Gene Selection and Ranking with Microarray Data

Alfred O. Hero
Dept. of Electrical Engineering & Computer Science
University of Michigan
1301 Beal Ave
Ann Arbor, MI

Abstract

Over the past decade there has been an explosion in the amount of genomic data available to biomedical researchers due to advances in biotechnology. For example, using gene microarrays, it is now possible to probe a person's gene expression profile over the more than 30,000 genes of the human genome. Signals extracted from gene microarray experiments can be linked to genetic factors underlying disease, development. and aging in a population. This has greatly accelerated the pace of gene discovery. However, the massive scale and experimental variability of genomic data makes extraction of biologically significant genetic information very challenging. One of the most important problems is to select a ranked list of genes which are both biologically and statistically significant based on a gene microarray experiment. We will describe multicriterion methods that we have developed for this gene selection and ranking problem.

1. INTRODUCTION

Since Watson and Crick discovered DNA more than fifty years ago, the field of genomics has progressed from a speculative science starved for data and computation cycles to one of the most thriving areas of current research and development. It was not until almost 45 years after Watson and Crick's discovery that the first entire genome was sequenced, the E Coli bacterium containing over 4000 genes, after several years of effort. In 2001 the first draft of the human genome, containing more than 30,000 genes, was obtained. In spring 2003 the genome for the SARS corona virus (SARS-CoV) was sequenced and authenticated in less than 2 months time.[1,2] These recent leaps in progress would not have been possible without significant advances in gene sequencing technology. One such technology, which is the main focus of this paper, are gene microarrays and their associated signal extraction and processing algorithms.

Gene microarrays provide a high throughput method to simultaneously probe a large number gene expression levels in a biological sample. Current state-of-the-art microarrays contain up to 50,000 gene probes that interact with the sample producing probe responses that can be measured as a multichannel signal. When the probes are suitably representative of the range of genetic variation of the organism, this signal specifies a unique gene expression signature of the sample. Gene microarrays are a very powerful tool which can be used to perform gene sequencing, gene mapping and gene expression profiling. They will be critical in determining the genetic circuits that regulate expression levels over time and genetic pathways that lead to specific biological function or dysfunction of an organism.

In this paper we will describe some signal processing challenges in gene microarray analysis and present a few approaches we have developed in interacting with our collaborators in molecular biology. The focus application of the paper is the analysis of temporal gene expression profiles and their role in exploring genetic factors underlying disease, regulatory pathways controlling cell function, organogenesis and development. In particular we and our collaborators in the Dept. of Human Genetics at the University of Michigan are interested in analyzing retinal data to determine genetic factors underlying dysfunction of the eye due to aging, glaucoma, macular degeneration, and diabetes. Our examples will be primarily drawn from these areas and we will focus on the problem of selection and ranking of genes that are both biologically and statistically significant from experimentally replicated microarray data.

In our past work on signal processing for gene microarrays[3–7] our primary goal has been be to develop statistically reliable methods for selecting and ranking temporal gene expression profiles. The work most closely related to this paper is our multi-criterion optimization approach to *gene ranking* using a statistical version of Pareto front analysis.[5,7] In this work two methods for ranking data from multiple microarray experiments were introduced: cross-validation leading to resistant Pareto front (RPF) analysis, and Bayes smoothing, leading to posterior Pareto front (PPF) analysis. In this paper we focus on the *gene selection* problem and adopt a statistical multiple criteria approach similar to our previous work. We then illustrate these methods for two Affymetrix GeneChip experiments for probing the genes of the retina. In these experiments we adopt pairs of criteria which trade-off high selectively for robustness. Specifically, one selection criterion is a (multivariate) paired t-test statistic for selecting gene profiles. This criterion has optimal gene selection properties under a Gaussian microarray probe response model. The other criterion is based on distribution-free rank order statistics. This criterion is robust to violations of distributional assumptions on the data.

The outline of the paper is as follows. In Sec. 2 we give some background on genomics and review gene microarrays in the context of temporal profile analysis. In Sec. 3 we motivate and describe the multicriterion selection and ranking approach. In Sec. 4 we apply false discovery rate (FDR) to multicriterion gene selection. Finally, in Sec. 5 we illustrate these techniques for experimental data.

2. GENOMICS BACKGROUND

We start with some definitions and a brief review of molecular biology and genetics. The genome refers to the genetic operating system which controls structure and function of cells in an organism. This genome consists of genes that lie on segments, called exons, of the double stranded DNA helix which lie on a number of chromosomes in the nucleus of every cell in the organism. The number of genes in the DNA of a given organism can range from a few thousand for simple organisms to tens of thousands for more sophisticated organisms. Each exon contains a gene which is encoded as a nucleotide sequence of symbols A,C,G,T forming a 4-ary alphabet.

Gene expression occurs when the DNA sheds certain of its genes in the cell nucleus in order to stimulate or inhibit various functions, e.g., cell growth or metabolism. This stimulation occurs through production of derivatives of DNA, the mRNA and tRNA, produced by a process called transcription and translation. Stimulated by mRNA and tRNA the ribosome of a cell produces specific amino acids in polypetide chains. These chains form proteins that carry out the intended function expressed by the DNA. While the DNA does not change, the specific genes expressed in this fashion can change over time, environmental conditions, and treatments. The objective of genomics is to identify the very large numbers of genes that are expressed by the organism.

Biotechnology, such as gene microarray hybridization, Northern hybridization, and gell electrophoresis, is essential to reliably probe the gene expression of a biological sample. Bioinformatics provides tools for computational extraction and analysis of the vast amounts of information in probe response data. As scientists and genetic engineers become increasingly interested in studies of gene expression profiles over time, signal processing will become a major bioinformatics tool. We next briefly describe the signals generated by gene microarrays.

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 references[8,9] for more details. After hybridization of an unknown tissue sample to the gene microarray, the abundance of each probe present in the sample can be estimated from the measured levels of hybridization. Two main types of gene microarrays are in wide use: photo-lithographic gene chips and fluorescent spotted cDNA arrays. An example of the former is the Affymetrix[10] product line. An example of the later is the cDNA microarray protocol of the National Human Genome Research Institute (NHGRI).[11] A suite of software tools are available from Affymetrix and elsewhere for extracting accurate estimates of abundance, called probe responses. When probe responses are to be compared across different microarray experiments they must also be normalized. Extraction and normalization methods can range from simple unweighted sample averaging, as in the Affymetrix MAS4 software, to more sophisticated model-based analyses, such as MAS5,[10] the Li-Wong method[12,13] and RMA.[14,15] Many of the more sophisticated packages are available as freeware, e.g., see Strimmer's website[16] for links to relevant software written in the R software language. When several microarray experiments are performed over time they can be combined in order to find genes with interesting temporal expression profiles (see Fig. 1). This is a data mining problem known variously as ”gene selection” and ”gene filtering” for which many methods have been proposed.[17–19] Crucial for gene ranking is the specification of a preference ordering for the ranking. A popular gene selection and ranking method is based on optimizing some single fitness criterion such as: the ratio of between-population-variation to within-population-variation; or the

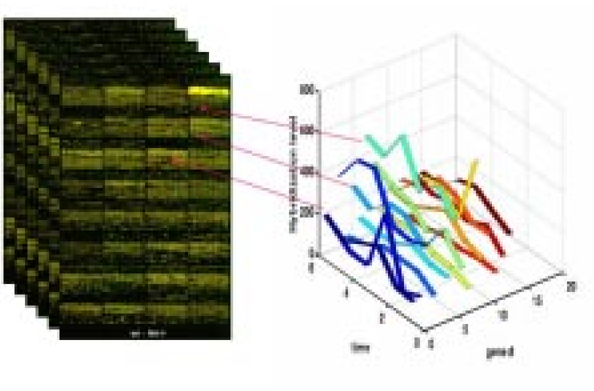

Figure 1. *Probing gene expression at several time points leads to a temporal sequence of gene microarrays (left). A few of the sequences can be extracted at specific probe locations on the microarrays and plotted as time signals (right).*

temporal correlation between a measured profile and a profile template. A problem with this single criterion ranking method is that it is often difficult for the molecular biologist to articulate what he is looking for in terms of a single quantitative criterion. It is for this reason that our group has proposed multiple criteria methods for selecting and ranking gene profiles.[3,5,7]

3. MULTICRITERIA SELECTION AND RANKING

As contrasted to maximizing *scalar* criteria, multiple objective gene filtering seeks gene profiles that strike an optimal compromise between maximizing several criteria. It is often easier for a molecular biologist to specify several criteria than a single criterion. For example the biologist might be interested in aging genes, which he might define as those genes having expression profiles that are increasing over time, have low curvature over time, and whose total increase from initial time to final time is large. As another example, one may have to deal with two biologists who each have different criteria for what features constitute an interesting aging gene.

Multicriterion Gene Selection: To illustrate, let fitness criteria $\xi_1(g), \ldots, \xi_p(g)$ be defined for each gene g in the microarray. A reasonable gene selection criterion would be that the fitness for each selected gene g lies in the quadrant $\xi_1(g) > u_1, \ldots, \xi_p(g) > u_p$. Here $u_1, \ldots, u_p$ are thresholds which are selected by the experimenter to reflect the biological significance of a particular level of measured gene fitness $\xi_k(g)$. This is illustrated in Fig. 2 where the selected sector for two aging criteria (the orthogonalized criteria described in Sec. 5.1) is superimposed over the scatter plot of fitness levels extracted for all the genes probe in the microarray. This scatter plot is called the multicriteria scattergram of the fitness responses.

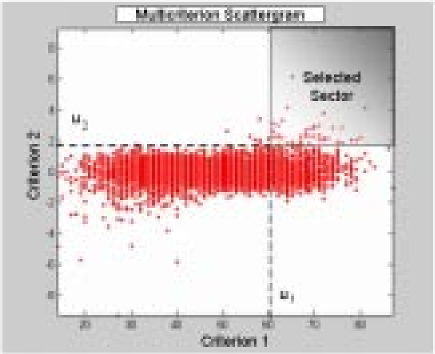

Figure 2. *Multicriteria scattergram of gene fitness responses with overlaid gene selection sector. The choice of position $[u_1, u_2]$ of the sector depends on the experimenter's chosen biological significance levels for gene discovery.*

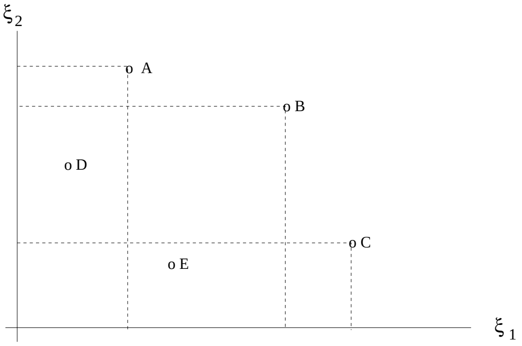

Figure 3. *A hypothetical multicriterion scattergram for genes A,B,C,D,E plotted as vectors in the plane described by a pair of fitness criteria ξ_1 and ξ_2. A, B, C are non-dominated genes and form the (first) Pareto front. A second Pareto front is formed by genes D,E.*

Multicriterion Gene Ranking: In a well designed gene microarray experiment, multicriterion (or other) methods of selection will generally result in a large number of genes and the biologist must next face the problem of selecting a few of most "promising genes" to investigate further. Resolution of this problem is of importance since validation of gene response requires more sensitive techniques, such as RT-PCR, which are much more time consuming and expensive. Some sort of rank ordering of the selected genes would help guide the biologist to a solution. As a linear ordering of set of vector quantities such as $\{[\xi_1(g), \ldots, \xi_p(g)]\}_g$ does not generally exist, an absolute ranking of the selected genes is of course generally impossible. However a partial ordering of these vectors is possible and such a "partial ranking" can be formulated as a multiple objective optimization problem. Multiple objective optimization captures the intrinsic compromises among possibly conflicting objectives in a natural way. Consider the multicriterion scattergram in Fig. 3 and suppose that fitness criteria ξ_1 and ξ_2 are to be maximized. Gene D is *dominated* by both gene A and gene B since gene D has lower fitness in both criteria ξ_1 and ξ_2. Likewise gene E is dominated by gene B and gene C. On the other hand genes A, B and C are not dominated by any other gene and are therefore preferable to genes D and E. Multi-objective filtering uses this non-dominated property as a way to establish a preference relation among genes given a set of criteria $\{\xi_q\}_q$. More formally, gene i is said to be dominated if there exists some other gene $g \neq i$ such that for at least one q

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

The set of non-dominated genes are defined as those genes that are not dominated. All the genes which are non-dominated constitute a set of points called the (first) Pareto front. A second Pareto front can be obtained by stripping off the points on the first front and computing the Pareto front on the remaining points. For the example in Fig. 3 the first Pareto front is $\{A, B, C\}$ and the second Pareto front is $\{D, E\}$.

The above multiple criterion selection and ranking methods are applicable to any set of criteria $\xi_1, \ldots, \xi_p$. However, these method do not account for any statistical uncertainty. The study of gene expression almost always requires hybridizing several microarrays from a population to capture and reduce response variability. This variability can be due to two factors: biological variability of the population and experimental variability. It is difficult to separate these two factors and most analysis is performed with a statistical model which lumps them together.

4. FDR FOR MULTIPLE CRITERIA

For comparing experiments in a way that accounts for statistical variations it is essential to report a figure of statistical significance of the each of the findings. Two important quantities indicative of statistical significance are the p-value, associated with testing a single gene response, and the false discovery rate (FDR), associated with testing all the gene probes simultaneously (multiple comparisons). In gene microarray experiments the biologist is always making multiple comparisons so FDR is the more appropriate quantity. Let each gene on the microarray have measured aggregate fitness $\xi_1(g) = u_1(g), \ldots, \xi_p(g) = u_p(g)$, e.g., a statistic computed as the average fitness of g over all of the microarray replicates. For ease of presentation, we assume that the statistical distribution P of $\xi_1(g), \ldots, \xi_p(g)$ is known when the probe responses are spatially independent and identically distributed (i.i.d.) random variables over the microarray. In other words the aggregate fitness statistic is distribution free under the null hypothesis that all probe responses are i.i.d. The p-value is computed for a single gene probe, say gene g_o, and is the probability that purely random effects, i.e., i.i.d. probe responses, would have caused g_o to be selected. More precisely the p-value for g_o is defined as:

$$\mathrm{pv}(g_o) = P(\xi_1 > u_1(g_o), \ldots, \xi_p > u_p(g_o))$$

where $\xi_1, \ldots, \xi_p$ are random variables are computed fitness levels of an i.i.d. random sample. If an experimenter were only interested in deciding on the biological significance of a single gene g_o based only on observing that gene, then reporting $p(g_o)$ would be sufficient for another biologist to assess the statistical significance of the experimenter's finding. In contrast to the p-value, FDR communicates statistical significance of an experimenters decisions made on the basis of many gene probes. The FDR is the probability that purely random effects would have caused specific genes to be selected among all probes on the microarray.

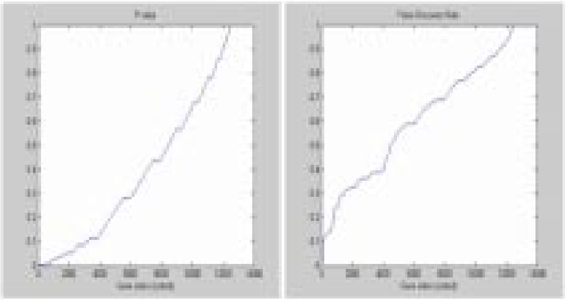

Figure 4. *The maximum p-value for multiple criteria gene selection in the aging gene mouse retina microarray experiment (left). The FDR, computed from the p-value using a well known formula,[20] for the same experiment (right). The genes are rank ordered in terms of their p-value and FDR probabilities, respectively.*

When the null distribution P is unknown, the p-value and the FDR can be computed empirically by simulation or resampling. More information on FDR can be found in the references.[20–22] In general an experimenter would like the maximum p-value and the FDR for his selected genes to be as low as possible to ensure a high level of statistical significance. However, as compared to the more conservative FDR, use of the maximum p-value gives an overoptimistic measure of significance. This is illustrated in Fig. 4 for the aging gene microarray study described in the next section. In terms of Fig. 7 the FDR is related to the probability that at least one of the many gene responses would fall into the selected sector.

5. APPLICATIONS

Here we illustrate statistical multi-criterion selection and ranking techniques for data from two gene microarray experiments. The biological significance of the experiment and the list of statistically significant genes found will be reported elsewhere. Our purpose here is simply to illustrate the application of our gene selection and ranking techniques. Both experiments used oligonucleotide-arrays, specifically the Affymetrix U74 mouse chips, and probe responses were extracted using the Affymetrix MAS5 data analysis package.[10]

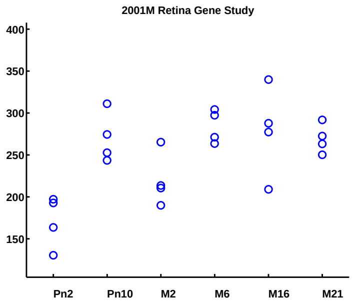

Figure 5. *24 data points (4 replicates at each 6 time points) for a specific gene extracted from GeneChips in mouse retina aging study.*

5.1. Strongly Increasing Profiles

The experiment consists of 24 retinal tissue samples taken from a population of age-sorted mice at 6 ages (time points) with 4 replicates per time point. These 6 time points consisted of 2 early development (Pn2-Pn10) and 4 late development (M2-M21) time points. DNA from each sample of retinal tissue was amplified and hybridized to the 12,422 probes on one of 24 Affymetrix U74 GeneChips. The data arrays from the GeneChips were processed by Affymetrix MAS5 software to yield probe response data. We eliminated from analysis all genes that MAS5 called out as "absent" from all chips, leaving 8826 genes for analysis. Figure 5 shows the 24 data points for a particular gene among the 8826 genes studied. Define the gene response datum extracted from the m-th microarray replicate at time t for the g-th gene probe location:

$$x_{t,m}(g), \quad g = 1, \ldots, G, \; m = 1, \ldots, M, \; t = 1, \ldots, T. \quad (1)$$

where $G = 8826$, $M = 4$, $T = 6$. Figure 5 shows the response data $\{x_{t,m}(g)\}_{t,m}$ for one of the genes extracted from the GeneChip. The scientific objective of the experiment is to find genes which are strongly associated with aging and development, i.e. those that are strongly monotonic over time. Template matching methods are not effective here since they require specification of a profile pattern and, due to variability in the experiment, this can miss genes that have the desirable monotonicity characteristics but do not agree with the specified pattern. Thus we adopted the following multi-criteria approach. We designed criteria to key onto three types of profiles: 1) those that are monotonically increasing; 2) those that are monotonically decreasing; 3) those that display a large end-to-end change. We only describe the gene selection method for monotonic increasing case as the treatment of the decreasing case is analogous. In order to tease out the monotonic increasing profiles we previously proposed a natural *virtual profile* criterion that counts the number of monotonic increasing trajectories among the $6^4 = 4096$ possible trajectories that could pass through the 24 data points[7] (See Fig. 6). However, even though it is arguably a more compelling monotonicity statistic, it has exponential computational complexity $O(M^T)$ and, to our knowledge, its p-values are not tabulated. For these reasons, for the gene selection application we prefered to use the well known Jonckheere-Terpstra (JT) test statistic[23] as criterion ξ_1. For end-to-end change we adopted the one sided paired t-test statistic[24] as criterion ξ_2. The JT statistic essentially counts the number of times that a sample at a future time point is larger than a sample at a previous time point and its computation is only of polynomial complexity ($O((T + 1)T/2M^2)$). The paired t-test statistic is an optimal end-to-end selection criterion when the extracted probe responses are Gaussian random variables with identical variances. An implicit assumption underlying the use of the JT and student-t test statistics is that the probe responses have identical distributions except for a possible shift in location, as measured by the mean or median. This assumption is reasonable after normalization of the gene microarrays, e.g. after using the RMA procedure.[14] As our collaborators are primarily interested in the genes that are implicated in late development or aging, we dropped the first two time points in the data set for the analysis described below.

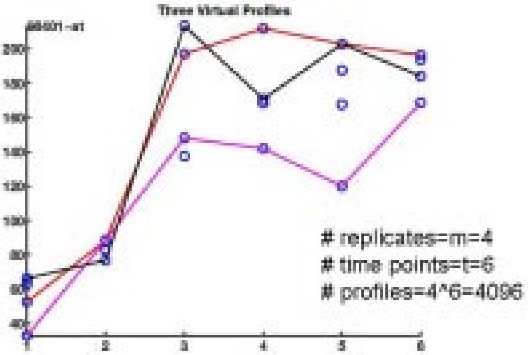

Figure 6. *3 of the $6^4 = 4096$ virtual profiles that can be drawn through the 24 gene responses in mouse retinal aging study. None of these 3 are monotonic. Label at top left denotes the gene's Unigene number.*

While the sampling distribution of the JT statistic is known exactly under an assumption of spatially i.i.d. probe responses, the sampling distribution for the paired t-test is not known exactly unless the responses are Gaussian distributed. Therefore we chose to generate the FDR contours empirically using a resampling method. In this method we

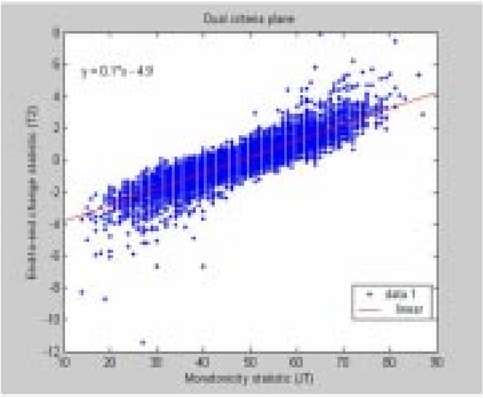

Figure 7. *The multicriterion scattergram of pairs $\{\xi_1(g), \xi_2(g)\}_{i=1}^{G}$ for i.i.d. resampled GeneChip probe responses appears approximately Gaussian distributed with regression line as indicated. Here ξ_1 is equal to the JT statistic and ξ_2 is equal to T2 which denotes the end-to-end paired t test statistic.*

simulated 500 sets of i.i.d. probe responses $\{x_{t,m}(g)\}_{t,m,g}$ for which the marginal distribution matches the empirical margin distribution of $\{x_{t,m}(g)\}_{m,g}$ at each time point t. Using these 500 simulated GeneChip data sets we determined FDR by computing the relative frequency that any gene fitness statistic $[\xi_1(g), \xi_2(g)]$ falls in a given sector. By varying the position $[u_1, u_2]$ of these sectors over the plane constant FDR contours were determined. To obtain the most discriminating multicriterion test we made an orthogonalizing transformation to data in the multicriterion plane. This transformation was motivated by the observation that the scattergrams of the resampled data (see Fig. 7) appeared to be a correlated approximately bivariate Gaussian sample. Using a regression of ξ_2 on ξ_1 we determined a monotonic transformation that converted these resampled scattergrams into approximately orthogonal bivariate Gaussian scatter plots. This transformation was then applied to the original data set to determine a set of monotonic increasing genes at a FDR level of 0.1 (see Fig. 8). Shown in Fig. 9 are the 9 top ranked monotone increasing gene profiles among the 16 genes selected.

5.2. Differentially Expressed Profiles

The second experiment we describe is concerned with finding genes whose expression profiles change significantly after a treatment. Such genes are called ”differentially expressed” after treatment. One variant of this experiment is

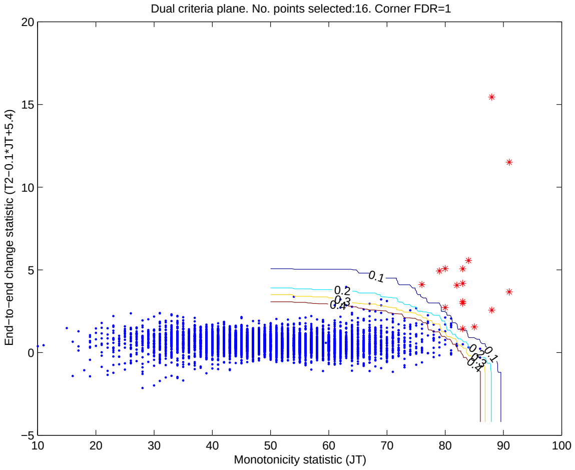

Figure 8. *Fitness criteria plotted in orthogonalized dual criteria plane of ξ_1=JT and ξ_2=T2 statistics for detecting increasing genes in aging study. Superimposed are the constant contours of FDR and highlighted genes (asterisks) that pass at a FDR level of 0.1.*

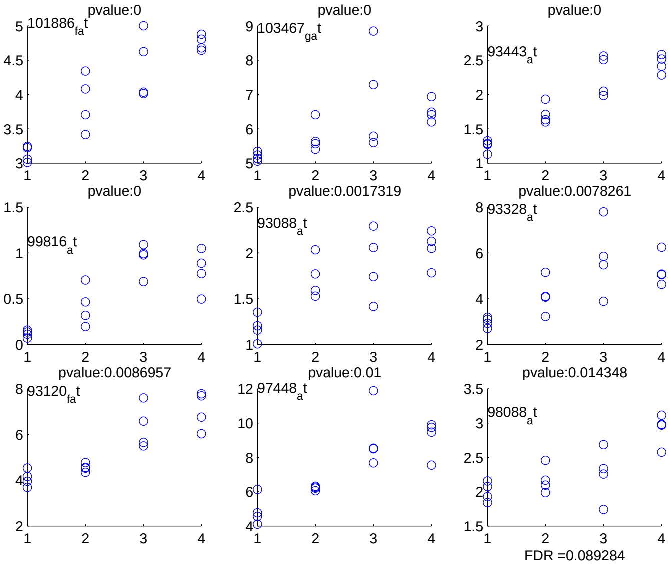

Figure 9. *Last 4 time points of gene trajectories associated with the top 9 ranked genes among those FDR = 0.1 genes shown by asterisks in Fig. 8.*

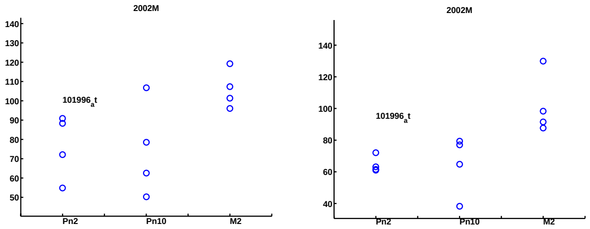

Figure 10. Responses for a gene in knockout mouse (left) vs wildtype mouse (right) for differential expression study.

called a wildtype vs knockout experiment. In this experiment one has a control population (wildtype) of subjects and a treated population (knockout) of subjects whose DNA has been altered in some way. One then collects cell samples from both populations at different times and generates microarray data sets to find any genes that are differentially expressed. Figure 10 shows gene probe responses from such a wildtype and knockout experiment. We label the wildtype and knockout responses $W_{t,m}(g)$ and $K_{t,m}(g)$ in a similar manner to (1) where here $M = 4, T = 3$.

The dual criteria chosen were: 1) a Mack-Skillings (MS) statistic for testing for parallel W vs. K responses (profiles) in a two way layout[23]; and 2) a multivariate paired t (MVPT) test statistic for quantifying the amount of difference in the W vs. K responses.[24] Similarly to the previous experiment these two criteria are complementary: the MS test is a distribution free rank-order statistical test while the MVPT is optimal under the Gaussian assumption. We applied nonlinear transformations to these two criteria to stabilize their variances. Similarly to before we used a resampling method to empirically compute FDR contours in the dual criteria plane. These contours were superimposed on the multicriterion scattergram (see Fig. 11) to find the set of genes that are differentially expressed at a FDR of prescribed level. Figure 12 shows the 9 top ranked differentially expresse gene profiles among the 142 genes selected.

6. CONCLUSION

Signal processing for analysis of gene microarray and other gene experiments is a growing area and there are enough challenges to keep the community busy for years. In our collaborations we have found it crucial to interact closely with our biology colleagues to ensure that our signal processing methods are relevant and capture the biological aims of the experimenter. To illustrate this point, in this paper we have described one of our projects involving gene selection and ranking. To respond to the needs of our collaborators we had to develop a flexible multi-criterion approach to gene selection and ranking. A single criterion would have much greater

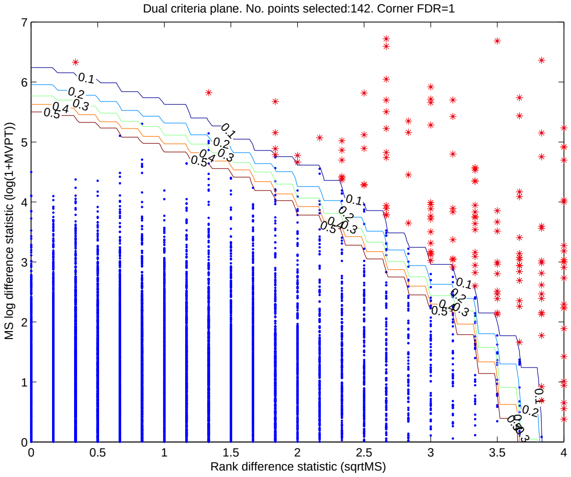

Figure 11. *Fitness of aging genes plotted in transformed dual criteria plane for detecting differentially expressed genes. Points on the plane are the square root Mack-Skillings (MS) statistic and the log of 1 plus the multivariate paired T test (MVPT). Superimposed are the constant contours of FDR and genes (asterisks) that pass the multi-criterion test at a FDR of 0.1.*

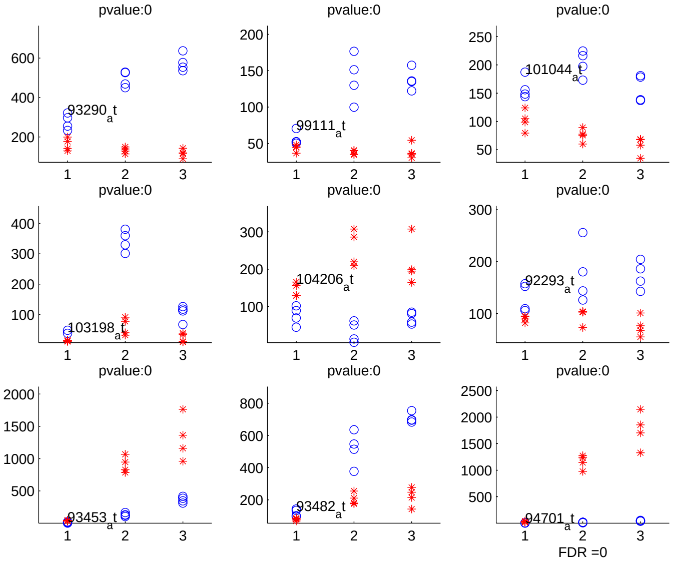

Figure 12. *Gene trajectories of top 9 ranked FDR≤0.1 genes in Fig. 11. Knockout "o" and Wildtype "*" are as indicated.*

difficulty in capturing the variety of properties that our collaborators considered biologically significant. To account for statistical variation, we had to extend multi-criterion optimization to a stochastic setting. We continue to refine our methods to meet the changing requirements of interacting with a very rapidly changing field.

Acknowledgement

The author would like to thank Prof. G. Fleury at Ecole Supérieure d'Electricité for the collaboration that has influenced work reported here. I would also like to thank Prof. A. Swaroop, Dr. S. Yosida, R. Farjo, and A. Mears in the Dept. of Human Genetics at University of Michigan for their guidance and support.

REFERENCES

1. P. A. Rota and *etal*, "Characterization of a novel coronavirus associated with severe acute respiratory syndrome," *Science*, vol. 10.1126, , May 1 2003. `www.scienceecpress.org`.
2. M. Marra and *etal*, "The genome sequence of the SARS-associated coronavirus," *Science Express*, vol. 10.1126, , May 1 2003. `www.scienceecpress.org`.
3. G. Fleury, A. O. Hero, S. Yosida, 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.
4. G. Fleury, A. O. Hero, S. Yosida, T. Carter, C. Barlow, and A. Swaroop, "Pareto analysis for gene filtering in microarray experiments," in *European Sig. Proc. Conf. (EUSIPCO)*, Toulouse, FRANCE, 2002.
5. A. Hero and G. Fleury, "Posterior pareto front analysis for gene filtering," in *Proc of Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, Raleigh-Durham, NC, 2002.
6. K. I. Siddiqui, A. Hero, and M. Siddiqui, "Mathematical morphology applied to spot segmentation and quantification of gene microarray images," in *Proc of ASILOMAR Conference on Signals and Systems*, Pacific Grove, CA, 2002.
7. A. Hero and G. Fleury, "Pareto-optimal methods for gene analysis," *Journ. of VLSI Signal Processing, Special Issue on Genomic Signal Processing*, vol. accepted, , 2003. `www.eecs.umich.edu/~hero/bioinfo.html`.
8. 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.
9. 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.
10. Affymetrix. *NetAffx User's Guide*, 2000. `www.netaffx.com/site/sitemap.jsp`.
11. National Human Genome Research Institute (NHGRI). *cDNA Microarrays*, 2001. `www.nhgri.nih.gov/DIR/Microarray`.
12. C. Li and W. Wong, "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection," *Proc. of Nat. Academy of Sci. (PNAS)*, vol. 98, pp. 31–36, 2001.
13. C. Li and W. Wong, "Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application," *Genome Biology*, vol. 2, pp. 1–11, 2001.
14. Y. H. Yang, S. Dudoit, P. Liu, and T. P. Speed, "Normalization for cdna microarray data," in *Proc of SPIE BIOS*, San Jose, California, 2001.
15. R. Irizarry, B. Hobbs, F. Collin, Y. Beazer-Barclay, K. Antonellis, U. Scherf, and T. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, To appear.
16. K. Strimmer. *R Packages for Gene Expression Analysis*. `www.stat.uni-muenchen.de/~strimmer/rexpress.html`.
17. 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.
18. A. A. Alizadeh and etal, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, 2000.
19. 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.
20. C. R. Genovese, N. A. Lazar, and T. E. Nichols, "Thresholding of statistical maps in functional neuroimaging using the false discovery rate," *NeuroImage*, vol. 15, pp. 772–786, 2002.
21. Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Royal Statistical Society*, vol. 57, pp. 289–300, 1995.
22. J. D. Storey and R. Tibshirani, "Estimating false discovery rates under dependence, with applications to dna microarrays," Technical Report 2001-28, Department of Statistics, Stanford University, 2001.
23. M. Hollander and D. A. Wolfe, *Nonparametric statistical methods*, Wiley, New York, 1991.
24. D. F. Morrison, *Multivariate statistical methods*, McGraw Hill, New York, 1967.