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
Kellogg 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: Oligonucleotide PM/MM layout (pathbox.wustl.edu).
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 4: 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 5:** 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 6: 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 7: Differential responses for a gene in mouse k (left) vs w (right) study.
Figure 8: 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 \( t \)

\[ y_{im}(n), \quad n = 1, \ldots, N, \quad t = 1, \ldots, T, \quad m = 1, \ldots, M. \]
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 $t = T$:
  \[ \xi_1(Y(n)) = |\bar{y}_{T*}(n) - \bar{y}_{1*}(n)|^2 \]

- Standard deviation at $t = 1$:
  \[ \xi_2(Y(n)) = (y_{1m}(n) - \bar{y}_{1*}(n))^2 \]

- Standard deviation at $t = T$:
  \[ \xi_3(Y(n)) = (y_{Tm}(n) - \bar{y}_{T*}(n))^2 \]
Some possible scalar functions:

- $t$-test statistic (Goss et al 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 et al 2000): $R^2(n) = \frac{T_n}{1+T_n}$

- $H$ statistic (Sinha et al 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”

\[
\sum_{p=1}^P W_p = 1, \quad W_i \geq 0
\]

Find optimal gene via:

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

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.

\[
\xi_p(Y(i)) \leq \xi_p(Y(j)), \quad p = 1, \ldots, P
\]
Figure 9: For 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.
Figure 10: *a)* Non-dominated property, and *b)* Pareto optimal fronts, in dual criteria plane.
Figure 11: $\xi_1 = \text{mean change vs } \xi_2 = \text{pooled standard deviation}$ for 8826 human retina genes (2001H). Superimposed are T-test boundaries.
Figure 12: 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: Posterior Pareto Analysis

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

\[
p(i|Y) = P\left( \bigcap_{j \neq i} \left\{ \xi(i) \leq \xi(j) \right\}^c | Y \right) \\
= \int dP(\xi(i)|Y) \prod_{j \neq i} P\left( \left\{ \xi(i) \leq \xi(j) \right\}^c | Y, \xi(i) \right)
\]

Case of two criteria \((P = 2)\)

\[
p(i|Y) = \int \int du_1 du_2 f_{\xi_1(i), \xi_2(i)|Y}(u_1, u_2) \prod_{j \neq i} [F_{\xi_1(j)|Y}(u_1) + F_{\xi_2(j)|Y}(u_2) - F_{\xi_1(j), \xi_2(j)|Y}(u_1, u_2)]
\]
Application to Fred Wright’s Mixture Study

Figure 13: 8 ranked monotone decreasing gene profiles.
Figure 14: *Multicriterion scattergram for first two rows of \( \tilde{A}_3 \).*
Figure 15: Multicriterion scattergram for $A = [-1, 1, 0; -1, -1, 2]$. 98 genes are non-linear profiles (p-value of 0.1).
Figure 16: The first five Pareto fronts for the genes with non-linear profiles shown in Fig. 15.
Figure 17: 17 genes in first Pareto front with non-zero probability by cross-validation.
Figure 18: The 8 top cross-validation ranked gene profiles remaining on the first Pareto front.
Figure 19: *PPF and posterior probabilities of belonging to the first Pareto front.*
Figure 20: The 8 top posterior ranked gene profiles remaining on the first Pareto front.
Non-parametric Pareto filter criterion: Virtual Profiles

Figure 21: 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.
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
    \[
    \xi_1(Y(n)) = \bar{y}_{T*}(n) - \bar{y}_{1*}(n) = \max
    \]
  - 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 22: Multicriterion mean scattergram for the virtual profile ranking and mean ene-to-end increase criteria.
Figure 23: The first five Pareto fronts.
Figure 24: The 8 top cross-validation ranked gene profiles.
<table>
<thead>
<tr>
<th>PPF linear contrast</th>
<th>P(IV)</th>
<th>RPF linear contrast</th>
<th>P(IV)</th>
<th>RPF non-parametric</th>
<th>P(IV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFFX-ThrX-5-at</td>
<td>0.999</td>
<td>AFFX-DapX-5-at</td>
<td>1</td>
<td>AFFX-LysX-3-at</td>
<td>1</td>
</tr>
<tr>
<td>HG3342-HT3519-s-at</td>
<td>0.998</td>
<td>AFFX-ThrX-5-at</td>
<td>1</td>
<td>D63880-at</td>
<td>1</td>
</tr>
<tr>
<td>AFFX-DapX-5-at</td>
<td>0.998</td>
<td>AFFX-ThrX-M-at</td>
<td>1</td>
<td>HG831-HT831-at</td>
<td>1</td>
</tr>
<tr>
<td>HG831-HT831-at</td>
<td>0.996</td>
<td>HG3342-HT3519-s-at</td>
<td>1</td>
<td>U7379-at</td>
<td>1</td>
</tr>
<tr>
<td>AFFX-ThrX-M-at</td>
<td>0.986</td>
<td>HG831-HT831-at</td>
<td>1</td>
<td>V00594-at</td>
<td>1</td>
</tr>
<tr>
<td>X69111-at</td>
<td>0.984</td>
<td>U14394-at</td>
<td>1</td>
<td>U14394-at</td>
<td>0.847</td>
</tr>
<tr>
<td>U14394-at</td>
<td>0.974</td>
<td>V00594-at</td>
<td>1</td>
<td>AFFX-ThrX-5-at</td>
<td>0.431</td>
</tr>
<tr>
<td>AFFX-LysX-3-at</td>
<td>0.962</td>
<td>X69111-at</td>
<td>1</td>
<td>AFFX-DapX-5-at</td>
<td>0.245</td>
</tr>
<tr>
<td>V00594-at</td>
<td>0.955</td>
<td>U45285-at</td>
<td>0.944</td>
<td>AFFX-PheX-3-at</td>
<td>0.222</td>
</tr>
<tr>
<td>U45285-at</td>
<td>0.932</td>
<td>AFFX-LysX-3-at</td>
<td>0.917</td>
<td>AFFX-HSAC07/X00351-5-at</td>
<td>0.208</td>
</tr>
<tr>
<td>AB000115-at</td>
<td>0.899</td>
<td>AFFX-HSAC07/X00351-5-at</td>
<td>0.806</td>
<td>AB000115-at</td>
<td>0.167</td>
</tr>
<tr>
<td>AFFX-HSAC07/X00351-5-at</td>
<td>0.866</td>
<td>AB000115-at</td>
<td>0.417</td>
<td>U00954-at</td>
<td>0.167</td>
</tr>
<tr>
<td>U73379-at</td>
<td>0.837</td>
<td>U73379-at</td>
<td>0.13</td>
<td>U45285-at</td>
<td>0.167</td>
</tr>
<tr>
<td>AFFX-DapX-M-at</td>
<td>0.678</td>
<td>V00594-s-at</td>
<td>0.074</td>
<td>U73379-at</td>
<td>0.167</td>
</tr>
<tr>
<td>Y09912-rna1-at</td>
<td>0.67</td>
<td>U73379-at</td>
<td>0.037</td>
<td>U73379-att</td>
<td>0.157</td>
</tr>
<tr>
<td>U73379-at</td>
<td>0.56</td>
<td>AFFX-PheX-5-at</td>
<td>0.028</td>
<td>HG1980-HT2023-at</td>
<td>0.032</td>
</tr>
<tr>
<td>AFFX-DapX-3-at</td>
<td>0.555</td>
<td>U03399-at</td>
<td>0.009</td>
<td>AFFX-PheX-M-at</td>
<td>0.028</td>
</tr>
<tr>
<td>V00594-s-at</td>
<td>0.554</td>
<td>U30998-at</td>
<td>0.028</td>
<td>U30998-at</td>
<td>0.028</td>
</tr>
<tr>
<td>HG1980-HT2023-at</td>
<td>0.483</td>
<td>Y09912-rna1-at</td>
<td>0.028</td>
<td>Y09912-rna1-at</td>
<td>0.028</td>
</tr>
<tr>
<td>HG3044-HT3742-s-at</td>
<td>0.441</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D43636-at</td>
<td>0.399</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L27624-s-at</td>
<td>0.387</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U03399-at</td>
<td>0.378</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S69370-s-at</td>
<td>0.321</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFFX-PheX-5-at</td>
<td>0.315</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 25: The top scoring genes (Affymetrix nomenclature).
Conclusions

1. Multi-criterion data mining can perform robust and flexible gene filtering

2. Non-informative priors can be used to find posterior front probabilities

3. Cross-validation can account for statistical sampling uncertainty when posterior analysis is intractible