Pareto Analysis for Gene Filtering in Microarray Experiments

G. Fleury[◇,], A. Hero[†], S. Yoshida[‡], T. Carter[#], C. Barlow[#] and A. Swaroop[‡]
[◇]Ecole Supérieure d'Electricité, Service des Mesures, 91192 Gif-sur-Yvette, France
[†]Depts. of EECS, BioMedical Eng., and Statistics, University of Michigan, Ann Arbor MI 49109, USA
[‡]Depts. of Ophthalmology and Human Genetics, University of Michigan, Ann Arbor MI 48105, USA
[#]The Salk Institute for Biological Studies, La Jolla CA 92037, USA [*]

ABSTRACT

We introduce a method for detecting strongly monotone evolutionary trends of gene expression from a temporal sequence of microarray data. In this method we perform gene filtering via multi-objective optimization to reveal genes which have the properties of: strong monotonic increase, high end-to-end slope and low slope deviation. Both a global Pareto optimization and a pair-wise local Pareto optimization are investigated. This gene filtering method is illustrated on mouse retinal genes acquired at different points over the lifetimes of a population of mice.

1 Introduction

Microarray analysis of gene expression profiles offers one of the most promising avenues for exploring genetic factors underlying disease, regulatory pathways controlling cell function, organogenesis and development [6, 4, 5]. Oligonucleotide-based microarrays allow researchers to accurately quantify the expression level of RNAs of thousands of genes in a tissue sample, thereby providing valuable information about complex gene expression patterns [7]. However, the massive scale and variability of such microarray expression data creates new and challenging problems of clustering and data mining: the so-called *gene filtering* problem.

This paper is an extension of a robust and flexible approach to gene filtering presented in [3]. We called this approach Pareto gene filtering which was based on optimizing two criteria for discovering monotonic gene trajectories. Here we will extend this analysis to three criteria. A more stringent gene filter can be designed by appropriately supplementing the former technique with additional filtering criteria. We compare the global Pareto fronts to the locally optimal pairwise Pareto fronts. The criteria, applied in pairs, give sets of Pareto fronts which can be combined by intersection. This strongly reduces the number of candidate genes which must be evaluated by RT-PCR analysis techniques.

The outline of the paper is as follows. In Sec. 2 a brief overview of microarrays is given. In Sec. 3 we describe the new gene evolution clustering algorithm and in Sec. 5 we apply it to analysis of a sequence of Affymetrix microarrays of mouse retina and we experimentally validate our analysis using real time RT-PCR techniques.

2 GeneChip Microarrays

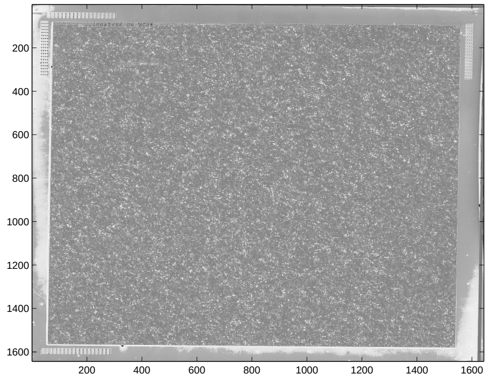

Figure 1: *Affymetrix GeneChip image.*

While the methods described herein are applicable to general genetic expression data, we focus here on analysis of the Affymetrix GeneChip oligonucleotide array. The GeneChip contains several thousand single stranded DNA oligonucleotide probe pairs, which are each 25 bases long and correspond to target genes of interest [6].

Each probe pair consists of an element containing oligonucleotides that perfectly match the target (PM probe) and an element containing oligonucleotides with a single base mismatch (MM probe). During hybridization the labeled RNA of interest binds the probe

[*]This research was partially supported by a NATO grant, supporting G.F.'s sabbatical at the University of Michigan during the summer of 2001, University of Michigan institutional funds and grants from the National Institute of Health (EY11115 (supplement), EY07961, EY07003 (core)), the Macula Vision Research Foundation, the Foundation Fighting Blindness, and Research to Prevent Blindness (RPB). A.S. is recipient of a Lew R. Wasserman Merit Award from RPB.

pair, and the level of binding to each element is determined through electronic scanning of the GeneChip post-hybridization and wash. The expression level of a target RNA is quantified by determining the difference between the PM and MM probes, and averaging this difference for all sixteen probe pairs that represent a given gene (avgdiff, or average difference). Affymetrix software is used to extract intensity information from the GeneChip image (see Fig. 1), and this data is summarized in the form of a spreadsheet with numbers, e.g. call, average difference and log average, indicating absence or presence of a strong hybridization and level of hybridization for each probe. As with any technology taking many thousands of measurements, even a low level of variability can result in many false positives or negatives, therefore replications of the experiment are required to minimize such variability.

The aging experiments described below consist of $M = 4$ samples in each of $K = 6$ different mouse populations. Each population corresponds to a different time point ranging from postnatal day 1-10 to 21 months of age. For each time point M different GeneChip microarrays were processed each containing over $N = 12,000$ probes. The objective is gene filtering: to detect and cluster interesting patterns of gene expression indicative of evolution of the gene over the K time points.

3 Filtering Genetic Signals

For the n-th probe, $n \in \{1, ..., N\}$ of m-th the mouse, $m \in \{1, ..., M\}$, sampled at the k-th time point, $k \in \{1, ..., K\}$ we define the GeneChip avgdiff response $y_n^m(k)$. When looking for genes which have significant non-constant trajectories it is natural to cluster genes based on two criteria: small population variability at each time point (intra-class dispersion) and large variability between populations at different time points (inter-class dispersion). Two natural measures of intra-class dispersion and inter-class dispersion are the (unnormalized) sample deviation of the n-th gene at time sample k

$$\xi_n^1(k) = \sum_{i \neq j} \|y_n^i(k) - y_n^j(k)\|, \tag{1}$$

and the sample deviation between the n-th gene at time samples $k1$ and $k2$

$$\xi_n^2(k1, k2) = \sum_{i,j} \|y_n^i(k1) - y_n^j(k2)\|, \tag{2}$$

where $\| \bullet \|$ denotes a norm, e.g. l_1, l_2 or l_∞. A simple test, analogous to the paired T-test [2], to separate the two time samples could be based on thresholding the ratio of the two dispersion measures:

$$T_n(k1, k2) = \frac{M-1}{2M} \frac{\xi_n^2(k1, k2)}{\xi_n^1(k1) + \xi_n^1(k2)} > \mathcal{T}^{-1}(1-\alpha), \tag{3}$$

where $\mathcal{T}^{-1}(1-\alpha)$ is a threshold chosen to ensure level of significance $\alpha \in [0, 1]$. Figure 2 shows boundaries of the critical region in the $\xi^1 \times \xi^2$ plane specified by (3) for the mouse gene microarray experiment described in Sec. **??**. These boundaries are straight lines corresponding to thresholding (3) at the respective levels of significance.

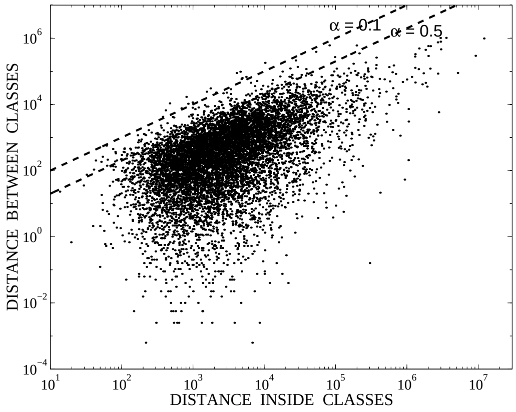

Figure 2: *Scatter plot of inter-class and intra class dispersion criteria (1) and (2) for 8826 mouse retina genes. Superimposed are T-test boundaries for levels of significance* $\alpha = 50\%$ *and* $\alpha = 10\%$.

4 Pareto Filtering Methods

The principle of multi-criterion optimization is different from scalar criteria for filtering and clustering genes such as the paired t-test (3). Rather than filtering by thresholding a scalar criterion, e.g. the t-test ratio on the left side of (3), multi-criterion filtering captures the intrinsic compromises among the conflicting objectives, e.g. dispersion criteria (1) and (2). Consider Fig. 3.a and suppose that ξ^1 is to be minimized and ξ^2 is to be maximized. Under this criterion it is obvious that gene A is "better" than gene C because both criteria are higher for A than for C. However it is not easy to specify a preference between A, B and D. Multi-objective clustering uses the "non-dominated" property as a way to establish such a preference relation. A and B are said to be non-dominated because a gain on one criterion in going from A to B corresponds to a loss on the other criterion. All the genes which are non-dominated constitute a curve which is called the Pareto front (Fig. 3.b). A second Pareto front is obtained by stripping off points on the first front and computing the Pareto front of the remaining points. Pareto analysis has been adopted for many applications including evolutionary computing and optimization [8, 10]. Figure 4 shows the first three Pareto fronts related to the classical criteria (1 & 2).

Pareto analysis provides a new non-parametric gene filtering method which we have used [3] for detecting

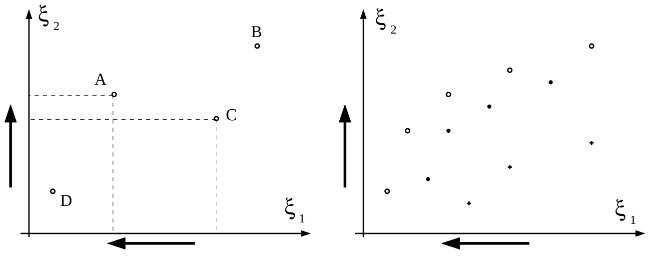

Figure 3: *a). Dominance property, and b). Pareto optimal fronts, in dual criteria plane.*

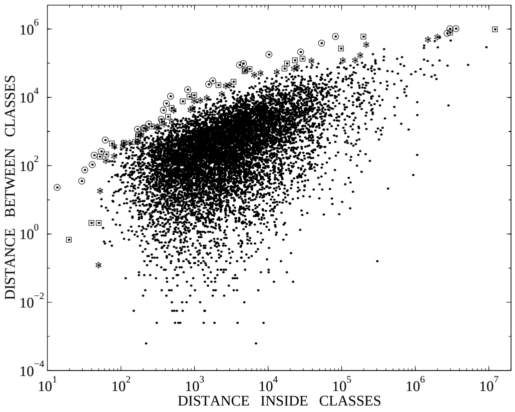

Figure 4: *First (circle) second (square) and third (hexagon) Pareto optimal fronts for same data as shown in Fig. 2.*

genes with specific patterns of temporal evolution. The method was based on joint-maximization of two criteria, namely *monotonicity* ξ^1 (eq. 4) and *end-to-end increase* ξ^2 (eq. 5) of the gene trajectories. The $\overline{y^\star}$ notation denotes the arithmetic average of y^i over i. Since different mice are sacrificed to form each time point, virtual time trajectories must be reconstructed. There are a total of K^M possible virtual trajectories. An example of a typical set of these trajectories is shown in figure 5.

$$\xi^1_n = \frac{1}{K^M}\sum_{i,j,k} \mathrm{sgn}\ \left(y^i_n(k+1) - y^j_n(k)\right), \qquad (4)$$

$$\xi^2_n = \frac{1}{M^2}\sum_{i,j}\left(y^i_n(K) - y^j_n(1)\right) = \overline{y^\star_n(K)} - \overline{y^\star_n(1)}, \qquad (5)$$

After steady monotonic increase, the gene shown on the figure 5 displays a plateau starting at time M2. This can be associated to a development gene as contrasted to an aging gene which are of particular interest to us. For that reason we introduce a third criterion to eliminate development genes from monotonic genes. This third criterion (eq. 6) minimizes the maximal slope difference within the set of trajectories associated with a particular gene.

$$\xi^3_n = \max_{i,j,k}\left(y^i_n(k+1) - y^j_n(k)\right) - \min_{i,j,k}\left(y^i_n(k+1) - y^j_n(k)\right), \qquad (6)$$

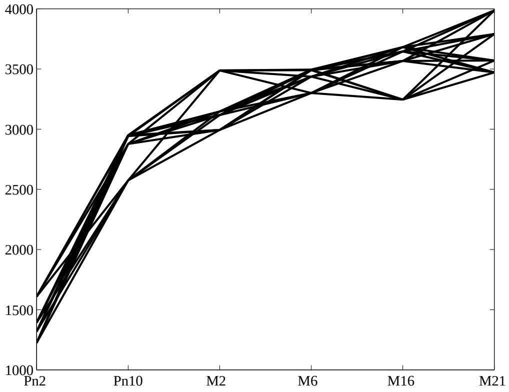

Figure 5: *Typical set of trajectories associated with a particular gene*

With these three criteria we can find the Pareto fronts of interesting genes. The most natural approach to extract these genes is to find a global Pareto front. This front is the set of non-dominated genes relative to all three criteria. An alternative is to find every local pairwise Pareto front and find the intersection. This is a far more stringent selection criterion.

5 Gene filtering application

As in [3] we applied the Pareto analysis described above to classifying patterns in mouse retina. The experiment consists of 6 time samples of retina material taken from a population of 24 mice. 4 mice were selected from the population at 6 different times including 2 early development (Pn2-Pn10) and 4 late development and aging (M2-M21) points. The 24 gene GeneChips were processed by Affymetrix software returning a Unigene-ordered list of 12,422 genes each labeled with Affymetrix attributes such as "call," "avgdiff," and "logavg" [1]. We eliminated from analysis all genes called out as "absent" from all chips, leaving 8826 genes whose expressions were analyzed using the "avgdiff" attribute. The total number of time trajectories for each gene is $6^4 = 4096$.

The figure 6 shows the solutions to the global Pareto optimization, using the three criteria discussed above. The arrow on the graph points in the preferred direction of the three criteria. There are more than one hundred genes on the first Pareto front shown in the figure. The figure 7 shows solutions to the Pareto optimization of pairs (ξ_1, ξ_2), (ξ_1, ξ_3) and (ξ_2, ξ_3). There is only one solution (called the Pareto cross-optimized gene) which lies on all three first Pareto fronts.

Quantitative real time PCR has been employed to independently validate this cross-optimized gene. RT-PCR analysis is highly accurate procedure for single

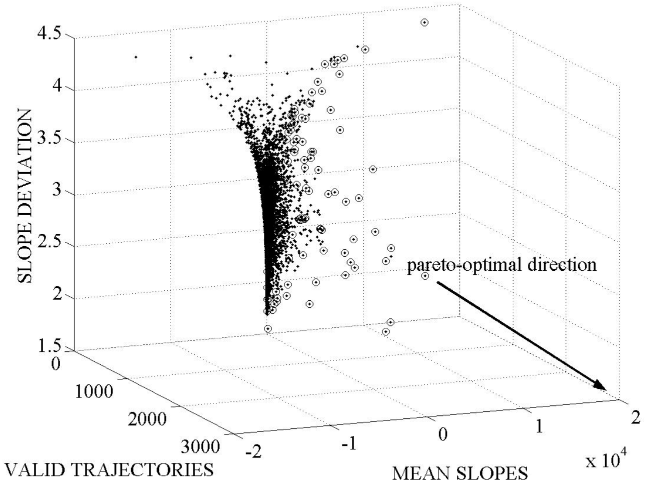

Figure 6: *First global Pareto front (o) for the three criteria (ξ_1, ξ_2 and ξ_3).*

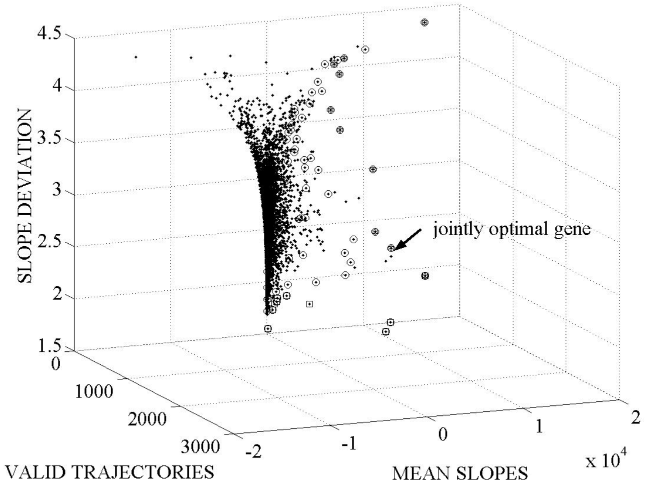

Figure 7: *First Pareto fronts for each pair of criteria taken from the set (ξ_1, ξ_2 and ξ_3). Each one of this front is denoted by squares, circles and stars, respectively.*

gene analysis. Oligonucleotide primers for exons of selected genes were designed to amplify PCR products of about 300 bp. The SYBR Green I dye which is a highly specific double-stranded DNA binding dye was used on real time quantitation. Detailed analysis and interpretation of this and other genes will be reported elsewhere.

6 Conclusion

We have introduced a Pareto method for gene filtering based on three criteria. Both globally optimized and pair-wise cross-optimized procedures have been used to filter "significant" sets of genes in a microarray experiment. The pair-wise cross-optimized procedure is more stringent, exposing a single significant gene among over a hundred globally optimal Pareto genes. Thus this pair-wise optimization procedure is a method which can zero-in on the most interesting genes in a large number of candidate genes. Cross-validation can be applied as discussed in [3] for testing the robustness of the procedure. This approach can be directly generalized to more than three criteria. Many signal processing challenges remain due to the increasingly high dimensionality of genetic data sets. It will be important to develop fast and high-throughput implementations of multi-objective gene clustering and filtering.

References

[1] Affymetrix. *NetAffx User's Guide*, 2000. `http://www.netaffx.com/site/sitemap.jsp`.

[2] P. J. Bickel and K. A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, San Francisco, 1977.

[3] G. Fleury, A. Hero, S. Yoshida, T. Carter, C. Barlow and A. Swaroop, "Clustering Gene Expression Signals from Retinal Microarray Data," *ICASSP'02*, to appear, May 2002.

[4] C. Lee, R. Klopp, R. Weindruch, and T. Prolla, "Gene expression profile of aging and its retardation by caloric restriction," *Science*, vol. 285, no. 5432, pp. 1390–1393, Aug 27 1999.

[5] F. Livesey, T. Furukawa, M. Steffen, G. Church, and C. Cepko, "Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene," *Crx. Curr Biol*, vol. 6, no. 10, pp. 301–10, Mar 23 2000.

[6] D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, E.L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nat. Biotechnol.*, vol. 14, no. 13, pp. 1675-80, Dec. 1996.

[7] D.J. Lockhart and E.A. Winzeler, Genomics, gene expression and DNA arrays, vol. 405, no. 6788, pp. 827-36, *Nature*, Jun 15 2000.

[8] R. E. Steuer, *Multi criteria optimization: theory, computation, and application*, Wiley, New York N.Y., 1986.

[9] S. Yoshida *etal*, , manuscript in preparation.

[10] E. Zitler and L. Thiele, "An evolutionary algorithm for multiobjective optimization: the strength Pareto approach," Technical report, Swiss Federal Institute of Technology (ETH), May 1998.