Gene Filtering with Multi-objective Optimization Criteria

A. O. Hero

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 clustering and filtering
2. Pareto filtering for gene pattern extraction
3. Application: development and aging in retina

Scientific Objectives

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

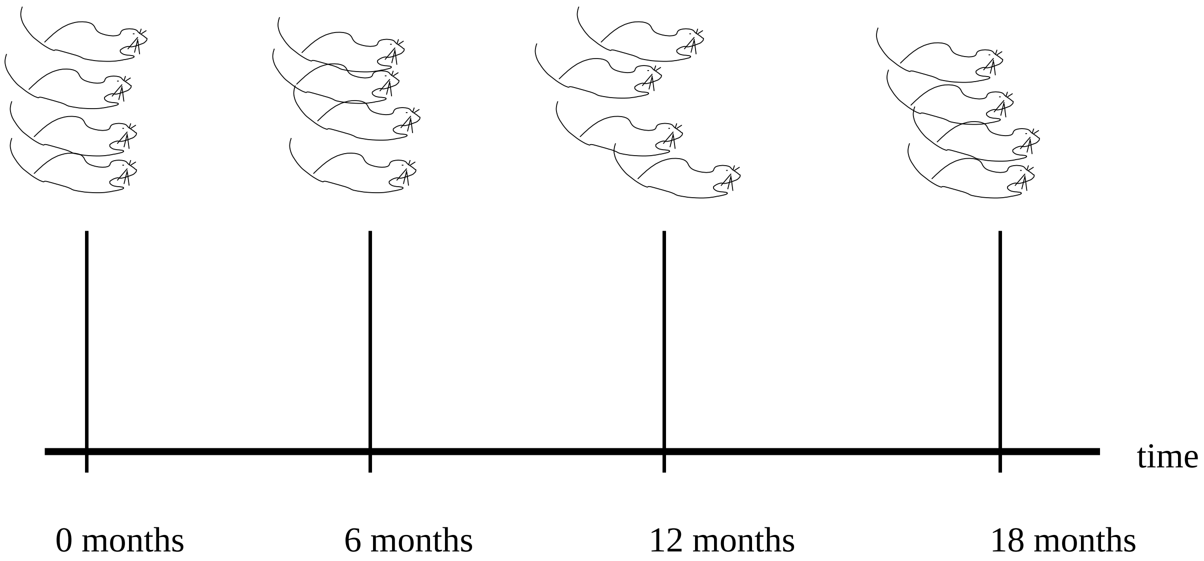
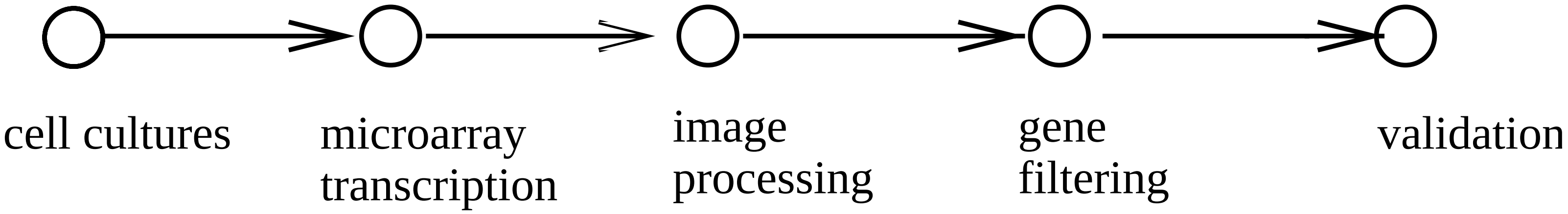


Figure 1: *Sample gene trajectories over time.*

Gene Microarrays



cell cultures

microarray transcription

image processing

gene filtering

validation

Figure 2: *Microarray experiment cycle.*

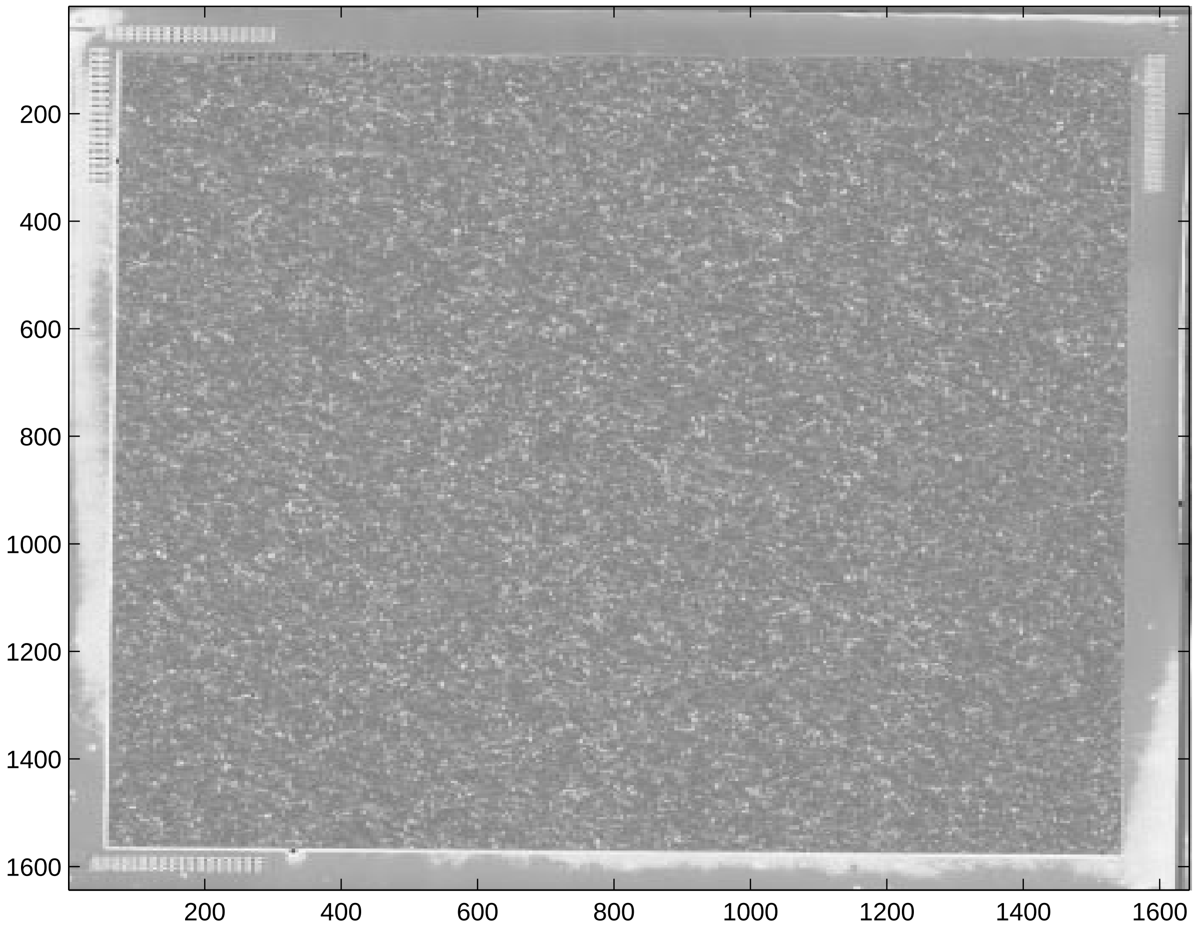


Figure 3: *Affymetrix GeneChip microarray.*

(U95 GeneChip) 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

Control Factors Influencing Variability

- **Sample preparation**: reagent quality, temperature variations
- **Slide manufacture**: slide surface quality, dust deposition
- **Hybridization**: sample concentration, wash conditions
- **Image formation**: scanner saturation, lens aberations, gain settings
- **Imaging and Extraction**: spot misalignment, discretization, clutter

→ account for data variability

- **Scaling factors**: universal intensity amplification factor for a chip
- **Raw Q**: noise and other random variations of a chip
- **Background**: avg of lowest 2% cell intensity values
- **% P**: percentage of transcripts present

Gene Clustering and Filtering

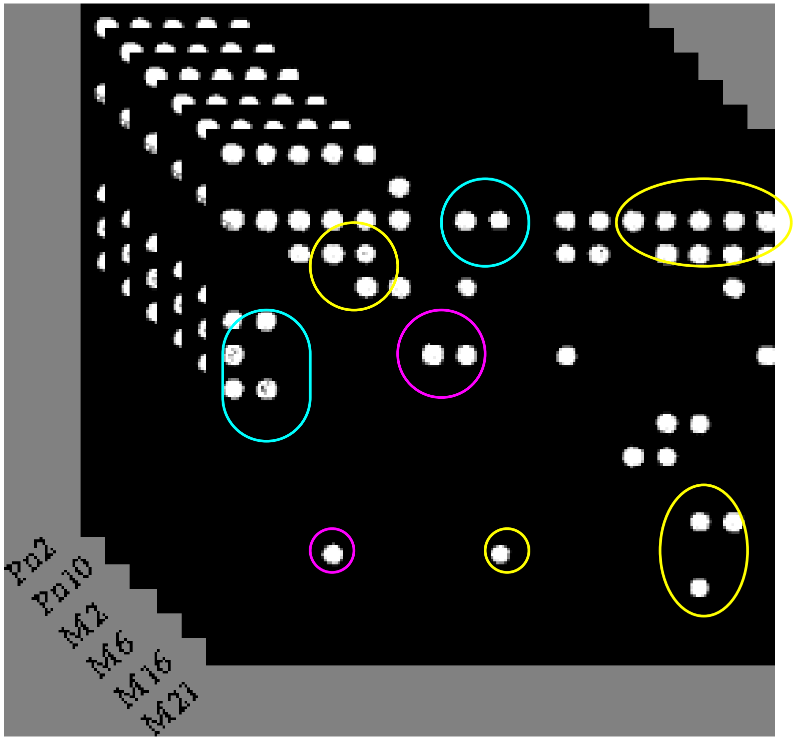


Figure 4: *Clustering on the Data Cube.*

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

Gene Trajectory Classification

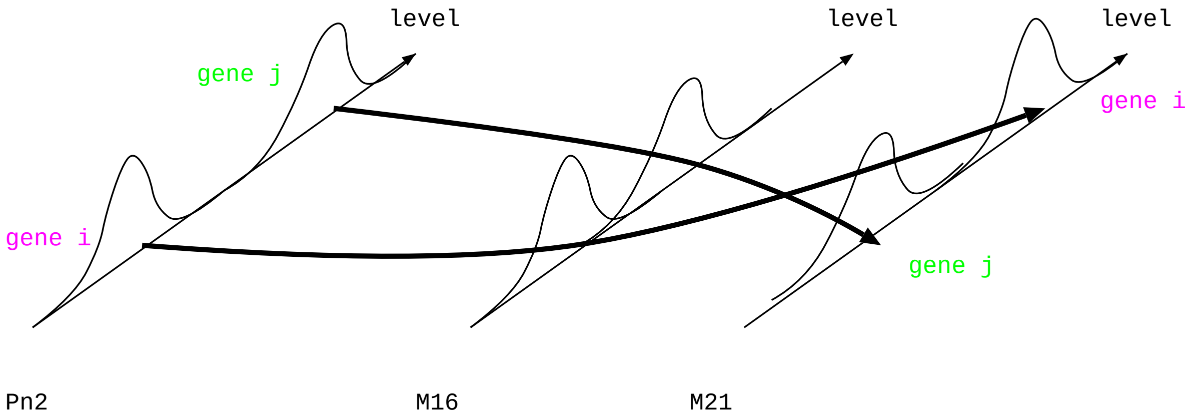


Figure 5: *Gene i is old dominant while gene j is young dominant*

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

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

Clustering and filtering Methods

Principal approaches:

- Relational database search - non-statistical (CheckMate, DiscoverIt)
- Hierarchical clustering (kdb trees, mixture models, gene shaving)
- K-means clustering
- Self organizing (Kohonen) maps (SOM)
- 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 i-th gene $\xi_1(\theta_i), \;\; , \ldots , \;\; \xi_P(\theta_i)$

Examples of $\xi_p(\theta_i)$:

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

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

- 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 cost "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

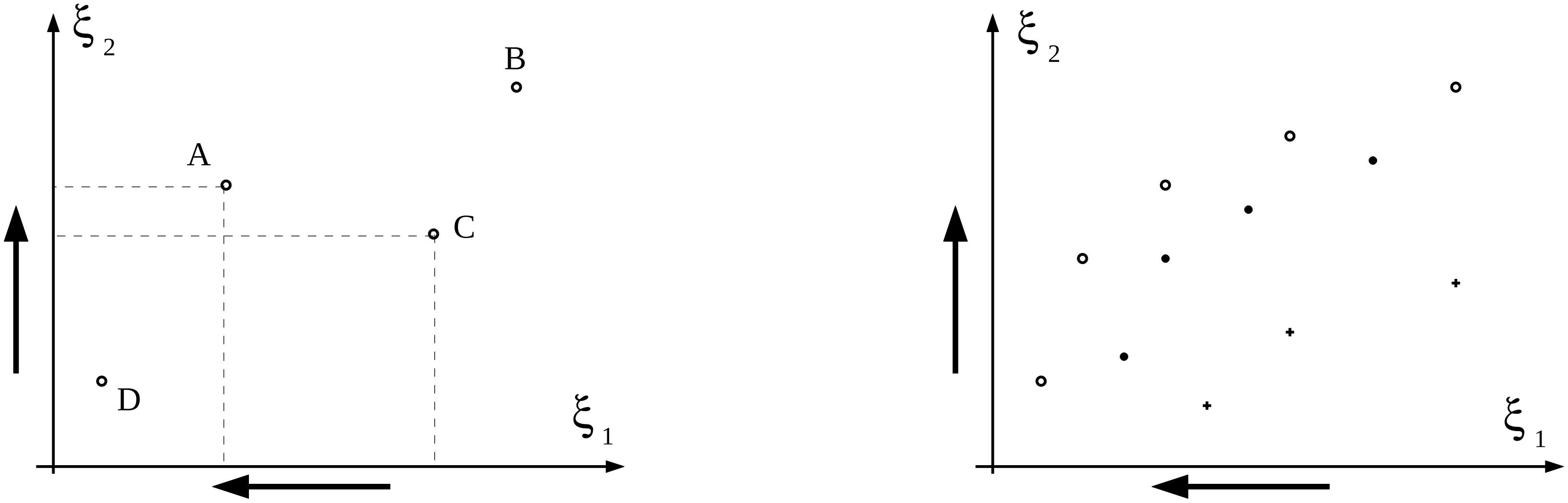


Figure 6: *a). Non-dominated property, and b). Pareto optimal fronts, in dual criteria plane.*

Pareto Gene Filtering vs. Paired T-test

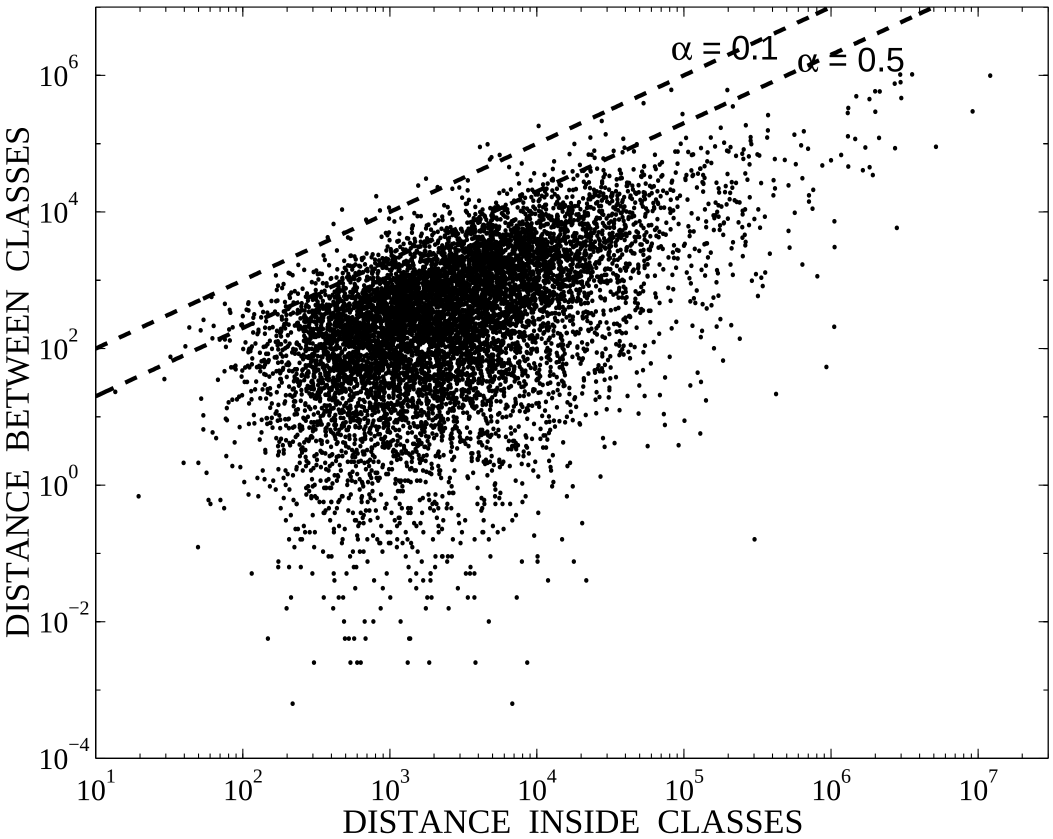


Figure 7: $\xi_1 =$ *mean change vs* $\xi_2 =$ *pooled standard deviation for 8826 mouse retina genes. Superimposed are T-test boundaries*

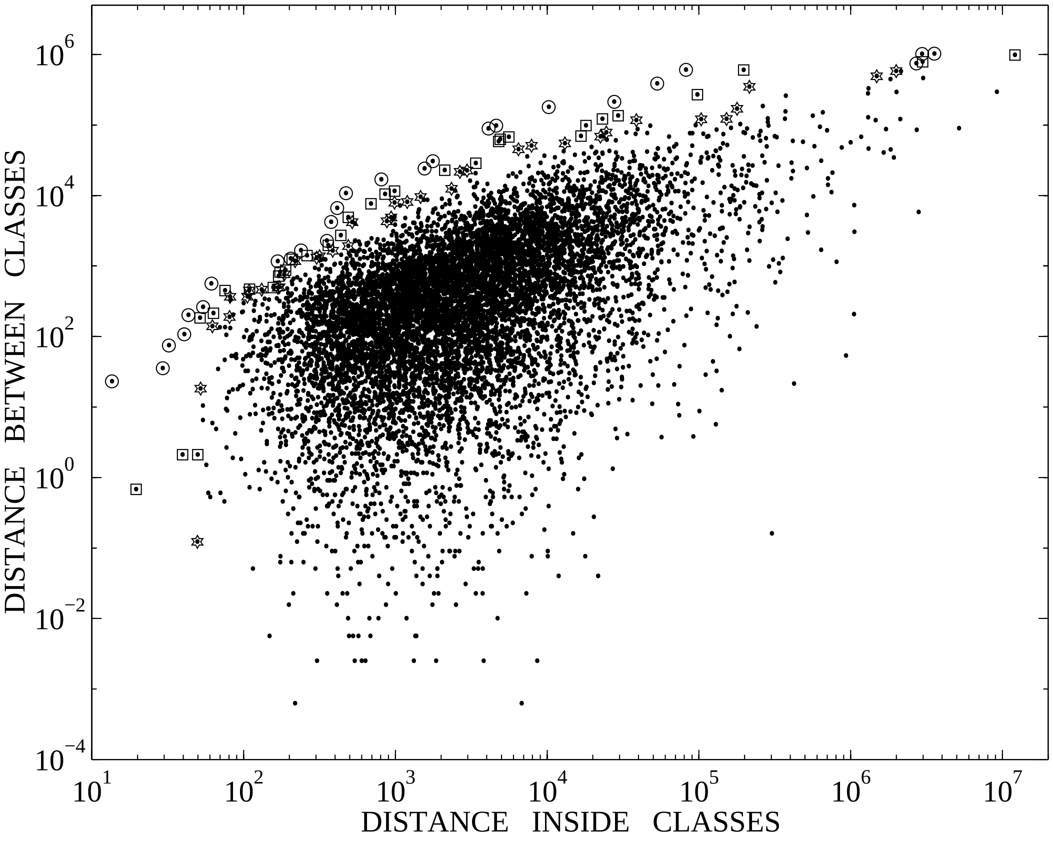


Figure 8: *First (circle) second (square) and third (hexagon) Pareto optimal fronts.*

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

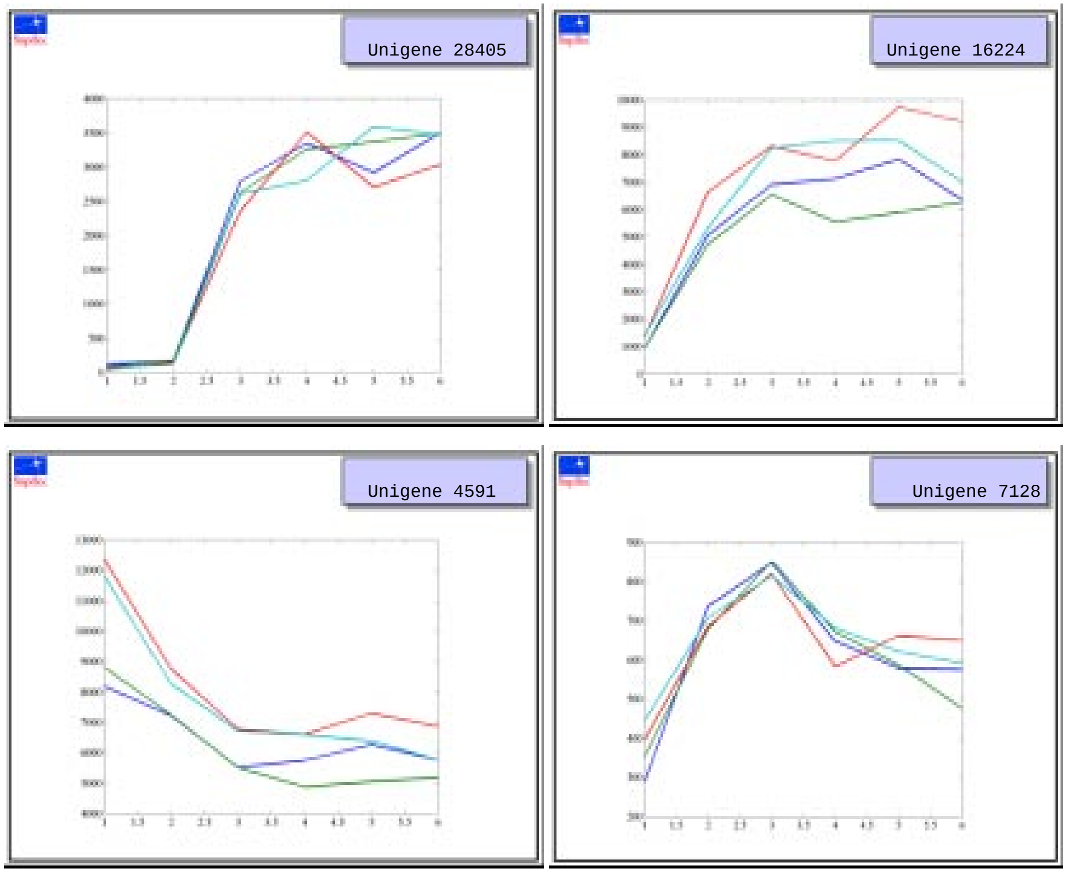


Figure 9: *Trajectories.*

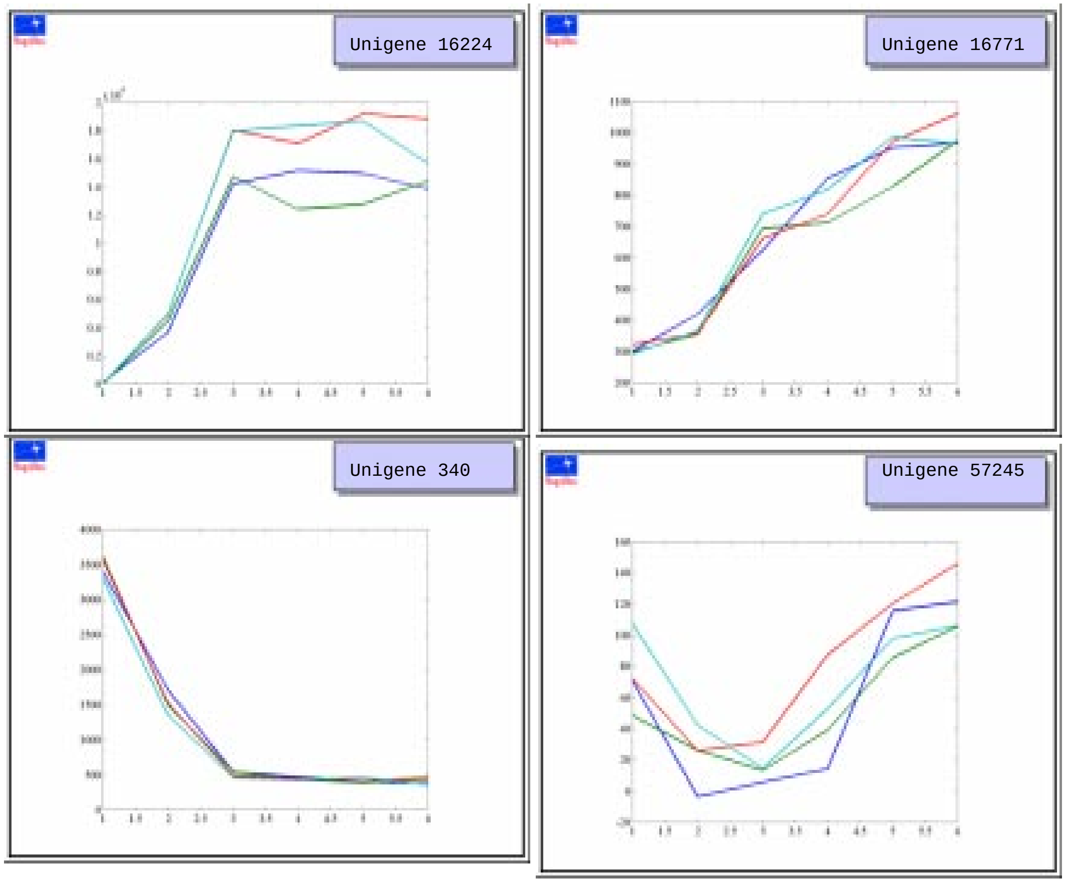


Figure 10: *Trajectories.*

Pairs of Trajectories for Replicated Segments

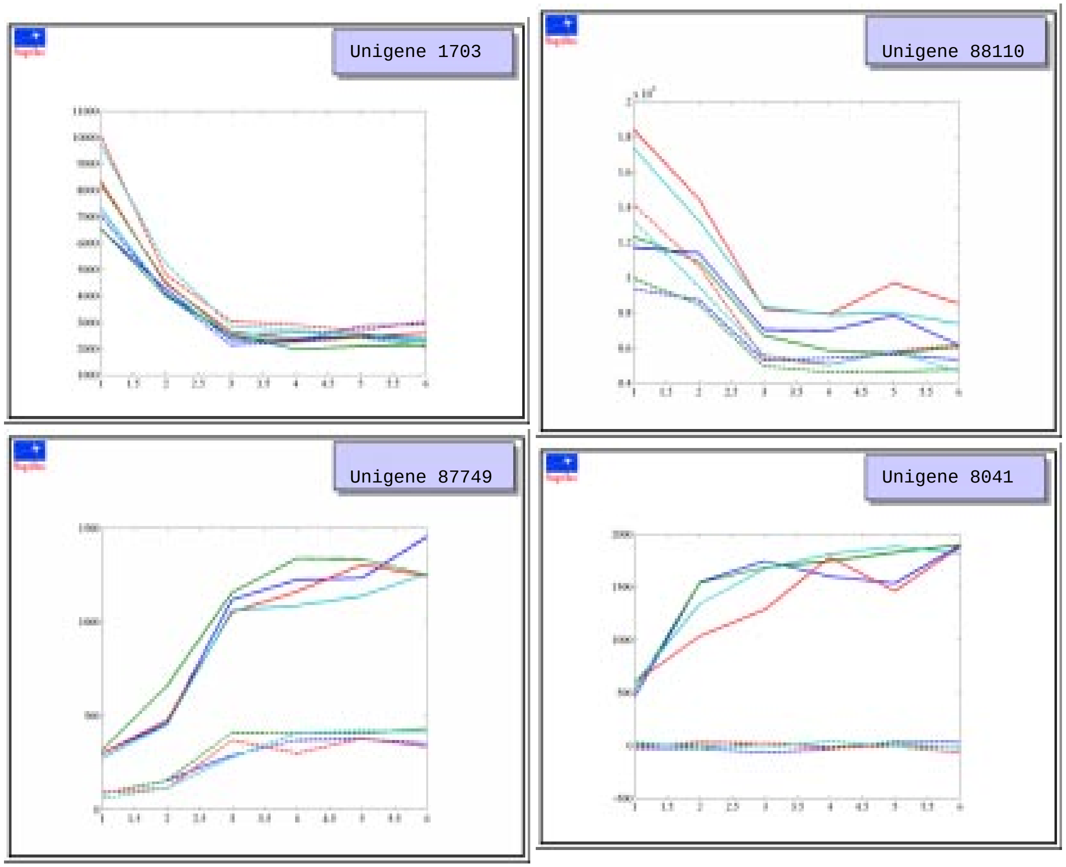


Figure 11: *Pairs of trajectories for replicated gene polynucleotide sequence.*

Popular Methods for Gene Profile Filtering

Unsupervised: Principal components analysis, hierarchical clustering

Supervised: least-squares template-clustering algorithm:

`Step 1:` define templates for temporal profiles of interest

`Step 2:` estimate gene profiles (LS/ANOVA)

`Step 3:` fit estimated gene profiles to templates

- **i**: best template-correlation (Chu&etal Science 1998)
- **ii**: best 95% confidence fit (Kerr&Churchill PNAS 2001)

`Step 4:` create clusters by pruning-off poorly fitting genes

`Step 5:` assess cluster reliability (bootstrap).

Drawbacks

- Gene profile variability requires difficult ”deformable” templates
- Most of these methods are sensitive to scaling and translation
- Linear fitting methods lose sensitivity for heavy-tailed noise
- ANOVA residual fitting errors are sensitive to outliers
- Bootstrap computation is a bottleneck
- scalar clustering and filtering criterion

Multi-objective Non-parametric Pareto Filtering

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

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

$$\xi_1(\theta_i) = \overline{\theta_i(T,*)} - \overline{\theta_i(1,*)} = \max$$

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

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

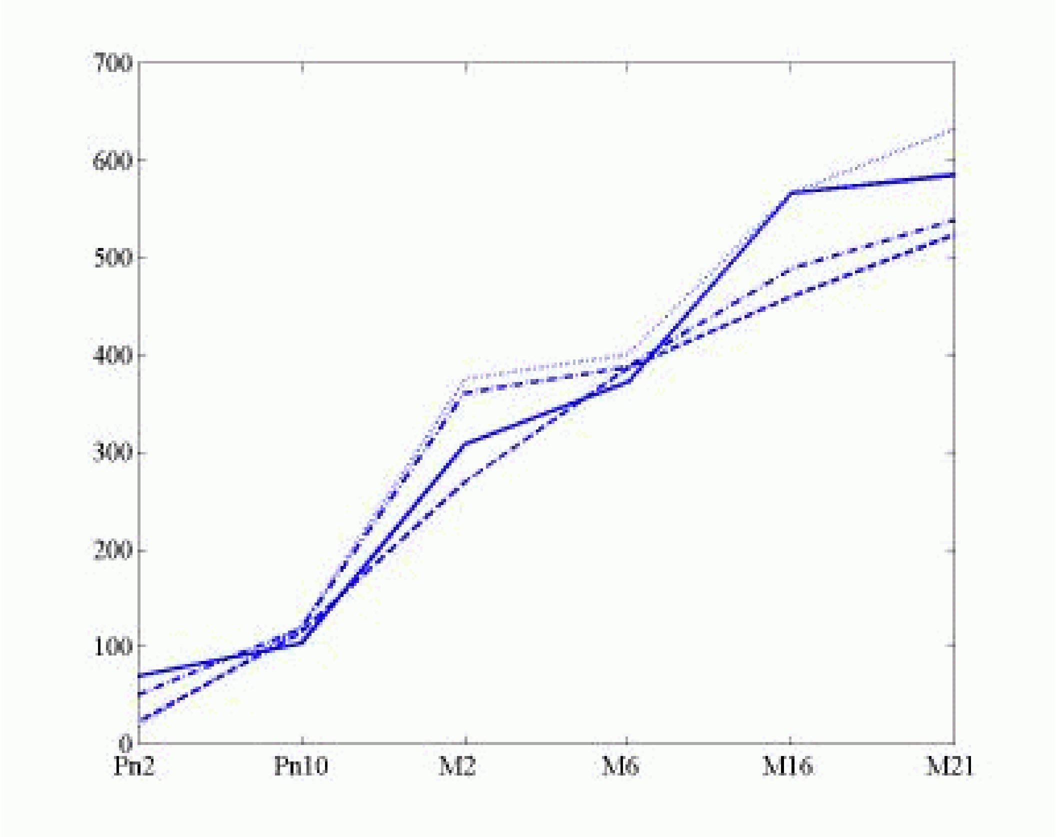


Figure 12: *4 candidate gene profiles from Mus musculus* $5'$ *end cDNA (Unigene 86632)*

Occurence Histogram

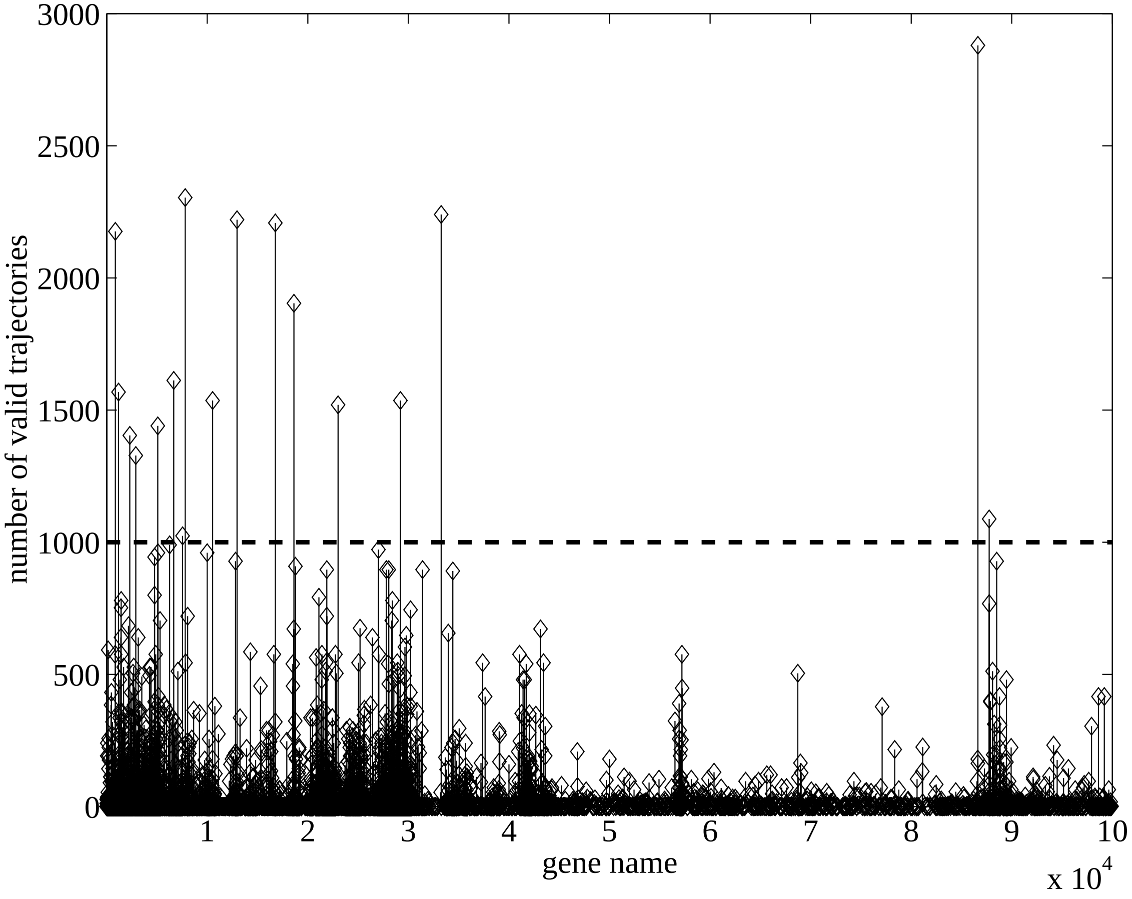


Figure 13: *Occurrence histogram with threshold.*

Old Dominant Pareto Fronts

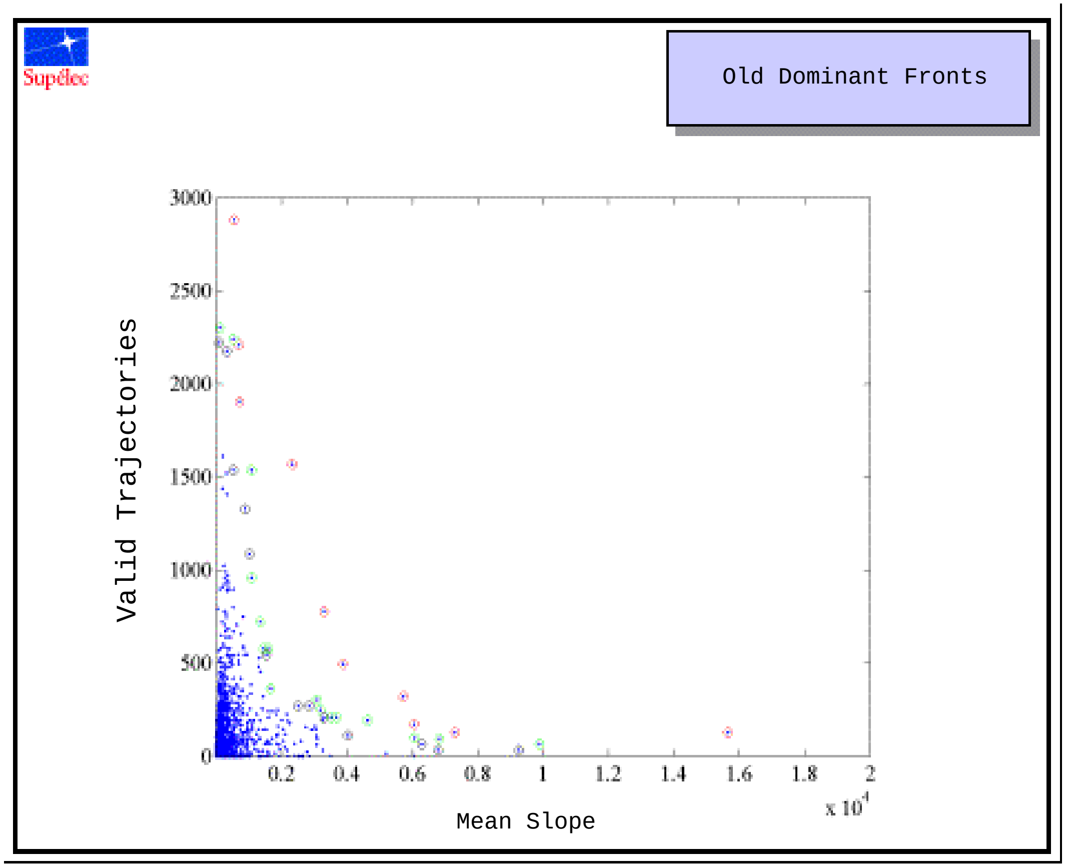


Figure 14: *Pareto fronts for old dominant genes.*

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

   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 $t = M21$

$$\xi_1(\theta_i) = \overline{\theta_i(T,*)} - \overline{\theta_i(1,*)} = \min$$

- 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

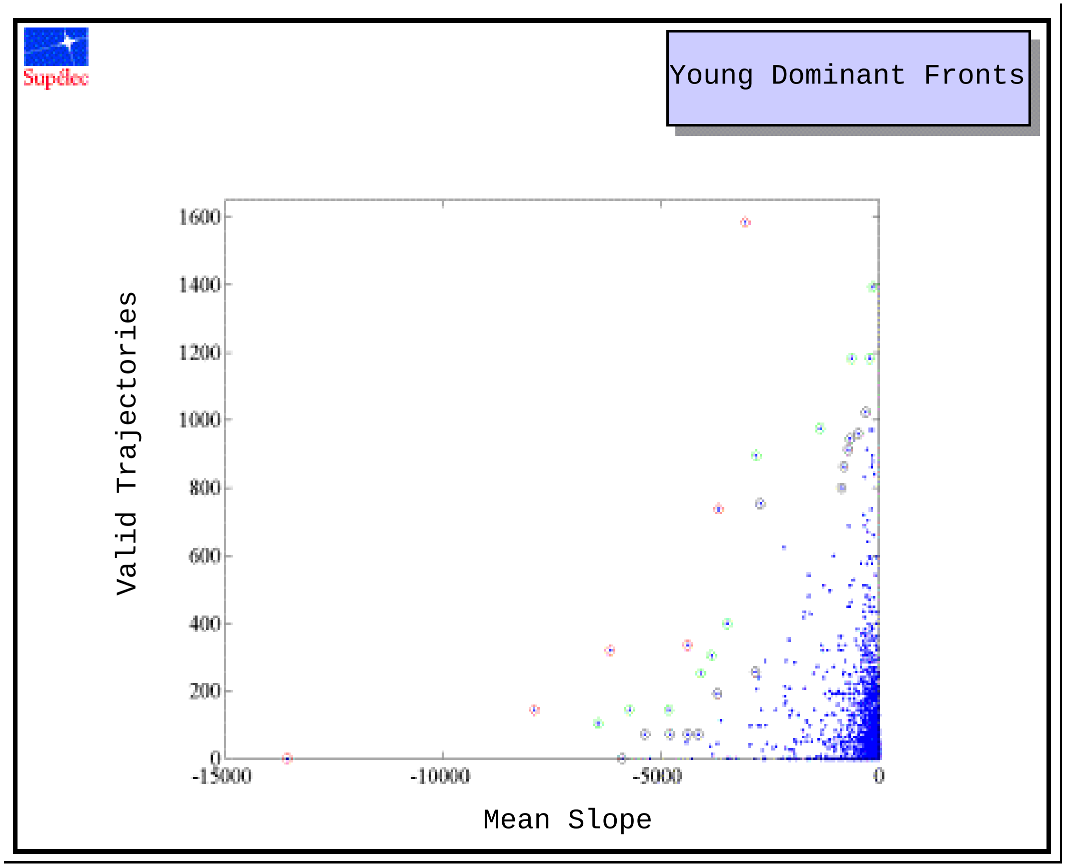

Young Dominant Fronts

Figure 15: *Pareto fronts for young dominant genes.*

Three-objective Pareto Filtering

Objective Extract “aging genes”

- Strictly increasing filtering criteria:
  - persistent positive trend from M2-M21

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

  - high consistency over $4^4 = 256$ possible combinations of trajectories

$$\xi_2(\psi_i) = \frac{\text{\# trajectories having } \psi_i = [1,\ldots,1]}{256} = \max$$

  - no plateau

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

Pareto Fronts

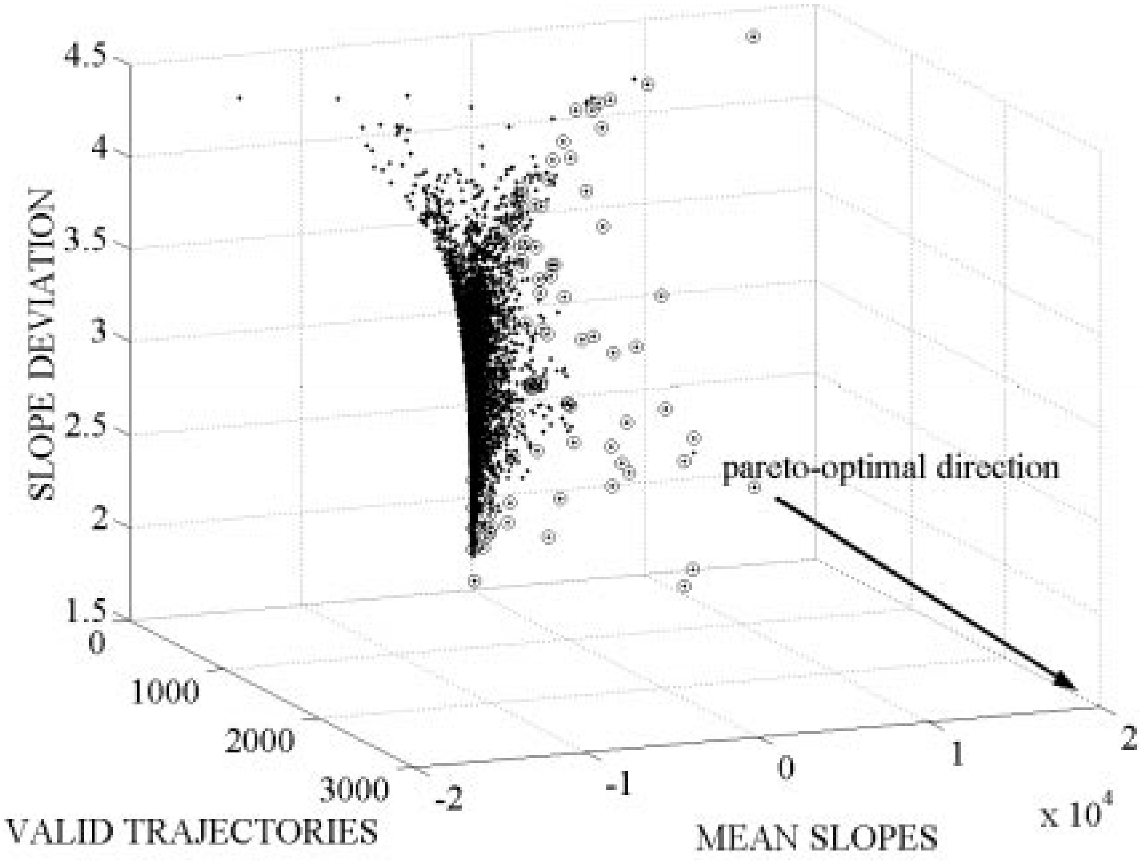


Figure 16: *First global Pareto front (o) for the three criteria (ξ_1, ξ_2 and ξ_3).*

Pairwise Pareto Fronts

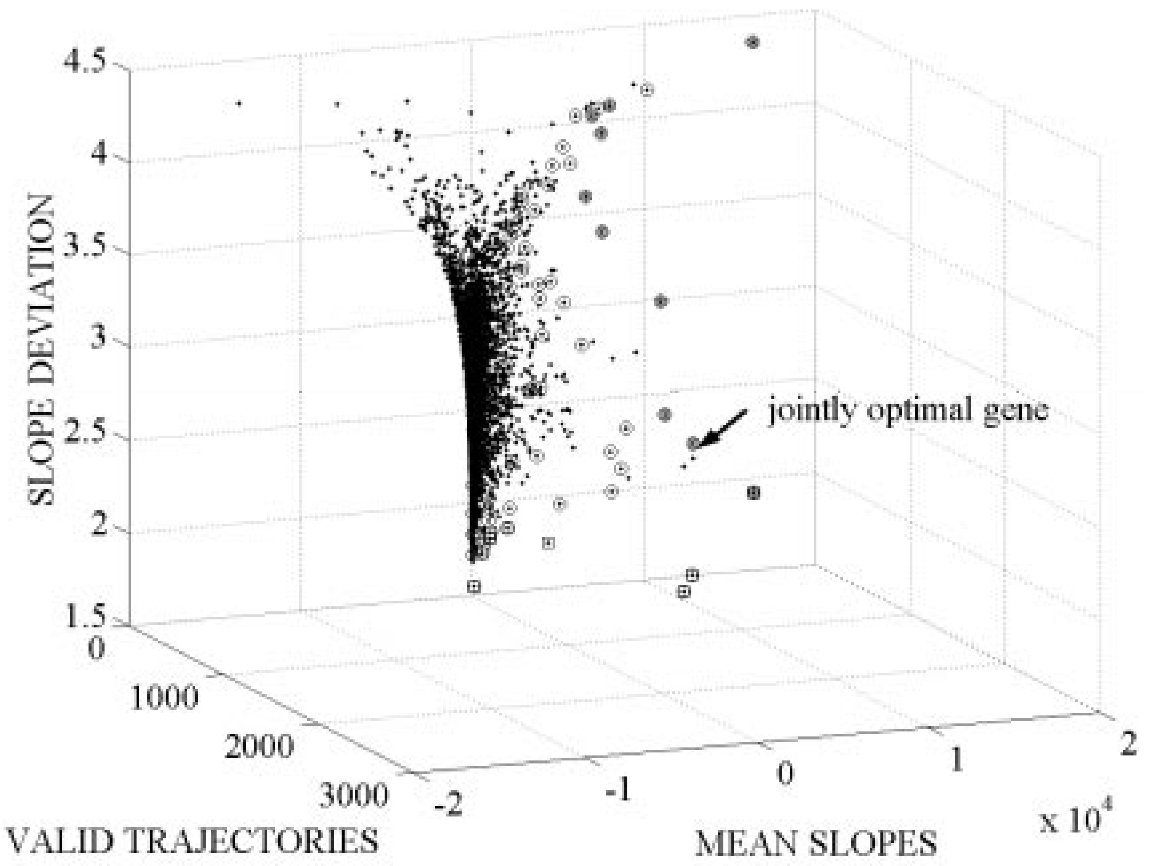


Figure 17: *First Pareto fronts for each pair of criteria taken from the set (ξ_1, ξ_2 and ξ_3). Each one of this front is denoted by squares, circles and stars, respectively.*

Aging Genes Found by Pareto Filter

Unigene #	Front	Description
7800	1st	Inositol triphosphate receptor type 2
86632	2nd	Histocompatibility 2, L Region
12956	2nd	Hyperpolarization-activated, cylcic nucleotide-gated K
29213	3rd	RIKEN cDNA 1200015F23 gene
33263	3rd	Histocompatibility 2, D region locus 1
29789	3rd	Expressed sequence A1430822
2289	3rd	RIKEN cDNA 1500015A01 gene
6671	3rd	RIKEN cDNA 1110027O12 gene
16771	4th	MHC class 1 antigen H-2K
34421	4th	Q4 class 1 MHC
6252	4th	Procollagen, type XIX, alpha 1
29357	4th	RIKEN cDNA 1300017C10 gene

Table 3: *Resistant aging genes remaining in first four Pareto fronts*

Conclusions

1. Pareto filtering performs robust and flexible gene filtering
2. Statistical sampling uncertainty can be reduced by cross-validation
3. Joint intensity extraction and gene filtering?
4. Evolutionary optimization algorithms for large data sets?
5. Large sample theory of Pareto fronts?