CLUSTERING GENE EXPRESSION SIGNALS FROM RETINAL MICROARRAY DATA

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 robust method for detecting evolutionary trends of gene expression from a temporal sequence of microarray data. In this method we perform gene clustering via multi-objective optimization to reveal genes with interesting and statistically significant temporal patterns. We illustrate this gene filtering methodology in the context of exploring the time trjaectories of mouse retinal genes acquired at different points over the lifetimes of a population of mice. For 6 time points sampled over 24 mouse subjects, our method can reliably reveal genes whose expression level increases or decreases monotonically, hits a peak or valley at birth, or exhibits other trends.

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 [5, 3, 4]. 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 [6]. 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 describes a robust and flexible approach to gene filtering and analysis for the purpose of detecting and validating temporal gene expression patterns from a series of microarray experiments. We call our approach *Pareto gene filtering* which is based on a novel multicriterion optimization and cross-validation clustering strategy. We apply this method to classifying gene trajectories in mouse retinal aging experiments.

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. 4 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

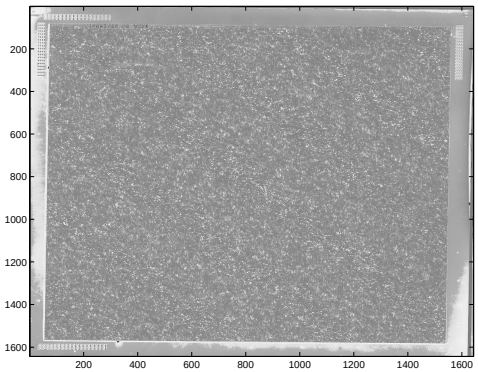

Fig. 1. *Affymetrix GeneChip image.*

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 [5].

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 pair, and the level of binding to each element is determined through electronic scanning of

This research was partially supported by a NATO grant, supporting G.F.'s sabatical 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 Foudation Fighting Blindness, and Research to Prevent Blindness (RPB). A.S. is recipient of a Lew R. Wasserman Merit Award from RPB.

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. CLUSTERING OF 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 (un-normalized) 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. 4. These boundaries are straight lines corresponding to thresholding (3) at the respective levels of significance.

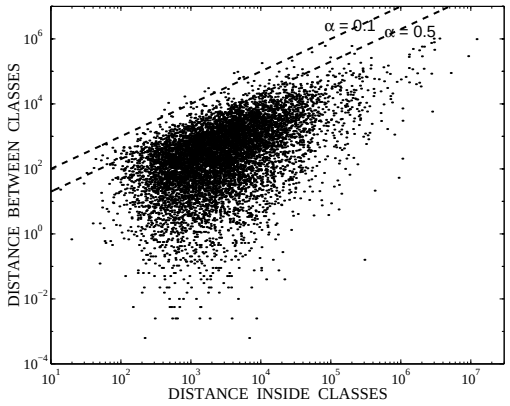

Fig. 2. *Scatter plot of inter-class and intra class dispersion criteria (1) and (2) for 8826 mouse retina genes. Superimposed are T-test boundaries for levels of significance $\alpha = 50\%$ and $\alpha = 10\%$.*

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 [7, 9]. 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 will use for detecting genes with

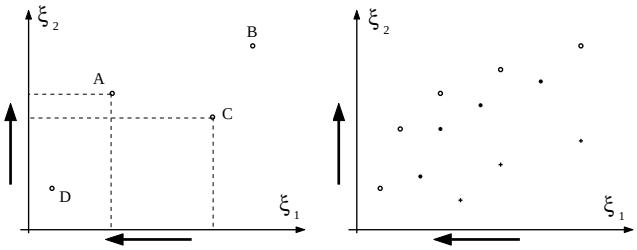

Fig. 3. *a). Dominance property, and b). Pareto optimal fronts, in dual criteria plane.*

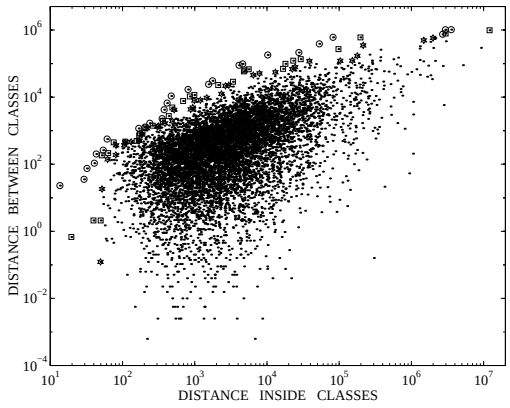

Fig. 4. *First (circle) second (square) and third (hexagon) Pareto optimal fronts for same data as shown in Fig. 2.*

specific patterns of temporal evolution. First a set of K^M time trajectories are defined for each gene corresponding to all possible time paths through one of M replicates at each K time point. For each trajectory we extract the sign of the slope between each time point to capture instantaneous increase or decrease of each gene trajectory. The set of K^M sign trajectories (K-dimensional vectors of signs) summarizes the *monotonicity* of a gene's evolution pattern. For each gene two criteria are then computed: the first is the proportion of the K^M trajectories satisfying a specific evolution pattern, e.g. response monotonicity over time; the second criterion is a measure of the *strength* of the evolution pattern, e.g. the gene response difference between first and last time points. The Pareto fronts are then computed and from these are extracted a list of "significant" genes displaying the pattern of interest.

4. EXPERIMENTAL RESULTS

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 $6^4 = 4096$ and 4 representative trajectories are shown in Fig. 6) for a specific gene. Figure 5 shows the first criterion (the number of trajectories among the 4096 which monotonically increase) as a function of gene number.

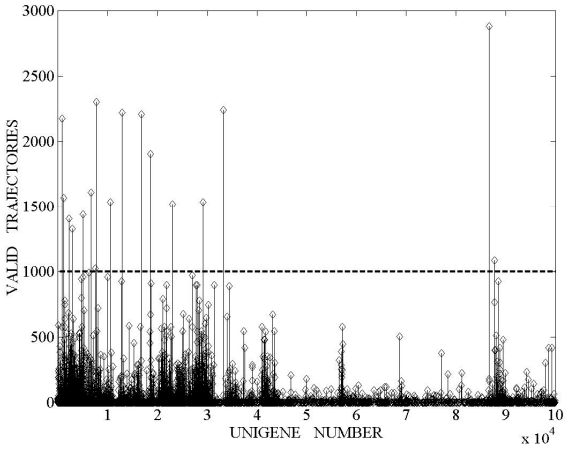

Fig. 5. *Occurrence histogram with threshold.*

The most monotonic gene (2880 trajectories out of 4096) is the gene shown in Fig. 6. This gene has been identified as a gene related to the immune-mediated process which may be associated with the aging process in retina [8].

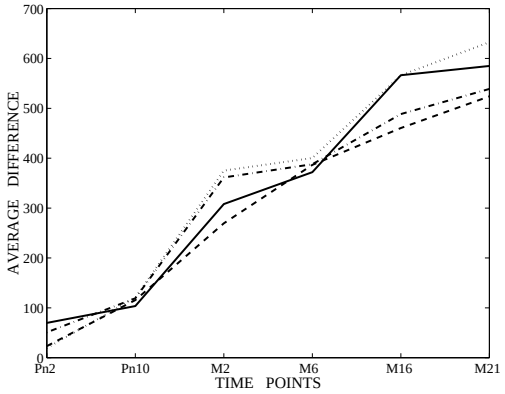

Fig. 6. *Most monotonic gene trajectory*

Choosing a significance threshold of 1000 trajectories upon 4096 is quite conservative, roughly corresponding to 100 sigma from the "random pattern" baseline. Note that the trajectories are statistically dependent, as they are based on reusing identical mice in several different trajectories. The Pareto fronts of the dual monotonicity and slope criteria were then computed and are shown in Fig. 7. The first three Pareto fronts contain 39 genes.

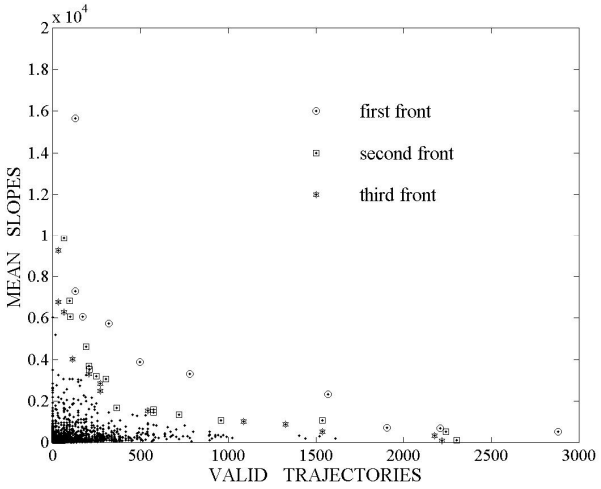

Fig. 7. *First three Pareto fronts for the double criterion (horizontal: number of valid trajectories satisfying the increasing criterion - vertical: slope between first and last time point).*

As the genetic data are strongly corrupted by measurement and other sources of variation, we applied a simple leave-one-out cross-validation procedure to the Pareto analysis. For each time point a mouse was omitted leaving 4096 sets of 729 trajectories to be tested. For each set of trajectories the first three Pareto fronts were computed. Eleven "resistant" genes remained in the first three fronts for all the 4096 tested trajectories. Among these resistant genes only 6 of the initial 39 genes survived the cross validation. Quantitative real time PCR has been employed to independently validate these 11 monotonic Pareto-resistant genes. RT-PCR analysis is highly accurate procedure for single 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. As of this writing, the RT-PCR anlaysis has confirmed the behavior of all the Pareto-resistant genes studied. Detailed analysis and interpretation of these genes will be reported elsewhere.

5. CONCLUSION

DNA microarray technology allows one to evaluate the expression profile of thousands of genes simultaneously. However, to take full advantage of these powerful tools, we need to find new methods to handle large amounts of data and information without becoming overwhelmed by the potentially large number of candidate genes. This paper shows that our new method of Pareto gene filtering can identify genes exhibiting interesting profiles. Additional genes discovered using this algorithm are now being cross-validated by other methods and the data obtained will be utilized to further refine the algorithms and analysis. 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.

6. 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] 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.

[4] 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.

[5] 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.

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

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

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

[9] 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.