Gene filtering and data mining for gene microarray experiments

A. Hero[†], G. Fleury[⋄]

[†]Depts. of EECS, BioMedical Eng., and Statistics, University of Michigan, Ann Arbor MI 49109, USA

[⋄]Ecole Supérieure d'Electricité, Service des Mesures, 91192 Gif-sur-Yvette, France

Abstract

The massive scale and variability of microarray gene data creates new and challenging problems of signal extraction, gene clustering, and data mining, especially for temporal studies. Most data mining methods for finding interesting gene expression patterns are based on thresholding single discriminants, e.g. the ratio of between-class to within-class variation or correlation to a template. Here a different approach is introduced for extracting information from gene microarrays. The approach is based on multiple objective optimization and we call it Pareto front (PF) analysis. This method establishes a ranking of genes according to estimated probabilities that each gene is Pareto-optimal, i.e., that it lies on the Pareto front of the multiple objective scattergram. For illustration the analysis will be illustrated in the context of ranking the most aberrant non-linear genes in Fred Wright's GeneChip study.

I. Introduction

One of the principal challenges in microarray analysis is to reliably extract genes exhibiting interesting expression profiles from the thousands of hybridization indices generated by the microarray. This is the so-called *gene filtering problem*, also called gene screening and gene selection, The most common approach to gene filtering are significance tests implemented by thresholding a set of test statistics, e.g. paired T-tests of mean differences, Fisher tests of variance, or Mann-Whitney rank tests. These can be found on most of the commercial and freeware packages used for statistical gene analysis such as the SAM MS Excel add-on [1], the Affymetrix Microarray Suite and Affymetric Data Mining Tool (DMT) [4], and others [3]. Such approaches can yield a list of genes that are ranked in order of statistical significance according to observed p-values. Methods of multiple comparisons are applicable to more than one filtering criterion [19] but use of such methods does not appear to be widespread in gene analysis.

This paper describes a different approach to gene selection, denoted Pareto-optimal filtering, that is based on the general theory of multiple objective optimization [21] to which the economist Vilfreda Pareto (1848-1923) made seminal contributions (see the website [2]). To apply this method the experimenter computes a number of ranking criteria for each gene, generating a point cloud of criteria vectors which is called the *multicriterion scattergram*. For example, to select the most monotonic profiles over time the ranking criteria might be chosen as the differences in gene expression level over successive time points. The objective of Pareto-optimal filtering is to isolate genes that achieve a compromise between maximizing (or minimizing) the competing gene-ranking criteria, i.e. to find the "winning" profiles. Such genes lie on the so-called *Pareto front* of the multicriterion scattergram and are the *non-dominated genes*, see Sec. III for definitions. Stripping off genes from successive Pareto fronts in the multicriterion scattergram yields a sequence of Pareto fronts at increasing depths in the data, called the first, second, third, ..., Pareto fronts, respectively. This sequence of fronts reveals a hierarchy of the highest scoring gene profiles. Some of the techniques described here have been applied to mouse retina studies [12], [13] and have been validated for mouse retina microarray experiments using RT-PCR techniques. The purpose of the present paper is to present the general Pareto filtering methodology, introduce a Bayesian formulation of Pareto filtering, and to illustrate by applying them to a widely available data set created expressly for testing gene filtering, classification, and differential expression estimation techniques [18].

As the gene indices are randomly sampled from multiple subjects there can exist substantial statistical sampling errors that complicate the Pareto-optimal analysis. These sampling errors can be handled by cross-validation, as in [12], [13], producing what can be called a *resistant Pareto front* (RPF) of genes, defined as those genes that land on the Pareto front with high relative frequency under resampling. As the RPF method does not rely on any distributional assumptions on the data it is very flexible, allowing treatment of arbitrary ranking criteria such as dependent and non-linear functions of the data. However, when the data distribution can be characterized, even approximately, RPF has obvious drawbacks. Principal among these drawbacks is the high computational load of cross-validation which can make RPF methods impractical to implement for large sample size. To address these drawbacks a Bayesian approach is presented for Pareto-optimal gene filtering: the *posterior Pareto front* (PPF) method.As contrasted to the RPF method, the PPF method ranks each gene according to its posterior probability that it belongs to the Pareto front. This probability is computed using prior densities on various unknown parameters in the sampling error distribution.

II. Gene Filtering in Microarrays

The ability to perform accurate genetic differentiation between two or more biological populations is a problem of great interest to geneticists and other researchers. For example, in a temporally sampled population of mice one is frequently interested in identifying genes that have interesting patterns of gene expression over time, called a gene expression profile. Gene microarrays have revolutionized the field of experimental genetics by offering to the experimenter the ability to simultaneously measure thousands of gene sequences simultaneously. A gene microarray consists of a large number N of known DNA probe sequences that are put in distinct locations on a slide. See one of the following refernences for more details [16], [9], [7], [11]. After hybridization of an unknown tissue sample to the gene microarrays, the abundance of each probe present in the sample can be estimated from the measured levels of hybridization (responses). Two main types of gene microarrays are in wide use: photo-lithographic gene chips and spotted fluorescence gene arrays. An example of the former is the Affymetrix product line [5]. An example of the later is the cDNA microarray protocol [10].

The study of differential gene expression between T populations requires hybridizing several (M) samples from each population to reduce response variability. Define the measured response at the n-th gene chip probe location for the m-th sample at time t

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

When several gene chip experiments are performed over time they can be combined in order to find genes with interesting expression profiles. This is a data mining problem for which many methods have been proposed including: multiple paired t-tests; linear discriminant analysis; self organizing (Kohonen) maps (SOM); principal components analysis (PCA); K-means clustering; hierarchical clustering (kdb trees, CART, gene shaving); and support vector machines (SVM) [14], [6], [8]. Validation methods have been widely used and include: significance analysis of microarrays (SAM); bootstrapping cluster analysis; and leave-one-out cross-validation [22], [17]. Most of these methods are based on filtering out profiles that maximize some criterion such as: the ratio of between-population-variation to within-population-variation; or the temporal correlation between a measured profile and a profile template.

III. Multiple Objective Gene Filtering

As contrasted to maximizing *scalar* criteria, multiple objective gene filtering seeks gene profiles that strike an optimal compromise between maximizing several criteria [12]. This is closely related to multiple objective optimization [21] in which the concept of Pareto-optimal solutions play a crucial role. These solutions are almost never unique and are variously called the Pareto-optimal set, the Pareto front, the Pareto frontier, and the Edgeworth-Pareto front [20]. Pareto optimality theory has been applied to a wide range of application areas including: economics, mathematical psychology, operations research, and evolutionary computing [24], [21].

Multiple objective optimization captures the intrinsic compromises among conflicting objectives. Consider Fig. 1 and suppose that ranking criteria ξ_1 and ξ_2 are to be maximized. The collection of points in the figure are called the multicriterion scattergram. It is obvious that genes A, B and C are "better" than genes D and E because both criteria are higher for the former than for the latter. Note that no gene among A, B and C dominates the other in both criteria ξ_1 and ξ_2. Multi-objective filtering uses this "non-dominated" property as a way to establish a preference relation among genes A, B, C, D and E. More formally, gene i is said to be dominated if there exists some other gene $g \neq i$ such that for some $p = p_o$

$$\xi_p(i) < \xi_{p_o}(g) \text{ and } \xi_p(i) \leq \xi_p(g),\ p \neq p_o.$$

The set of non-dominated genes are defined as those genes that are not dominated. All the genes which are non-dominated constitute a curve which is called the (first) Pareto front. A second Pareto front can obtained by stripping off points on the first front and computing the Pareto front of the remaining points - which for the example in Fig. 1 would be genes D and E.

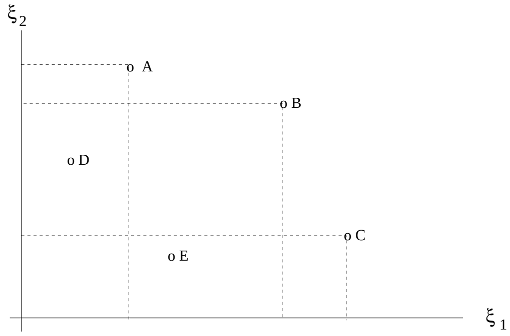

Fig. 1. *A hypothetical multicriterion scattergram for genes A,B,C,D,E plotted as vectors in the plane described by a pair of ranking criteria ξ_1 and ξ_2. A, B, C are non-dominated genes and form the (first) Pareto front. A second Pareto front is formed by genes D,E.*

The above methods are applicable when the criteria ξ_1 through ξ_P are perfectly observable. However, as these criteria depend on the true mean values $\underline{\mu}(i)$ of the i-th gene profile, the criteria are only partially observed through a random sample from the underlying population. A natural way around this is to apply Pareto front analysis to the sample means but this does not deal with statistical uncertainty due to random sampling errors. The way we propose to deal with this is to compute the probability that a given gene lies on the first Pareto front given the measurements. Two methods have been developed to compute this probability: 1) resampling and cross-validation and 2) computation of Bayes posterior probabilities under non-informative priors on the means and variances of the samples. Due to lack of space these methods are discussed elsewhere [15], [12].

To account for sampling uncertainty we assume the additive model for the gene profile measurement

$$y_{mt}(i) = \mu_t(i) + \epsilon_{mt}(i)$$

where $\epsilon_{mt}(i)$ are zero mean noise samples and $m = 1, \ldots, M$, $t = 1, \ldots, T$ and $i = 1, \ldots, N$. Let the vector criterion $\underline{\xi}(i) = [\xi_1(i), \ldots, \xi_P(i)]^T$ be defined as a linear function of the mean profile vector:

$$\underline{\xi}(i) = A\underline{\mu}(i),$$

where $A = ((a_{ij}))$ is a $P \times T$ *contrast matrix.* The vector $\underline{\xi}(i)$ will be called the *profile contrasts* for gene i. The profile contrast matrix must satisfy some simple properties, such as orthogonality and positivity, to compute the posterior Pareto probabilities $\{p(i|Y)\}_i$ that gene i belongs to the first Pareto front. Of course no such assumptions are necessary for resampling and cross-validation methods, described below, which consist of computing the relative frequency, also denoted $p(i|Y)$, that gene i belongs to the first Pareto front as data is resampled over subsets of the samples.

IV. Application to Fred Wright's Data

We applied our methods to Fred Wright's dataset described in the paper [18] and available at the web address provided in the citation. The analysis software was written and implemented in Matlab. Fred Wright's data set was obtained from a mixing experiment which the authors designed for empirically validating and comparing various differential gene expression methods of analysis. As explained in [18] three populations of genes were hybridized to Affymetrix HuGeneFL chips: serum starved human fibroblast cells; serum stimulated human fibroblast cells; and a 50-50 mixture of these cells. A total of 18 chips were processed corresponding to 6 replications within each of the three populations mentioned above. Each HuGeneFL chip contains the same 7129 gene probes. For each gene probe the sequence of hybridization levels from the "stimulated(t=1)," "50-50(t=2)," and "starved(t=3)," populations was defined, in that order, as a gene expression profile. This provides a suitable test dataset since the true profiles should be linearly increasing or decreasing over the three "time points." Any extracted non-monotone gene profiles must either be due to statistical estimation errors, uncontrolled fluctuations in sample concentrations during hybridization, or other experimental errors.

The objective is to determine the most aberrant inverted V-shaped gene profiles. As a preprocessing step a standard non-linear profile filter was applied using a Fisher test to screen gene profiles having large residual linear regression errors inconsistent with a linearity hypothesis. This preprocessing eliminated all but 98 genes from the 7129 total number of genes studied. In the sequel these will be called the "non-linear" gene profiles. A simple modification of the sign-based Pareto analysis method used in [23], [13], [12] can be applied to finding the most aberrant non-linear profiles. In Fig. 2 the multicriterion scattergram is displayed for Li-Wong indices downloaded from Fred Wright's website. The non-linear genes are displayed with crosses. The first criterion in the figure is the contrast defined by $A = [-1, 2, -1]$, which measures twice the difference between the middle point and the average of the two other points in each profile. The second criterion is the number of "virtual" profiles whose shapes match a convex cap profile. Specifically, for each gene generate all $6^3 = 216$ possible trajectories through the 3 sets of 6 replicated measurements of hybridization levels. The ranking is defined from the proportions of these trajectories which have slope of positive sign followed by slope of negative sign. This ranking procedure will be called the "non-parametric" resistant pareto front (RPF) method since the sign-based shape criterion does not depend on the sharpness or assymetry of the inverted V profile shape. Figure 3 shows the first five Pareto fronts computed on the full set of 3×6 non-linear gene samples indicated as crosses on Figure 2. These fronts were computed by successively stripping off genes found to lie on the previous Pareto fronts and rerunning Pareto analysis on the remaining points. Finally leave-one-out cross validation was performed to determine the resistant genes that for which a high proportion of the 216 resampled 3×5 trajectories remained on the first Pareto front. Fig. 4 shows the top 8 resistant profiles ranked in terms of relative frequency of remaining on the first front.

Linear contrast criteria on the non-linear gene profiles were also implemented to determine a ranking of the most aberrant inverted V-shaped profiles. For this the following contrast matrix is adopted

$$A = \begin{bmatrix} -1 & 1 & 0 \\ 1 & 1 & -2 \end{bmatrix},$$

which takes large values for the inverted-V shaped profiles. Figure 5 displays the multicriterion scattergram.

The crosses in the figure indicate the 98 non-linear genes. Figure 6 shows the first five Pareto fronts computed on the full data set without any cross-validation. Figures 7 and 9 show the results RPF and investigated Bayesian posterior Pareto front (PPF) analysis. These plots illustrate how statistical uncertainty in the multiple criteria plane (standard error contours) translates to probability that a gene lies on the first Pareto front. Figures 8 and 10 show the eight top scoring trajectories under PPF and cross-validated RPF analysis, respectively. In each sub-panel the indicated piecewise linear line passes through the means of the 6 replicates for each of the 3 time samples.

V. Conclusion

This paper introduced a new method of gene filtering based on analysis of the Pareto fronts of a specified multiple criterion objective function applied to each gene. These techniques also have applicability to general data mining problems involving shape analysis and general selection criteria.

Acknowledgements

The authors are grateful for illuminating discussions of this work with Dr. Anand Swaroop and Dr. Shigeo Yoshida, of the Depts. of Ophthalmology and Human Genetics, with Prof. Debashis Ghosh of the Dept. of Biostatistics at UM Ann Arbor. The authors also thank Prof. Terry Speed, of UC Berkeley, for making us aware of Frank Wright's data. This research was partially supported by a NATO grant that funded Gilles Fleury's sabatical at the University of Michigan during the summer of 2001.

References

[1] *SAM: Significance analysis of microarrays.* Stanford Office of Technology and Licencing, 2001. `http://www-stat.stanford.edu/~tibs/SAM/`.
[2] *Vilfred Pareto, 1848-1923.* History of Economic Thought, New School, 2001. `http://cepa.newschool.edu/het/profiles/pareto.htm`.
[3] *Data Analysis.* Human Genome Mapping Project Resource Center, 2002. `http://www.hgmp.mrc.ac.uk/Research/Microarray/dataanalysis.`
[4] *Genechip software.* Affymetrix, Inc, 2002. `http://www.affymetrix.com/products/software/index.affx`.
[5] Affymetrix. *NetAffx User's Guide*, 2000. `http://www.netaffx.com/site/sitemap.jsp`.
[6] A. A. Allzadeh and etal, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, 2000.
[7] D. Bassett, M. Eisen, and M. Boguski, "Gene expression informatics–it's all in your mine," *Nature Genetics*, vol. 21, no. 1 Suppl, pp. 51–55, Jan 1999.
[8] M. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. Sugent, T. Furey, M. Ares, and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. of Nat. Academy of Sci. (PNAS)*, vol. 97, no. 1, pp. 262–267, 2000.
[9] P. O. Brown and D. Botstein, "Exploring the new world of the genome with DNA microarrays," *Nature Genetics*, vol. 21, no. 1 Suppl, pp. 33–37, Jan 1999.
[10] cDNA Microarray Protocols. *Protocols*, 2001. `http://www.nhgri.nih.gov/DIR/Microarray/protocols.html`.
[11] P. Fitch and B. Sokhansanj, "Genomic engineering: moving beyond DNA sequence to function," *IEEE Proceedings*, vol. 88, no. 12, pp. 1949–1971, Dec 2000.
[12] G. Fleury, A. O. Hero, S. Yoshida, T. Carter, C. Barlow, and A. Swaroop, "Clustering gene expression signals from retinal microarray data," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, Orlando, FL, 2002.
[13] G. Fleury, A. O. Hero, S. Yoshida, T. Carter, C. Barlow, and A. Swaroop, "Pareto analysis for gene filtering in microarray experiments," in *European Sig. Proc. Conf. (EUSIPCO)*, Toulouse, FRANCE, 2002.
[14] T. Hastie, R. Tibshirani, M. Eisen, P. Brown, D. Ross, U. Scherf, J. Weinstein, A. Alizadeh, L. Staudt, and D. Botstein, "Gene shaving: a new class of clustering methods for expression arrays," Technical report, Stanford University, 2000.
[15] A. O. Hero and G. Fleury, "Pareto front analysis for gene filtering," *J. Am. Statist. Assoc.*, p. submitted, 2002.
[16] K. Kadota, R. Miki, H. Bono, K. Shimizu, Y. Okazaki, and Y. Hayashizaki, "Preprocessing implementation for microarray (prim): an efficient method for processing cdna microarray data," *Physiol Genomics*, vol. 4, no. 3, pp. 183–188, Jan 19 2001.
[17] K. Kerr and G. Churchill, "Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments," *Proc. of Nat. Academy of Sci. (PNAS)*, vol. 98, pp. 8961–8965. `citeseer.nj.nec.com/414709.html`.
[18] W. J. Lemon, J. T. Palatini, R. Krahe, and F. A. Wright, "Theoretical and experimental comparison of gene expression estimators for oligonucleotide arrays," *Bioinformatics*, To appear 2002. `http://thinker.med.ohio-state.edu/projects/fbss/index.html`.
[19] R. G. Miller, *Simultaneous Statistical Inference*, Springer-Verlag, NY, 1981.

[20] W. Stadler, *Multicriteria optimization in engineering and the sciences*, chapter Fundamentals of multicriteria optimization, Plenum, New York, 1988.

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

[22] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. of Nat. Academy of Sci. (PNAS)*, vol. 98, pp. 5116–5121, 2001.

[23] S. Yoshida, T. Carter, G. Fleury, S. Hiriyanna, M. Othman, D. Lockhart, A. Hero, C. Barlow, and A. Swaroop, "Altered expression of immune- and stress- response genes in aging retina; implications for aging associated retinopathies," *Nature Genetics*, vol. submitted, , May 2002.

[24] 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.

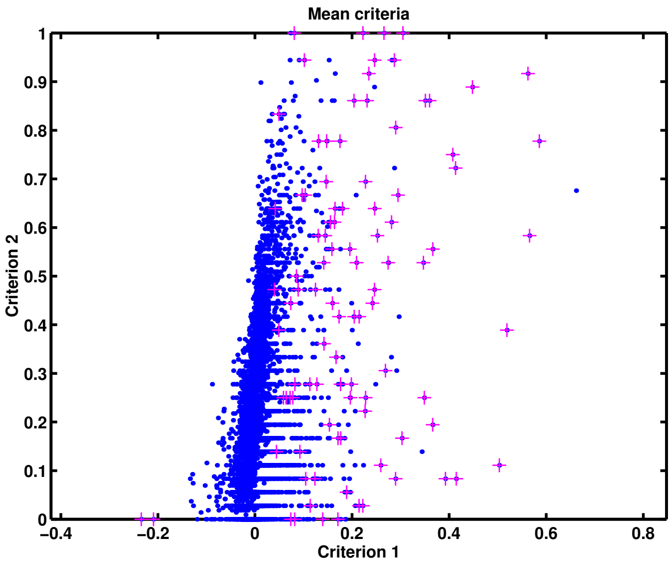

Fig. 2. *Multicriterion mean scattergram of the non-parametric slope-sign ranking criterion for filtering the most aberrant inverted V-shaped gene profiles for Li-Wong reduced indices in Fred Wright's HuGeneFL mixture study. Crosses denote the 98 non-linear genes failing the Fisher linear profile test at a p-value of 0.1.*

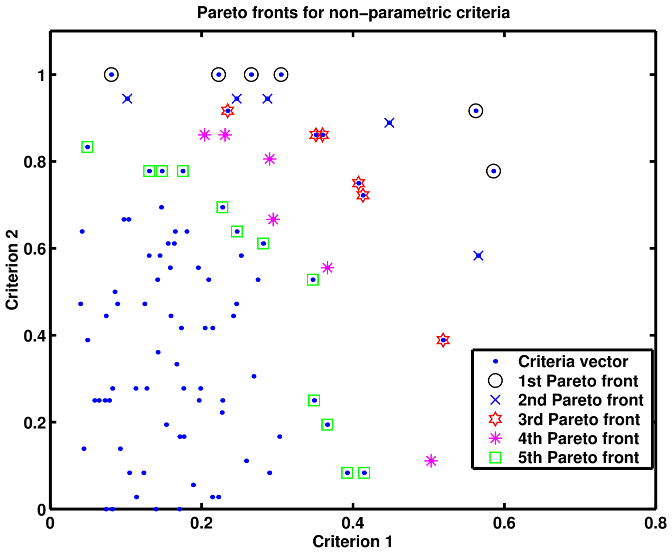

Fig. 3. *The first five Pareto fronts (no cross-validation) of the non-parametric inverted V-shape criteria for the non-linear genes indicated by crosses in Fig. 2.*

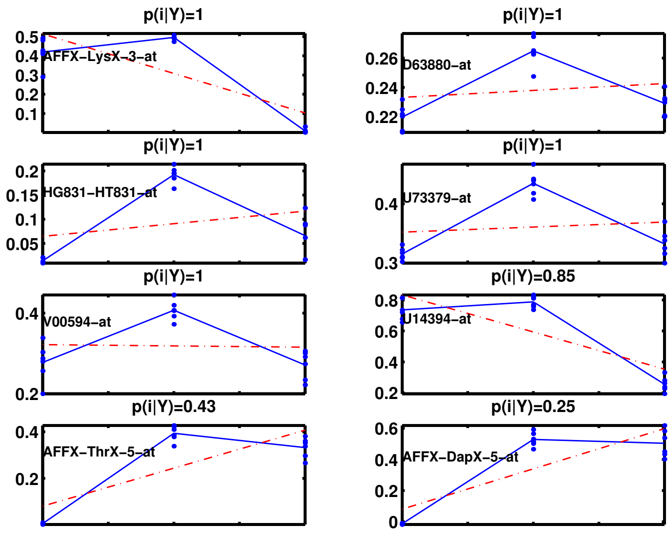

Fig. 4. *The 8 top ranked cross-validated gene profiles remaining on the first Pareto front among the non-linear genes in Fig. 3. $P(i|Y)$ denotes the relative frequency that each resampled (leave-one-out resamping) profile is Pareto-optimal according to the non-parametric slope-sign criteria. Dashed line is the linear regression line.*

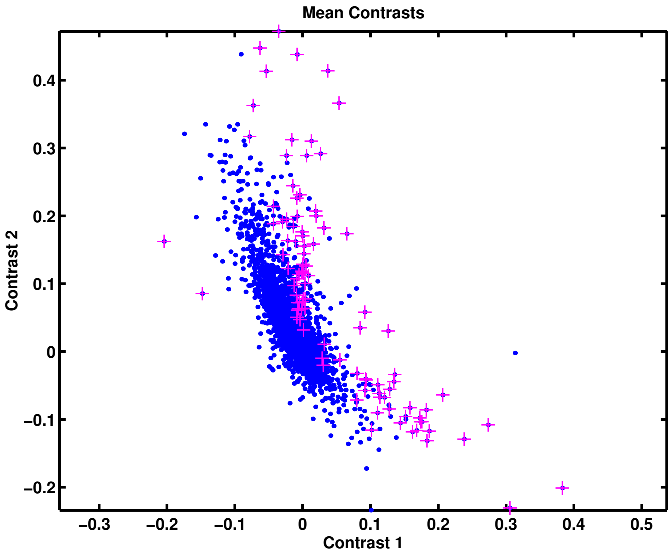

Fig. 5. *Multicriterion scattergram corresponding to contrast matrix* $A = [-1, 1, 0; -1, -1, 2]$. *Crosses again indicate the 98 genes having non-linear profiles at a p-value of 0.1.*

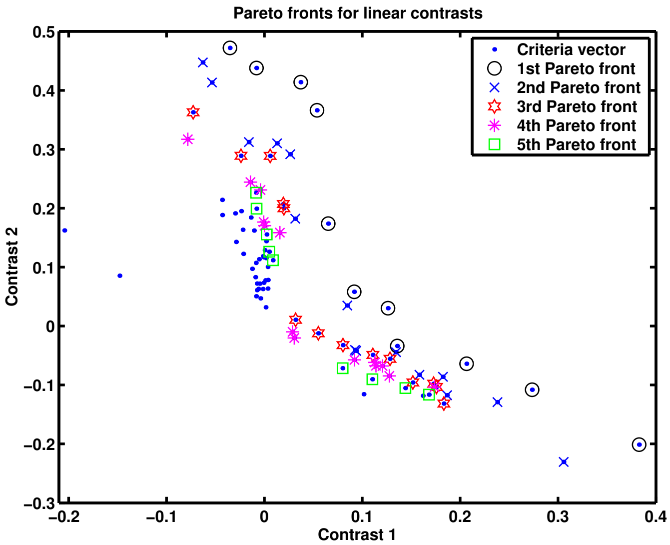

Fig. 6. *The first five Pareto fronts for the genes with non-linear profiles shown in Fig. 5.*

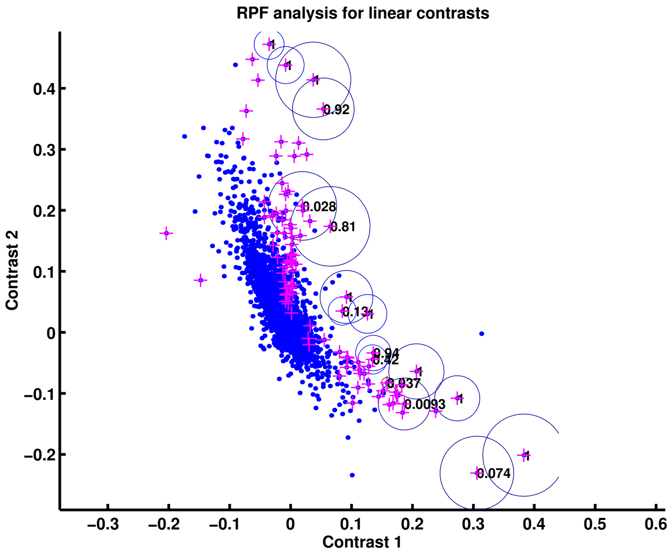

Fig. 7. *17 genes which belong to the first Pareto front with non-zero probability, computed by cross-validation analysis applied to Fig. 6. Constant contours around each point indicate standard errors under equal variance hypothesis and the relative frequencies of lying on the first Pareto front are indicated at the center of relevant contours.*

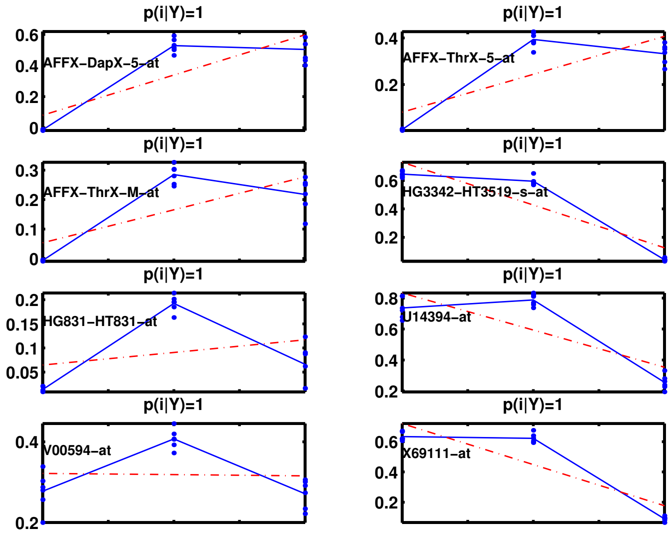

Fig. 8. *The 8 top ranked cross-validated gene profiles remaining on the first Pareto front among the non-linear genes in Fig. 5. $P(i|Y)$ denotes the relative frequency that each resampled (leave-one-out resamping) profile is Pareto-optimal according to the two linear contrast criteria. Dashed line is the linear regression line.*

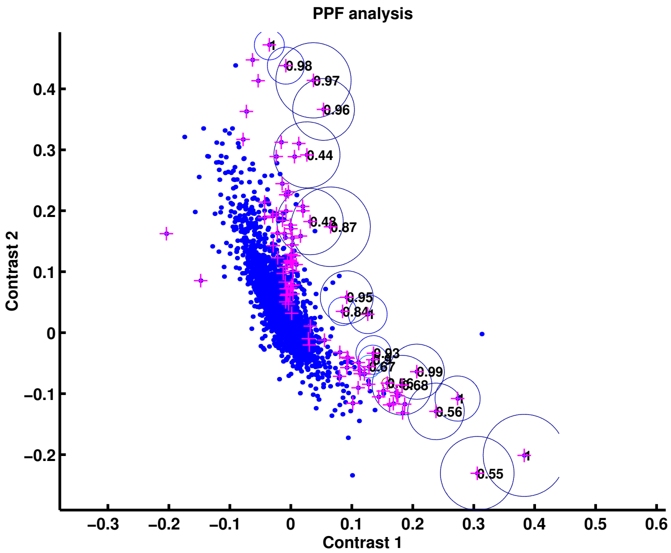

Fig. 9. *Same as in Fig. 7 for the linear contrast PPF method along with standard error constant contours and posterior probabilities of belonging to the first Pareto front. For clarity, only the first 20 top ranking genes are shown.*

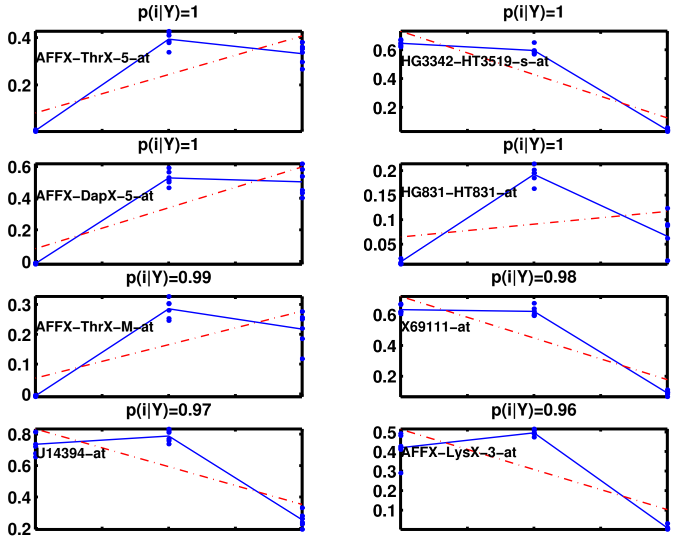

Fig. 10. *Same as Fig. 8 except that gene profile ranking is according to computed PPF posterior probabilities shown on Fig. 9. $P(i|Y)$ denotes the Bayes posterior probability that each profile is Pareto-optimal according to the two linear contrast criteria.*