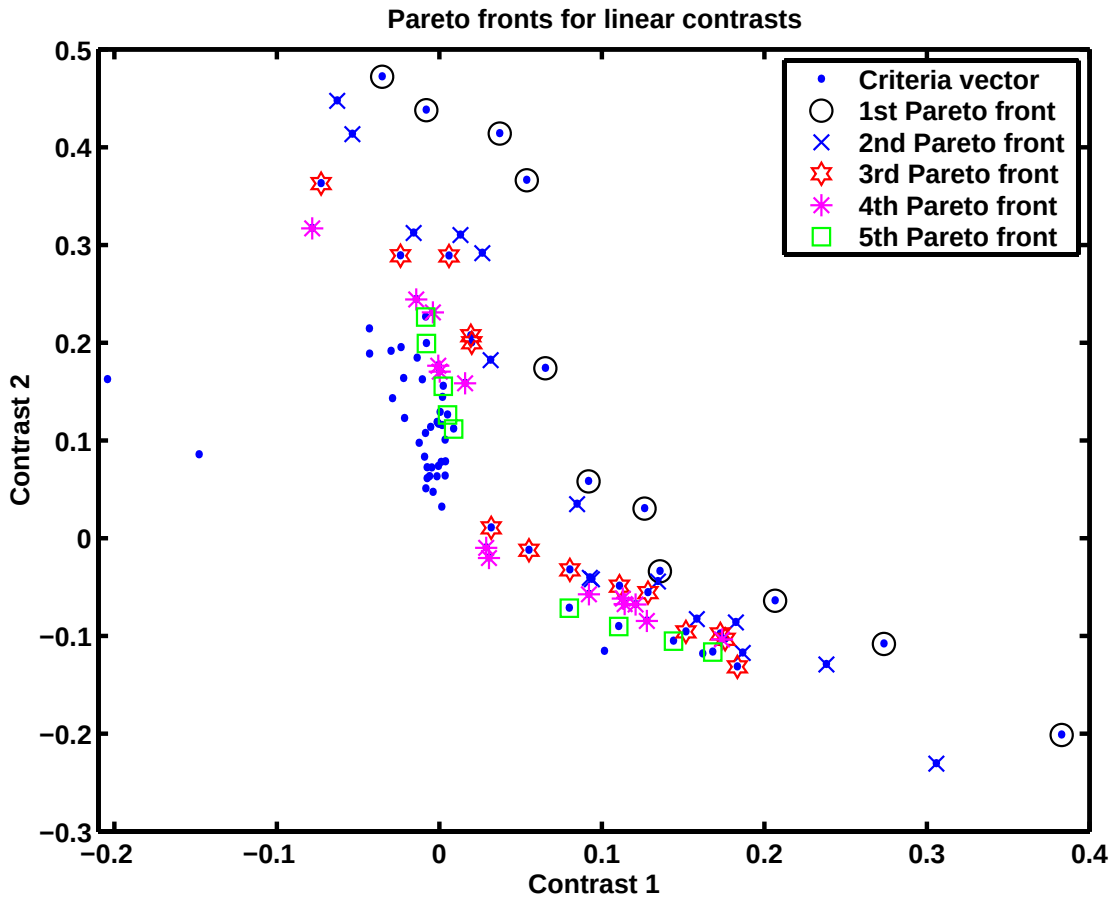

Figure 19: *The first five Pareto fronts for the genes with non-linear profiles shown in Fig. 18.*

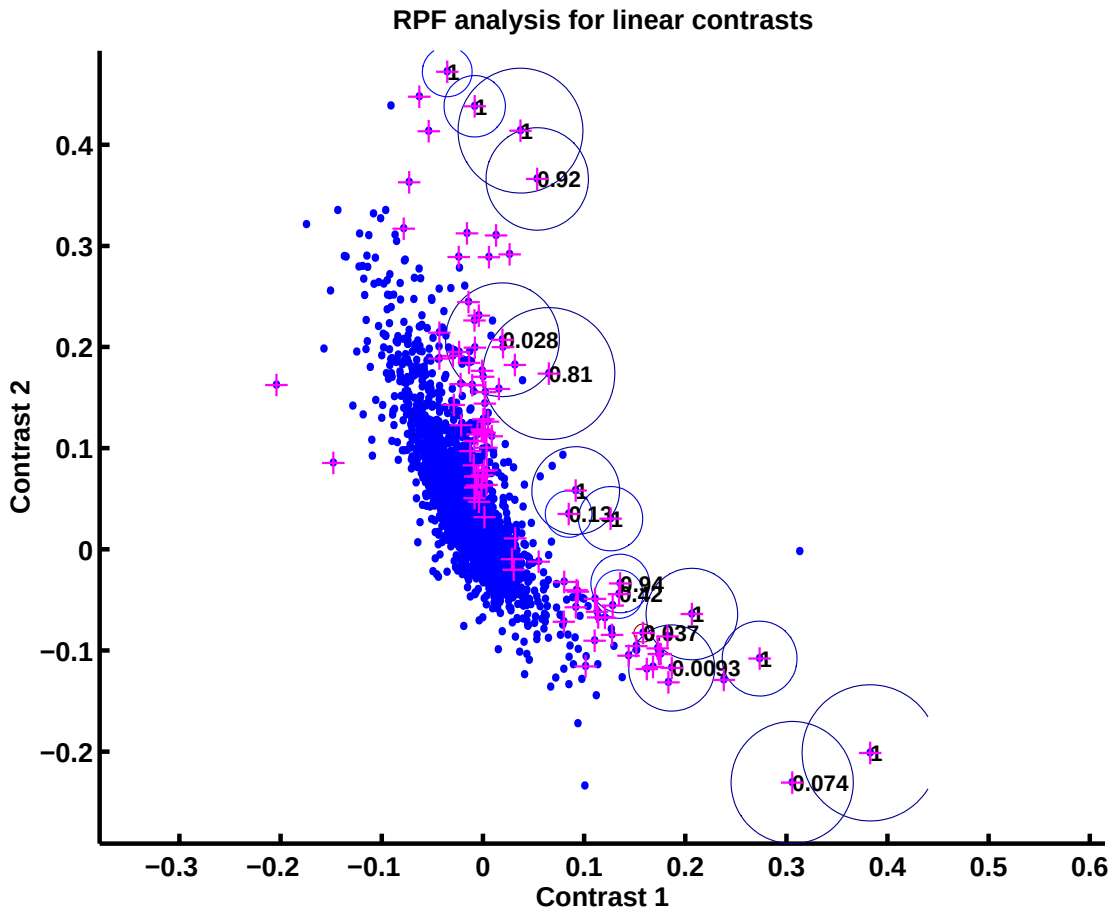

Figure 20: *17 genes in first Pareto front with non-zero probability by cross-validation.*

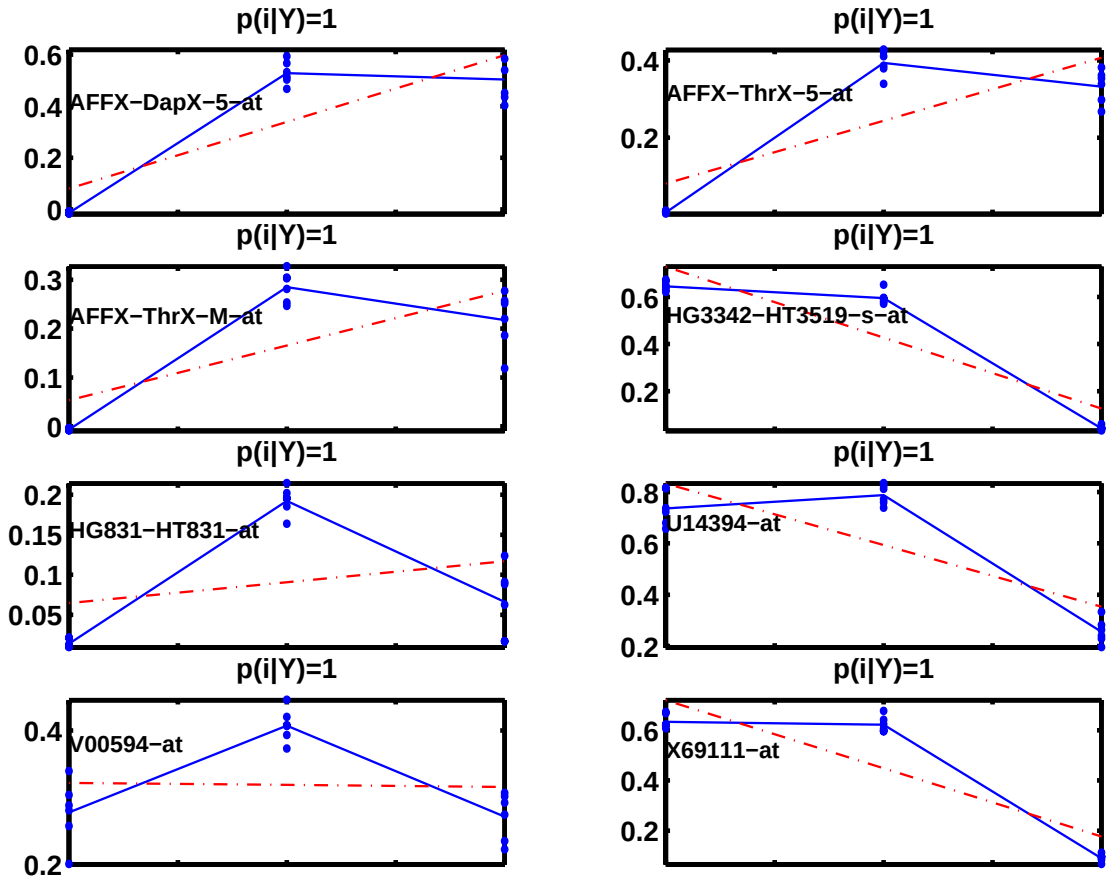

Figure 21: *The 8 top cross-validation ranked gene profiles remaining on the first Pareto front.*

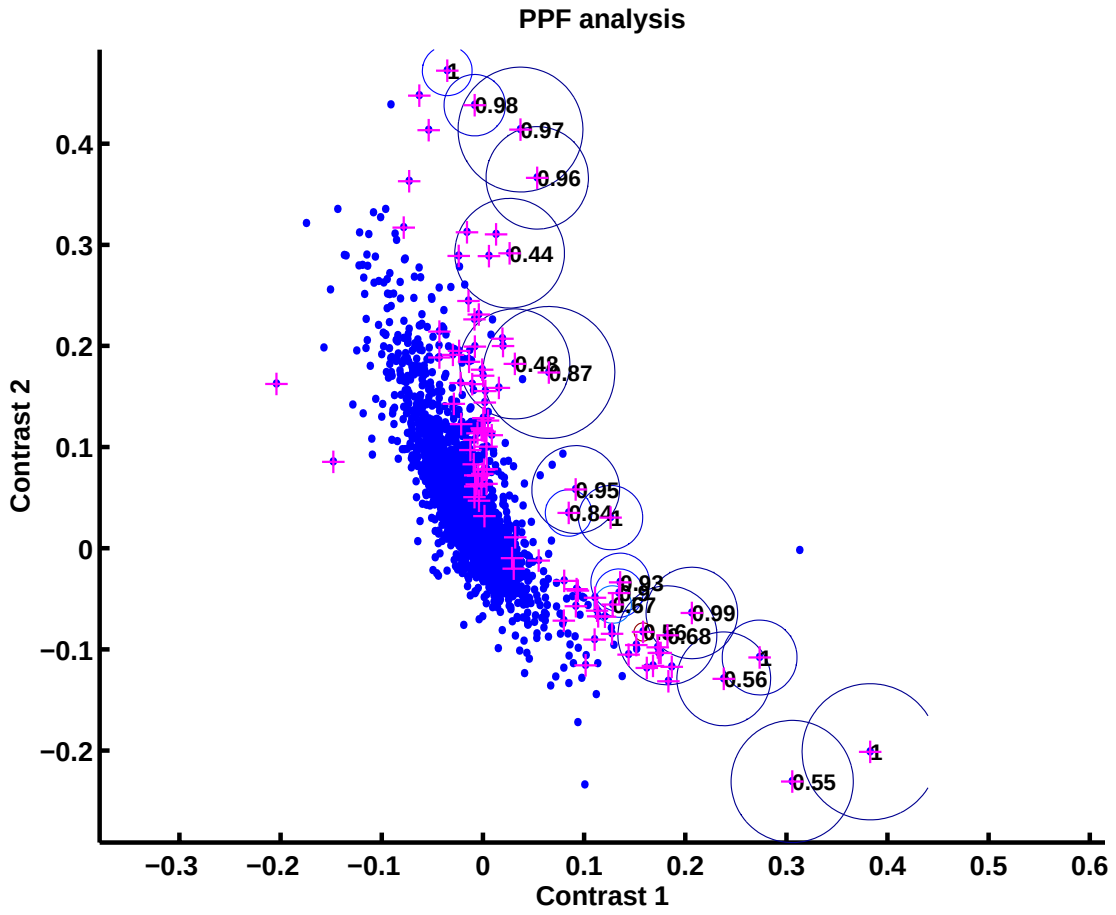

Figure 22: *PPF and posterior probabilities of belonging to the first Pareto front.*

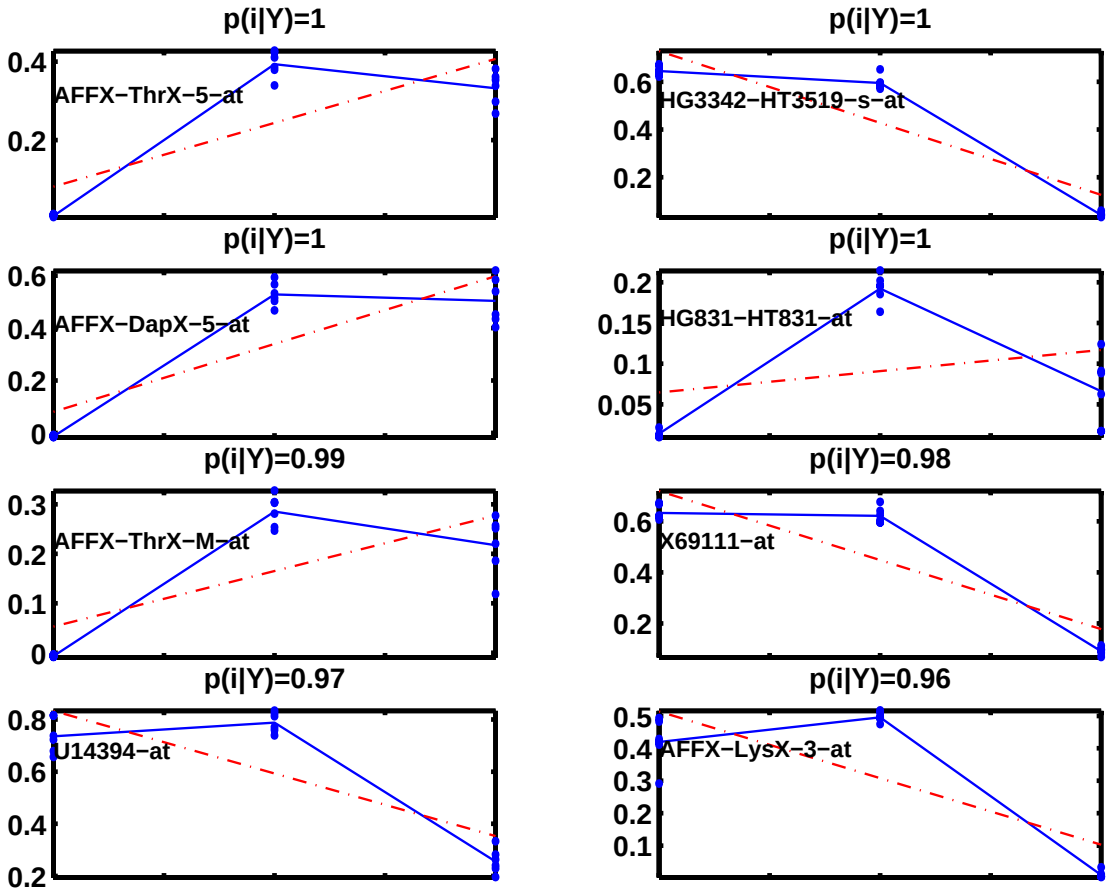

Figure 23: *The 8 top posterior ranked gene profiles remaining on the first Pareto front.*

Non-parametric Pareto filter criterion: Virtual Profiles

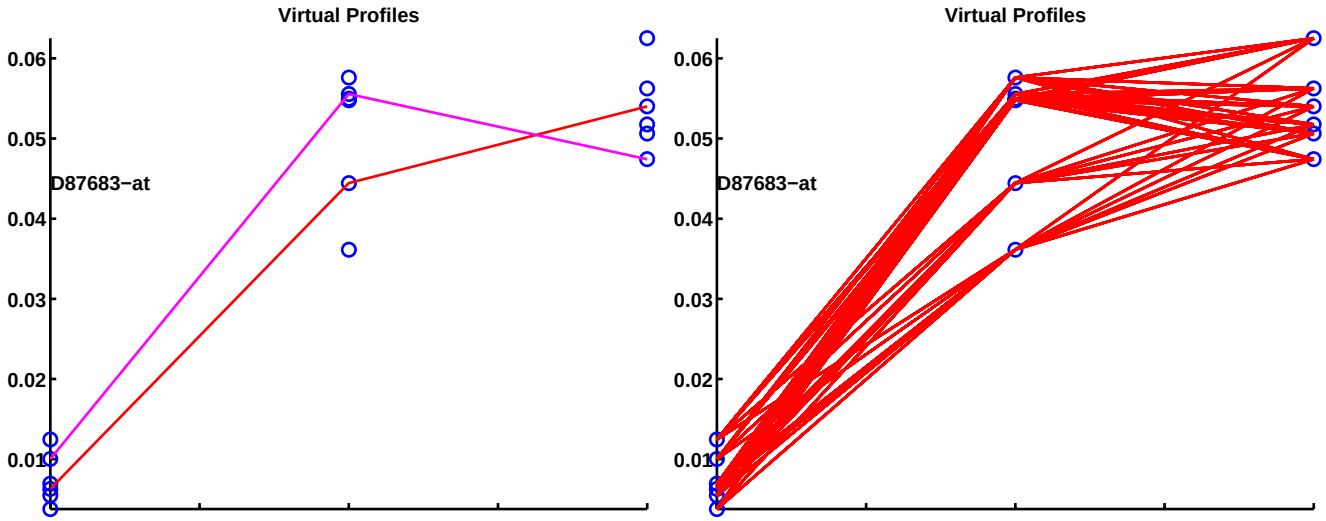

Figure 24: *Left: two virtual profiles in the data set. Right: the set of all* $3^6 = 729$ *virtual profiles for a gene in Fred Wright's dataset.*

Pareto Filtering using Virtual Sign-Profiles

Define *trend vector*: $\psi(n) = [b_1, \ldots, b_{T-1}]$, $b_i \in \{0, 1\}$

- Old dominant filtering criteria:
  - Maximum end-to-end increase

$$\xi_1(Y(n)) = \overline{y}_{T*}(n) - \overline{y}_{1*}(n) = \max$$

  - Maximum number of monotone increasing T^M virtual time profiles

$$\xi_2(Y(n)) = \frac{\text{\# virtual profiles having } \psi(n) = [1, \ldots, 1]}{T^M}$$

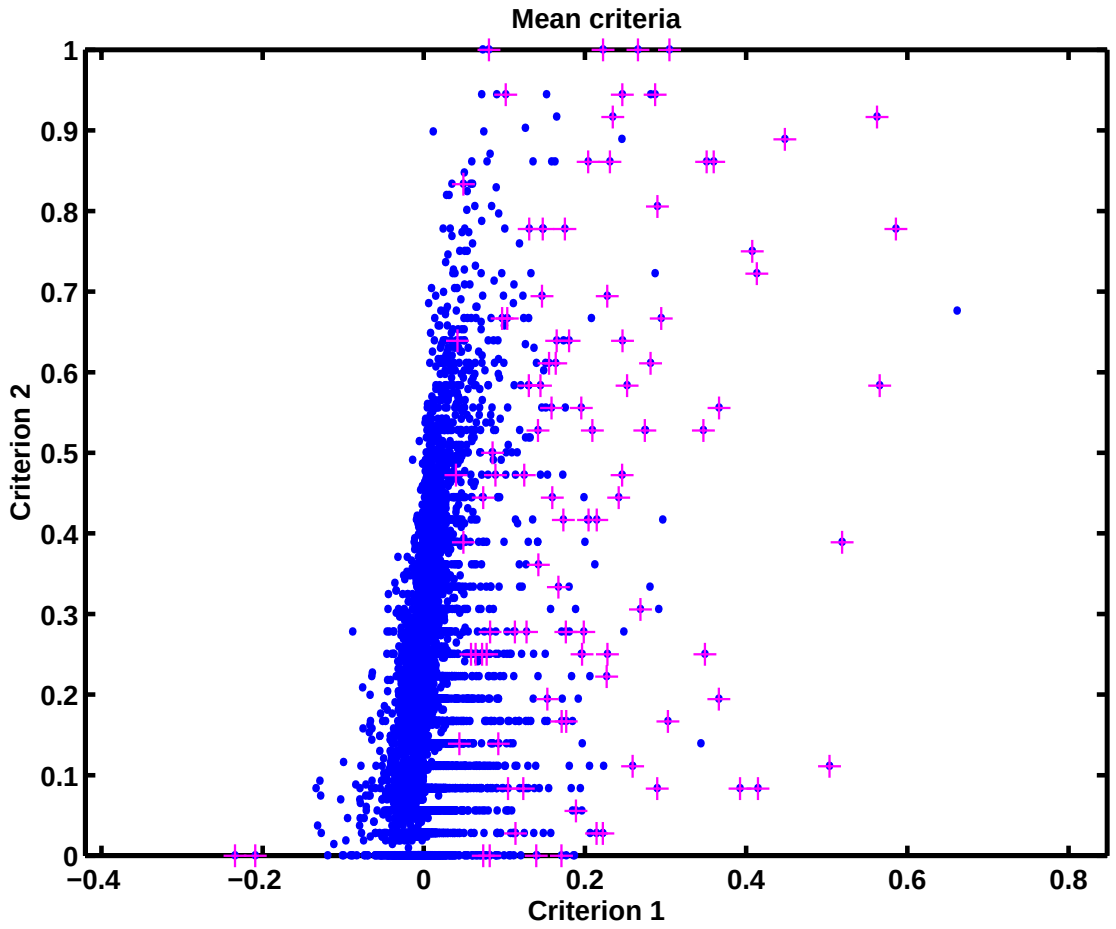

Figure 25: *Multicriterion mean scattergram for the virtual profile ranking and mean ene-to-end increase criteria.*

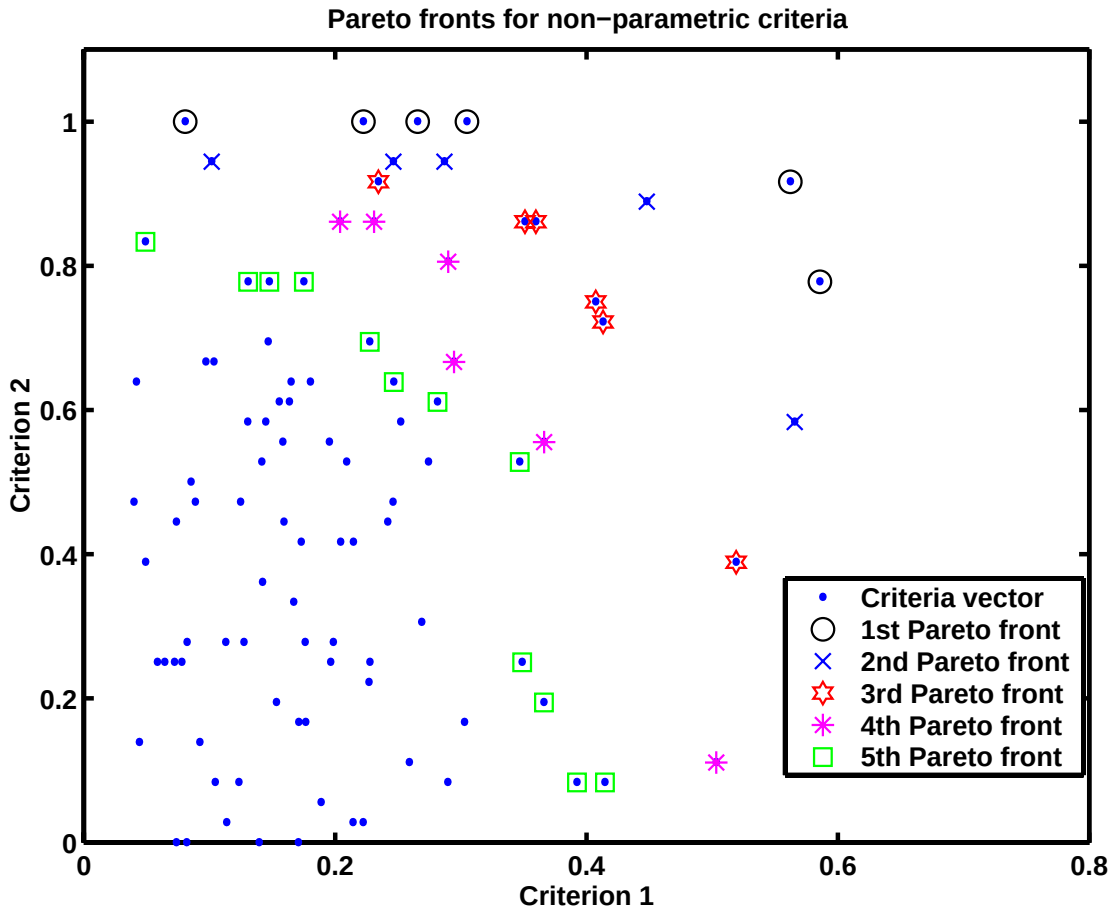

Figure 26: *The first five Pareto fronts.*

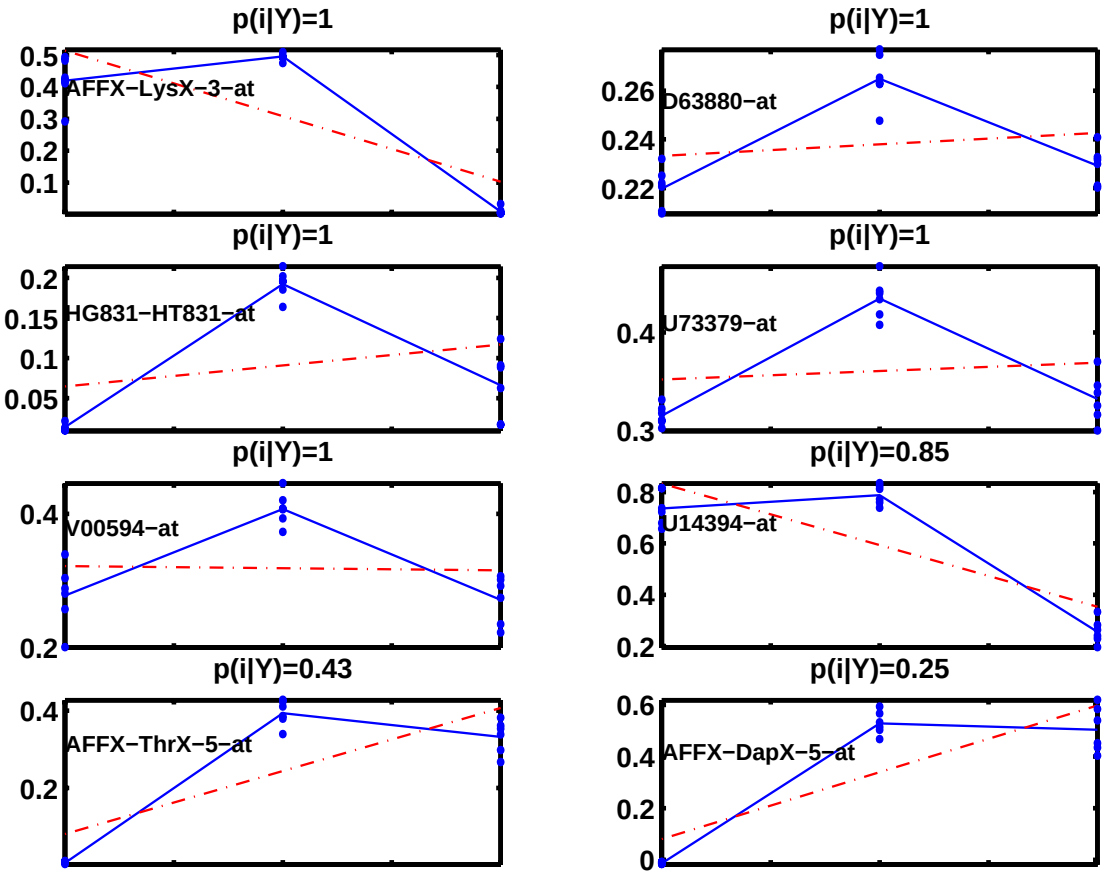

Figure 27: *The 8 top cross-validation ranked gene profiles.*

PPF linear contrast	P(I|Y)	RPF linear contrast	P(I|Y)	RPF non-parametric	P(I|Y)
AFFX-ThrX-5-at	0.999	AFFX-DapX-5-at	1	AFFX-LysX-3-at	1
HG3342-HT3519-s-at	0.998	AFFX-ThrX-5-at	1	D63880-at	1
AFFX-DapX-5-at	0.998	AFFX-ThrX-M-at	1	HG831-HT831-at	1
HG831-HT831-at	0.996	HG3342-HT3519-s-at	1	U73379-at	1
AFFX-ThrX-M-at	0.986	HG831-HT831-at	1	V00594-at	1
X69111-at	0.984	U14394-at	1	U14394-at	0.847
U14394-at	0.974	V00594-at	1	AFFX-ThrX-5-at	0.431
AFFX-LysX-3-at	0.962	X69111-at	1	AFFX-DapX-5-at	0.245
V00594-at	0.955	U45285-at	0.944	AFFX-PheX-3-at	0.222
U45285-at	0.932	AFFX-LysX-3-at	0.917	AFFX-HSAC07/X00351-5-at	0.208
AB000115-at	0.899	AFFX-HSAC07/X00351-5-at	0.806	AB000115-at	0.167
AFFX-HSAC07/X00351-5-at	0.866	AB000115-at	0.417	U00954-at	0.167
U73379-at	0.837	U73379-at	0.13	U45285-at	0.167
AFFX-DapX-M-at	0.678	V00594-s-at	0.074	U75362-at	0.167
Y09912-rna1-at	0.67	U75362-at	0.037	AFFX-ThrX-M-at	0.157
U75362-at	0.56	AFFX-PheX-5-at	0.028	HG1980-HT2023-at	0.032
AFFX-DapX-3-at	0.555	U03399-at	0.009	AFFX-PheX-M-at	0.028
V00594-s-at	0.554			U30998-at	0.028
HG1980-HT2023-at	0.483			Y09912-rna1-at	0.028
HG3044-HT3742-s-at	0.441				
D43636-at	0.389				
L27624-s-at	0.387				
U03399-at	0.378				
S69370-s-at	0.321				
AFFX-PheX-5-at	0.315				

Figure 28: *The top scoring genes (Affymetrix nomenclature).*

Mouse Retina Aging Study (2001M)

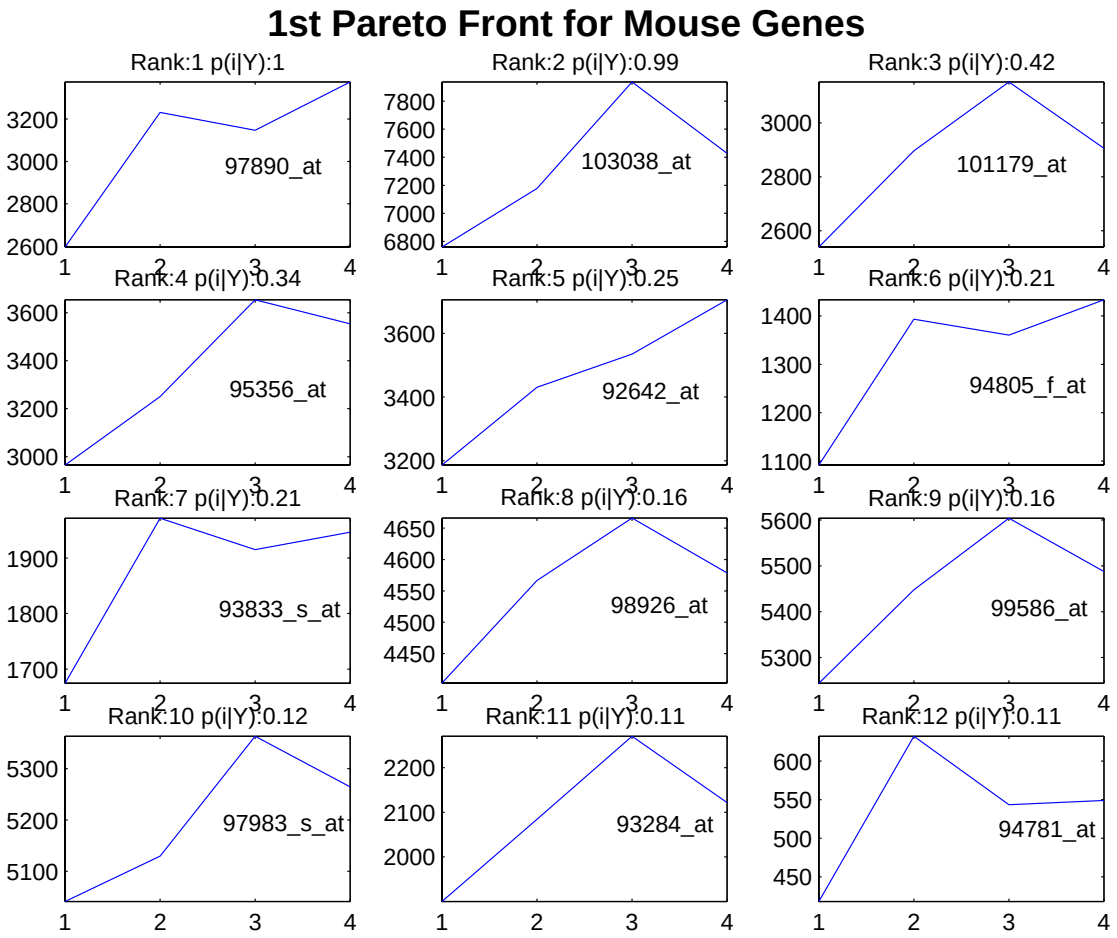

Figure 29: *Ranked first posterior Pareto front gene trajectories (Affy mouse study).*

Three-objective Pareto Filtering

Objective Extract "aging genes" in (2001M) study

- Strictly increasing filtering criteria:
  - Maximum end-to-end increase

$$\xi_1(Y(n)) = \overline{y}_{T*}(n) - \overline{y}_{1*}(n) = \max$$

  - Maximum number of monotone increasing T^M virtual time profiles

$$\xi_2(Y(n)) = \frac{\#\text{ virtual profiles having } \psi(n) = [1,\ldots,1]}{T^M}$$

  - no plateau

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

Pairwise Pareto Fronts

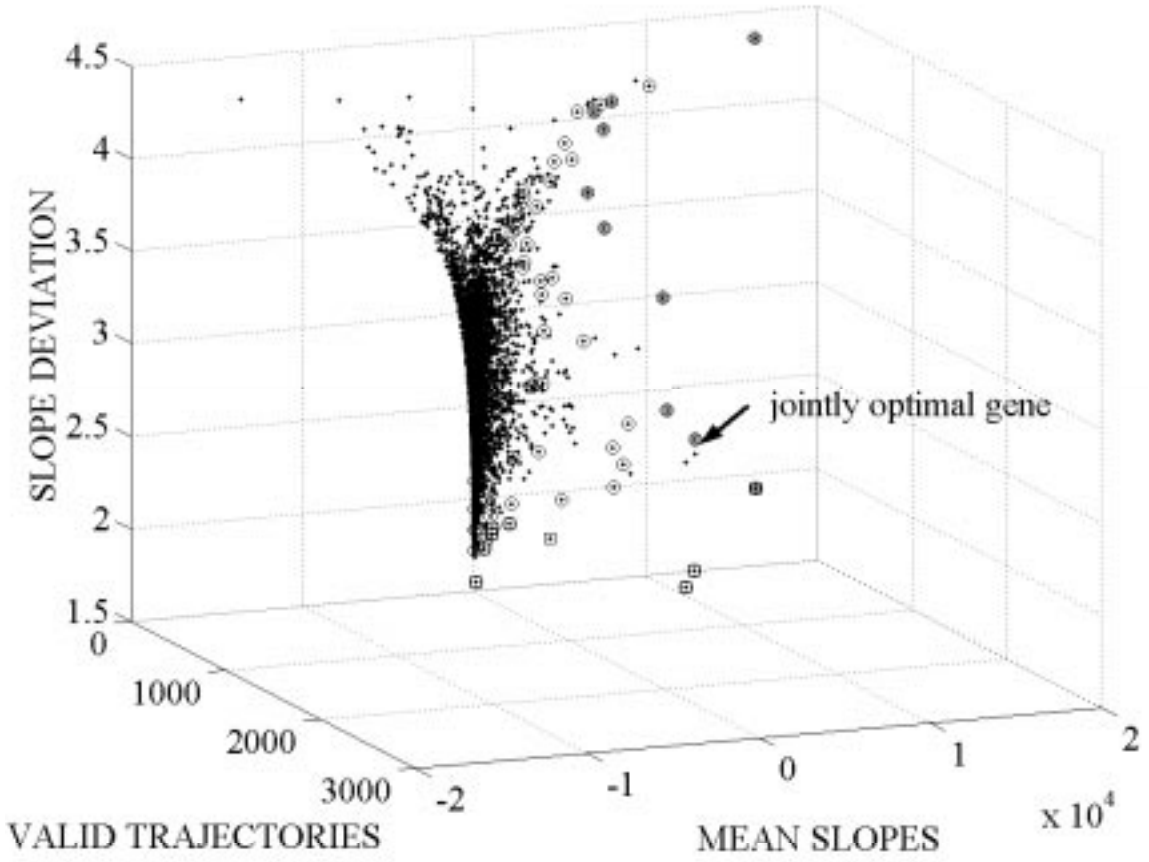

Figure 30: *First Pareto fronts for each pair of criteria taken from the set (ξ_1, ξ_2 and ξ_3).*

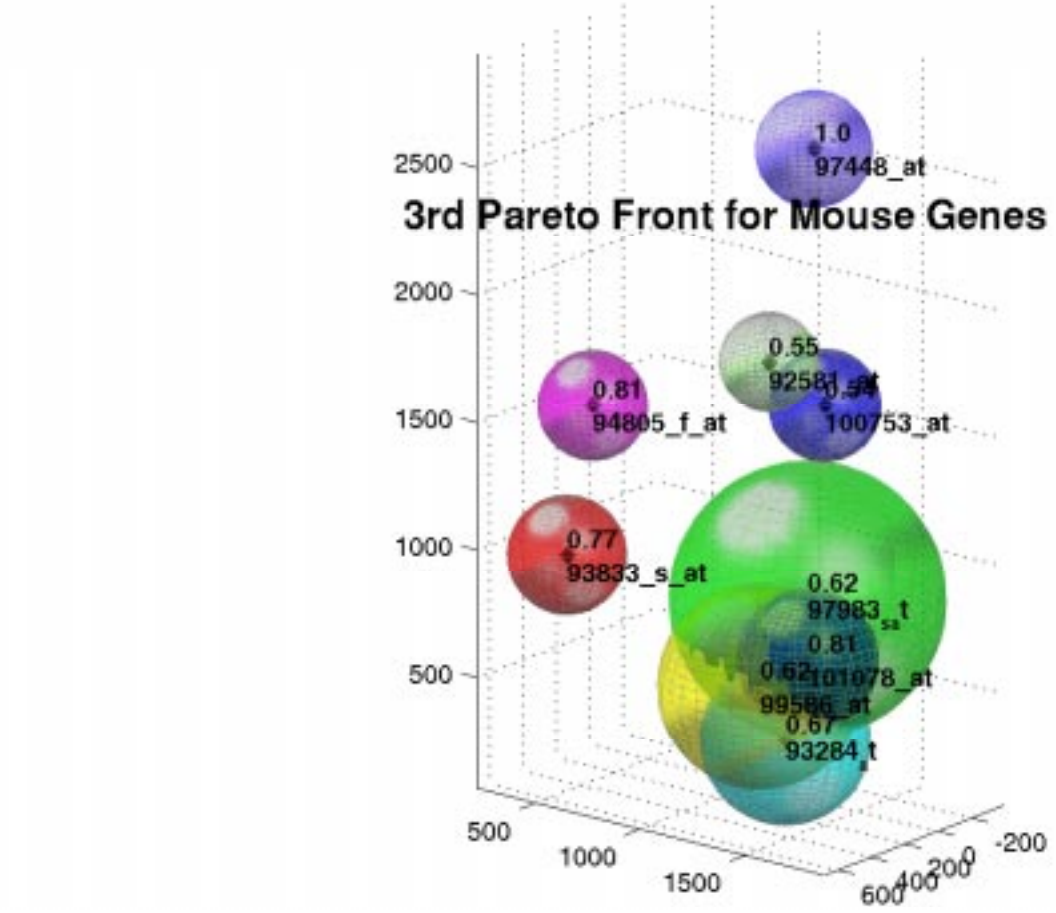

Figure 31: *Third posterior Pareto front for (affy mouse study).*

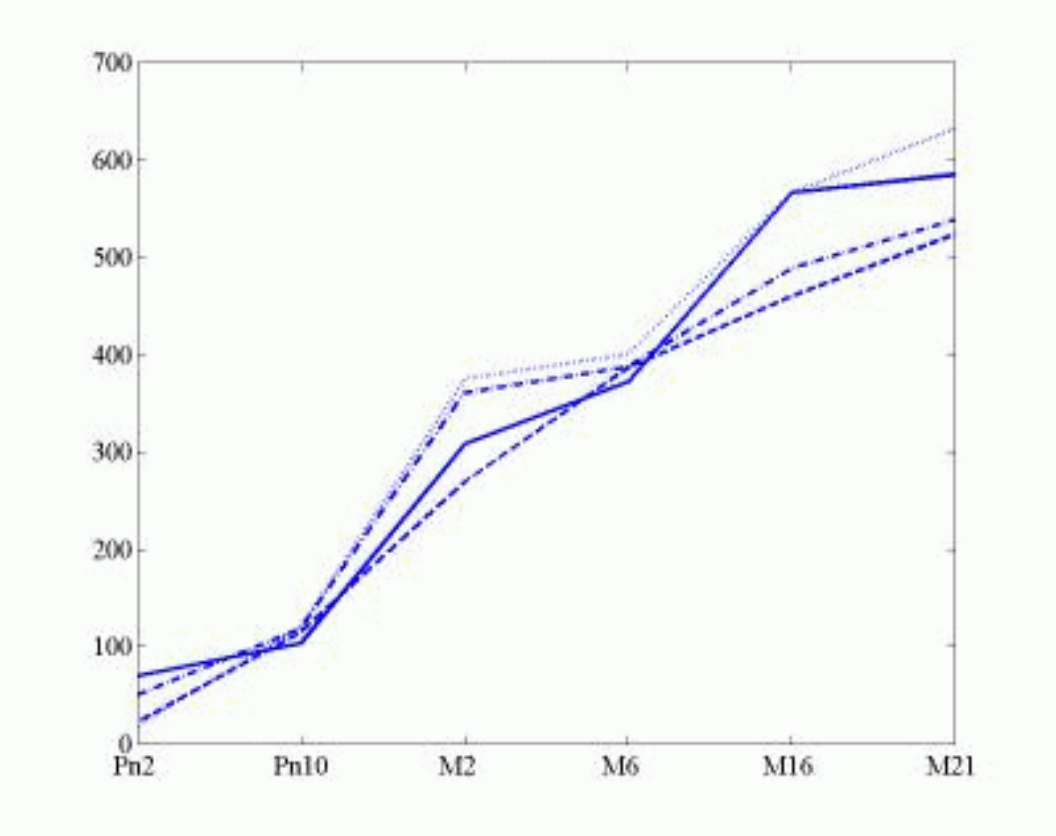

Figure 32: *Top ranked gene profile is Mus musculus* $5'$ *end cDNA (Unigene 86632)*

Conclusions

1. Multi-criterion data mining can perform robust and flexible gene filtering
2. Cross-validation can account for statistical sampling uncertainty
3. Non-informative priors can be used to find posterior front probabilities
4. Genetic priors: phylogenetic trees, BLAST database, etc?