Multicriteria Gene Screening for Microarray Experiments

Alfred O. Hero[†], Gilles Fleury[*], and Sébastien Cerbourg[†]
[†]Dept. of Electrical Engineering & Computer Science
University of Michigan
1301 Beal Ave
Ann Arbor MI, USA
[*]Service des Mesures
Ecole Supérieure d'Electricité
Gif-sur-Yvette, France

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 20,000 genes in 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 list of genes which are both biologically and statistically significant based on the outcomes of gene microarray experiments. We will describe a novel multicriteria method that we have developed for this gene selection problem that allows tight control of both minimum observable differential change (biological significance) and familywise error rate (statistical significance) and also provides a set of simultaneous confidence intervals for the differences.

Keywords: bioinformatics, gene filtering, multicriteria scattergram, familywise error rates.

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.[32] It was not until almost 45 years after Watson and Crick's discovery that the first entire bacterial genome was sequenced, the E Coli bacterium containing over 4000 genes, after many years of effort. In 2000, under the auspices of the international Human Genome Project (HGP), the first draft human genome was obtained, identifying the genes in 90% of 30,000 tagged sites along the DNA double helix. In Spring 2003, and almost two years ahead of schedule, the HGP was declared complete with 99% of the human genome sequenced with 99.9% accuracy.[10] In spring 2003 the genome for the SARS corona virus (SARS-CoV) was sequenced and authenticated in less than 2 months time.[28,23] 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 present a new multicriteria approach to analyzing gene microarray data that we have developed while 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 of genes that are both biologically significant, in terms of exhibiting large foldchange over time or over treatment, and statistically significant, in terms of controlling the rate of false positives.

In our past work on signal processing for gene microarrays[12,13,16,29,17] our primary goal has been to develop statistically reliable methods for ranking temporal gene expression profiles. The work most closely related to this paper is our multicriteria optimization approach to *gene ranking* using a statistical version of Pareto front analysis.[16,17] In that 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 multicriteria approach similar to our previous work. The novelty of our gene selection method is the use of a two stage procedure: 1) perform preliminary screening using multicriteria tests of significance; and 2) perform secondary screening using false discovery rate confidence intervals (FDRCI) on foldchange. The two stage procedure allows the experimenter to simultaneously impose a minimum foldchange requirement and a prescribed family wise error rate (FER) on the set of genes selected.

We illustrate our two stage methods for two Affymetrix GeneChip experiments designed to probe the genes of the retina. In these experiments we adopt pairs of criteria for stage 1 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. Stage 2 is implemented by thresholding simultaneous confidence intervals on foldchange constructed from adjusted Student-t quantiles. The purpose of this article is to illustrate methodology and not to report scientific findings. However, as presented in,[35,24] application of our procedure has resulted in discovery of many novel genes which have been experimentally validated by more sensitive foldchange quantitation methods (RT-PCR).

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 multicriteria selection and ranking approach. In Secs. 4 and 5 we discuss familywise error rate (FER), false discovery rate (FDR), and false discovery rate confidence intervals (FDRCI) for multicriteria gene screening. Finally, in Sec. 6 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 gel 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[9,5] 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[3] product line. An example of the later is the cDNA microarray protocol of the National Human Genome Research Institute (NHGRI).[27] 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 analysis, such as MAS5,[3] the Li-Wong method,[21,22] RMA oligo-chip analysis,[20] and SMA cDNA-chip analysis.[34,2] Many of these packages are available as freeware, e.g., see websites[1,31] for links to relevant software written in the R software language.[19] 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," "gene screening," and "gene filtering" for which many methods have been proposed.[15,4,8] 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 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 multicriteria methods for selecting and ranking gene profiles.[12,16,17]

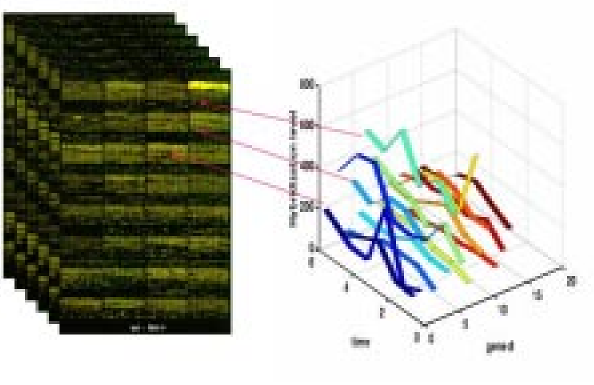

Figure 1. *Probing gene expression at several time points leads to a temporal sequence of gene microarrays (left). A few of the sequences can be extracted at specific probe locations on the microarrays and plotted as time signals (right).*

3. MULTICRITERIA SELECTION AND RANKING

As contrasted to maximizing *scalar* criteria, multicriteria gene screening 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. Or one may have to deal with two biologists who each have different criteria for what features constitute an interesting aging gene. As another example, which reflects the applications discussed below, one may wish to use two different statistical criteria; one quantitative foldchange criterion matched to an assumed model and another qualitative monotonicity criterion that is robust to violations in model assumptions.

Multicriteria Gene Selection: We define the fitness of a gene g using the vector $\underline{\xi}(g) = [\xi_1(g), \ldots, \xi_p(g)]$. Any genes whose fitness vector lies in the positive quadrant $\xi_1(g) > u_1, \ldots, \xi_p(g) > u_p$ will be said to have a "positive response." Here $u_1, \ldots, u_p$ are thresholds which could be selected by the experimenter to reflect the biological significance of a particular level of measured gene fitness $\underline{\xi}(g)$. This is illustrated in Fig. 2 where the selected sector for two aging criteria 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.

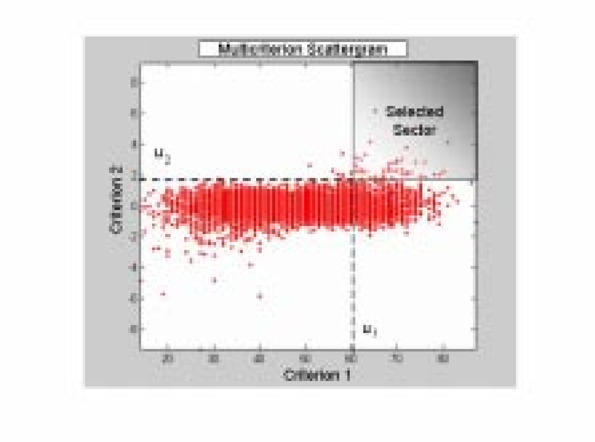

Figure 2. *Multicriteria scattergram of gene fitness responses for aging study with overlaid gene selection sector. Genes falling in this sector are declared "positive responses". The choice of position $[u_1, u_2]$ of the sector could depend on the experimenter's chosen biological significance levels. The two criteria are the JT (horizontal axis) and paired T-test (vertical axis) statistics described in Section 6.*

Multicriteria Gene Ranking: In a well designed gene microarray experiment, multicriteria (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 the most "promising genes" to investigate further. Resolution of this problem is of importance since validation of gene response requires running more sensitive amplification protocols, such as quantitative real-time reverse-transcription polymerase-chain-reaction (RT-PCR). As compared to microarray experiments, RT-PCR's higher sensitivity is offset by its lower throughput and its higher cost-per-probe. Some sort of rank ordering of the selected genes would help guide the biologist to a solution of the validation problem. As a linear ordering 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 multicriteria optimization problem. Further details on multicriteria optimization approaches to gene ranking were presented in[17] to which the reader is referred for more details. In this paper we concentrate on the gene selection problem.

Multicriteria gene selection and ranking methods are related to multicriteria optimization, also called multiobjective optimization and vector optimization,[11] which are applicable to any user defined set of criteria. However, these methods 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. ERROR RATES FOR MULTIPLE SCREENING CRITERIA

For comparing experiments in a way that accounts for statistical variations it is essential for an experimenter to report a figure of statistical significance of his findings. Three important quantities indicative of statistical significance are the p-value, associated with testing a single gene response, the familywise error rate (FER) 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 FER or FDR must be controlled. Define the aggregate fitness vector $\underline{\xi}(g) = [\xi_1(g), \ldots, \xi_p(g)]^T$ as a statistic computed by sample averaging over all of the microarray replicates of the g-th gene response. The null hypothesis is that that the response vectors $\{\underline{\xi}(g)\}$ are independent and identically distributed (i.i.d.). The objective is to detect positive gene responses which deviate from the null hypothesis by detecting gene fitness vectors lying in a positive quadrant of the multicriteria scattergram. In order to control false positive rates one needs to estimate them and this requires either knowing the statistical distribution P of the responses under the null hypothesis or implementing bootstrap procedures. For concreteness in this section we assume that P is known.

Let the measured aggregate fitness of a particular gene g be $\xi_1(g) = u_1(g), \ldots, \xi_p(g) = u_p(g)$. The p-value is computed for a single gene probe, say gene g_o, and is the probability that purely random effects would have caused g_o to be erroneously selected, generating a "false positive," based on observing microarray responses for gene g_o only. More precisely the p-value for g_o is defined as:

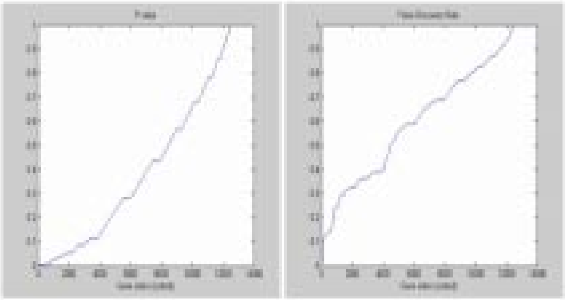

Figure 3. *The maximum p-value for multicriteria gene selection in the aging gene mouse retina microarray experiment (left). The FDR, computed from the p-value using a well known formula,[14] for the same experiment (right). The genes are rank ordered in terms of their p-value and FDR probabilities, respectively.*

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

where $\xi_1, \ldots, \xi_p$ are random variables equal to fitness levels of an i.i.d. random sample and $u_1(g_o), \ldots, u_p(g_o)$ are considered as fixed and non-random. If an experimenter were only interested in deciding on the biological significance of a single gene g_o based only on observing probes for that gene, then reporting $p(g_o)$ would be sufficient for another biologist to assess the statistical significance of the experimenter's statement that g_o exhibits a positive response. In contrast to the p-value, FER and FDR communicate statistical significance of an experimenter's finding of biological significance after observing all gene responses. The FER is the probability that there are any false positives among the set of genes selected. On the other hand, the FDR refers to the expected proportion of false positives among the selected genes. The

FDR is a less stringent criterion than the FER and weakly controls the FER.[14,6,30]

When the p-values are known the FER and the FDR can be upper bounded using Bonferroni-type methods.[25] Otherwise, the p-value, FER and FDR can be computed empirically by simulation or resampling methods[33] and this is the method we have used here. In general an experimenter would like the p-values, the FER and the FDR for his selected genes to be as low as possible in order to ensure a high level of statistical significance. However, as compared to the more conservative FER and FDR constraints, screening by the maximum p-value gives an overly optimistic measure of significance. This is illustrated for the FDR in Fig. 3 for the aging gene microarray study described in Section 6.

5. INCLUSION OF MINIMUM FOLD CHANGE CRITERION

The methods described above are applicable to discovering genes with any non-zero differential response at a prescribed level of significance. Frequently the experimenter is only interested in genes whose differential response over time or over treatment exceeds some threshold. This threshold is generally expressed in terms of log base two of the ratio of two responses and has units of "foldchange." The experimenters choice of minimum foldchange is commonly determined by the sensitivity of follow-up validation techniques such as RT-PCR. For example, our experimental collaborators commonly work with a minimum validatable fold change somewhere between 1.0 and 2.0.

For screening genes with a minimum foldchange criterion we have adopted a two-stage procedure based on the method of False Discovery Rate Confidence Intervals (FDRCI) of Benjamini and Yekutieli.[7] The first stage of this procedure uses multicriteria screening techniques, described in the previous section, to find a set of genes which are differentially expressed at a prescribed FDR level q. The second stage constructs simultaneous $(1-q)\%$ confidence intervals for the foldchanges at each time point for each gene discovered in the first stage. These confidence intervals are constructed on the time points using the FDRCI procedure of Benjamini and Yekutieli. A gene is declared "foldchange-significant" at foldchange level $f_{\min}$ and significance level q if it has at least one time point for which the foldchange confidence interval is greater than $f_{\min}$ or less than $-f_{\min}$. This procedure has the advantage of providing simultaneous confidence intervals on fold changes of each gene selected as foldchange-significant.

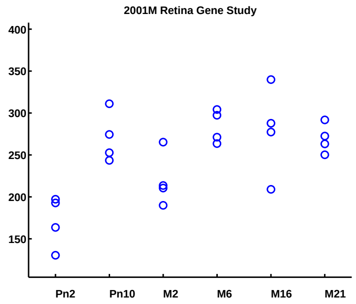

Figure 4. *24 data points (4 replicates at each 6 time points) for a specific gene extracted from 24 GeneChips in mouse retina aging study.*

6. APPLICATIONS

Here we illustrate multicriteria screening techniques for data from two gene microarray experiments. The biological significance of the experiment and the list of foldchange-significant genes found will be reported elsewhere.[35,24] Our purpose here is simply to illustrate the application of our gene selection and ranking techniques on real data. Both experiments used oligonucleotide-arrays, specifically the Affymetrix U74 mouse chips, and probe responses were extracted using the Affymetrix MAS5[3] and RMA[20] software packages.

6.1. Strongly Increasing Profiles

The experiment consists of 24 retinal tissue samples taken from each of 24 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, M6, M16, 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 log2 probe response data. We eliminated from analysis all genes that MAS5 called out as "absent" from all chips in addition to the Affymetrix housekeeping genes, leaving 6931 genes for analysis. Define the gene response datum extracted from the m-th microarray replicate at time t for the g-th gene probe location (Figure 4):

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

where $G=6931$, $M=4$, $T=6$. Figure 4 shows the response data $\{x_{t,m}(g)\}_{t,m}$ for one of the genes extracted from the Affymetrix GeneChip. The scientific objective of the experiment is to find genes which are strongly associated with aging and development, i.e. those that are monotonic over time and have large end-to-end foldchange. 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 multicriteria 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 end-to-end foldchange magnitudes greater than 1.0. We only describe the gene selection method for the monotonic increasing case as the treatment of the decreasing case is completely analogous. In order to tease out the monotonic increasing profiles we use a non-parametric distribution free statistic. In previous gene ranking work we 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.[17] However, even though it is arguably a more compelling monotonicity statistic, the virtual profile criterion has exponential computational complexity $O(M^T)$. Thus for this screening application we prefered to use the well known Jonckheere-Terpstra (JT) test statistic[18] as criterion ξ_1.

$$\xi_1(g) = \sum_{t=1}^{T} \sum_{t'>t} \sum_{m \neq m'} \text{sign}(x_{t',m'}(g) - x_{t,m}(g))$$

For end-to-end change we adopted a modified one sided paired t-test statistic[26] as criterion ξ_2.

$$\xi_2(g) = \sqrt{M/2}\frac{\overline{x_T}(g) - \overline{x_1}(g)}{\text{s}(g)} \tag{2}$$

where

$$\overline{x_t}(g) = \frac{1}{M}\sum_{m=1}^{M} x_{t,m}(g)$$

and

$$\text{s}^2(g) = \frac{1}{T(M-1)}\sum_{t=1}^{T}\sum_{m=1}^{M}(x_{t,m}(g) - \overline{x_t}(g))^2. \tag{3}$$

The null distribution of the statistic $\xi_2(g)$ is Student-t with $T(M-1)$ degrees of freedom (d.f.). The statistic (2) differs from the standard $2(M-1)$-d.f. paired t-test statistic in that we exploit the assumed homeoscedasticity ($\sigma^2_{tm}(g) = \sigma^2(g)$) of each of the probe responses to derive a more accurate pooled variance estimate (3). The p-values of the JT and paired-t statistics $\xi_1(g)$ and $\xi_2(g)$ are tabulated in[18] and,[26] respectively.

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.[34] 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.

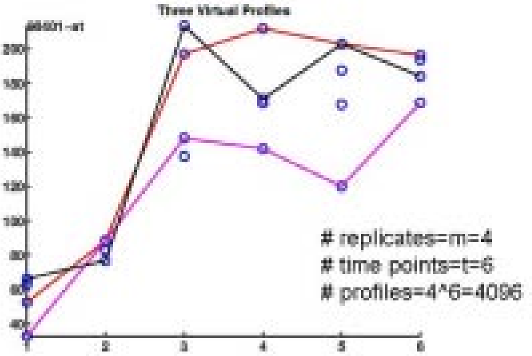

Figure 5. *3 of the $6^4 = 4096$ virtual profiles that can be drawn through the 24 gene responses in mouse retinal aging study. None of these 3 are monotonic. Label at top left denotes the gene's Affymetrix probe id number.*

Since the joint sampling null distribution of the JT and paired t-test statistics is unknown, we chose to generate FER contours empirically using a resampling method similar to the bootstrap. Specifically, we randomly permuted the probe responses $\{x_{t,m}(g)\}_{t,m,g}$ to generate 500 resampled sets of i.i.d. probe responses $\{x^{'}_{t,m}(g)\}_{t,m,g}$ for which the marginal distribution matches the empirical marginal distribution of $\{x_{t,m}(g)\}_{t,m,g}$. Using these 500 simulated GeneChip data sets we determined FER by computing the relative frequency that any gene fitness statistic $[\xi_1(g), \xi_2(g)]$ computed from $\{x^{'}_{t,m}(g)\}_{t,m,g}$ falls in a given sector as explained in Sec. 4. By varying the lower left endpoint $[u_1, u_2]$ of these sectors over the plane constant FER contours were determined.

To obtain the most discriminating multicriteria test we made an orthogonalizing transformation to data in the multicriteria plane. This transformation was motivated by the observation that the scattergrams of the resampled data (see Fig. 6) appeared to be a correlated approximately bivariate Gaussian sample. Using a regression of ξ_2 on ξ_1 we determined

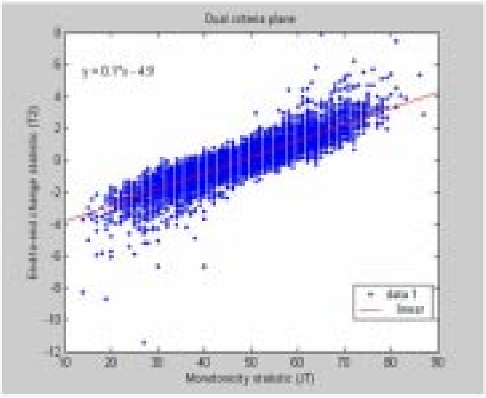

Figure 6. *The multicriteria scattergram of pairs $\{\xi_1(g), \xi_2(g)\}_{i=1}^G$ for i.i.d. resampled GeneChip probe responses appears approximately Gaussian distributed with regression line as indicated. Here ξ_1 is equal to the JT statistic and ξ_2 is equal to T2 which denotes the end-to-end paired t test statistic.*

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 of $G = 6931$ gene responses to determine a set of monotonic increasing genes at a FER level of q (see Fig. 7). This first stage of screening results in a set $\mathcal{G}_1$ of G_1 genes with declared positive responses. The second stage of screening consists of constructing the following level $(1-q)100\%$ simultaneous FDR confidence intervals on the foldchanges fc for these G_1 genes:

$$\overline{x_T}(g) - \overline{x_1}(g) - \mathrm{s}(g)/\sqrt{M/2}\,\mathcal{T}^{-1}_{T(M-1)}(1-q'/2) \leq \mathrm{fc}(g)$$
$$\leq \overline{x_T}(g) - \overline{x_1}(g) + \mathrm{s}(g)/\sqrt{M/2}\,\mathcal{T}^{-1}_{T(M-1)}(1-q'/2),$$

where $g \in \mathcal{G}_1$. Here $q' = qG_1/G$ is the adjusted FDRCI significance level,[7] $\overline{x_t}(g) = M^{-1}\sum_{m=1}^{M} x_{t,m}(g)$, and $\mathcal{T}_\nu^{-1}(\alpha)$ is the α quantile of the Student-t distibution with ν d.f. The second stage is completed by retaining those genes in $\mathcal{G}_1$ whose M2-to-M21 foldchange confidence intervals do not intersect the interval $(-\infty, f_{\min}]$ (See Fig. 8).

Shown in Fig. 9 are the profiles of the 2 genes who have monotone increasing gene profiles with foldchange at least $fcmin = 0.5$ at FDRCI level 0.1. The stringency of this screening procedure is reflected by the fact that the FER's for each of these gene are substantially below the FDRCI level.

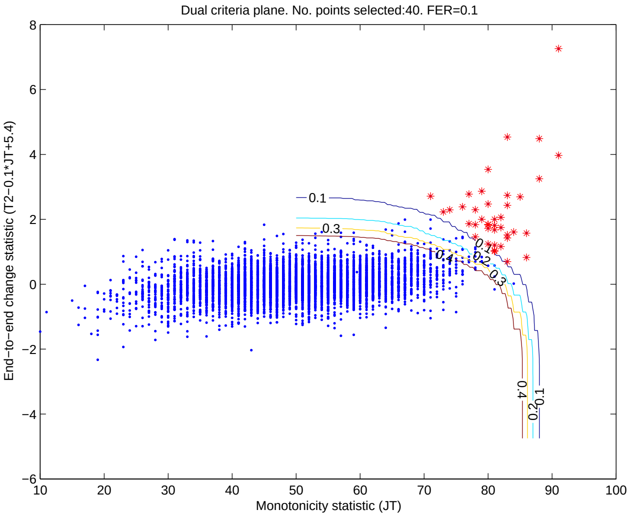

Figure 7. *Fitness criteria plotted in orthogonalized dual criteria plane of ξ_1=JT and ξ_2=T2 statistics for detecting increasing genes in aging study. Superimposed are the constant contours of FER and 40 highlighted genes (asterisks) that pass the first stage of screening for monotonic-increasing profiles at FER level 10%.*

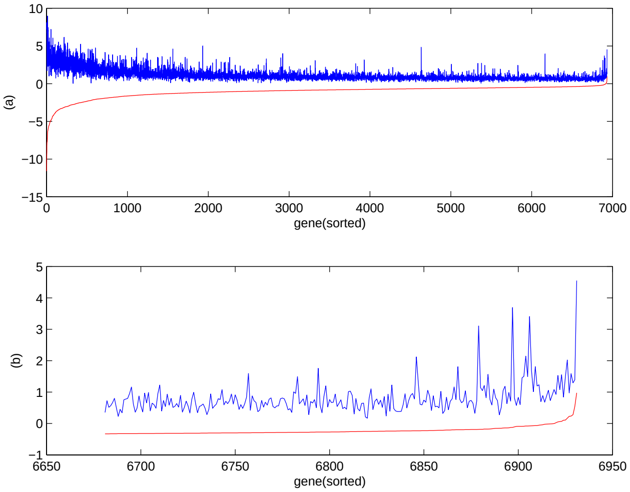

Figure 8. *(a) Plot of the upper and lower end-points of the 10% FDR confidence intervals (FDRCI) on M2-to-M21 foldchanges $\{\mathrm{fc}(g)\}$ sorted by lower endpoint (lower curve). (b) blowup of (a) over the 250 largest lower endpoint values. Only the two genes whose lower FDRCI endpoint is greater than the minimum foldchange $f_{\min} = 0.5$ pass the second stage of screening.*

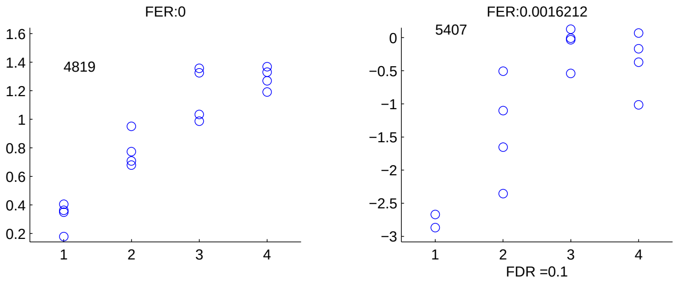

Figure 9. *Last 4 time points of the two gene trajectories (log2) with foldchange at least $fcmin = 0.5$ at FDRCI level 0.1 in the mouse retinal aging study. FER's for each gene are substantially below the FdRCI level. The numbers at top left of each plot simply identify these two genes in our library.*

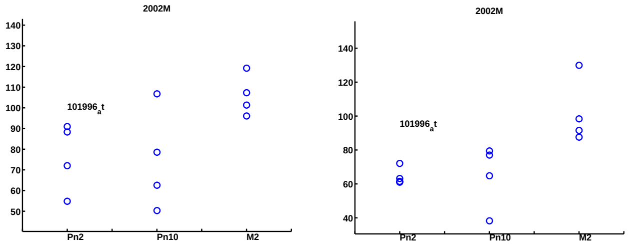

Figure 10. *Responses for a gene in knockout mouse (left) vs wildtype mouse (right) for differential expression study.*

6.2. 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 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 performed on two populations of mice by collaborators at the Sensory Gene Microarray Node at the University of Michigan. The population consisted of 12 knockout and 12 wildtype mice each divided into 3 subgroups of 4 mice. The 3 subgroups correspond to different time points: postnatal 2 days (Pn2), postnatal 10 days (Pn10), and postnatal 2 months (M2). The log2 probe responses were extracted from the Affymetrix GeneChips using the RMA algorithm. The scientific objective of the experiment is to find genes whose temporal expression profiles in the wildtype and knowckout population are significantly different. We label the wildtype and knockout responses $W_{t,m}(g)$ and $K_{t,m}(g)$ in a similar manner to (1) where here $M = 4$ and $T = 3$.

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[18]; and 2) a multivariate paired t (MVPT) test statistic for quantifying the amount of average difference in the W vs. K responses.[26] 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. To reduce dynamic range of the multicriteria scattergram we applied non-linear transformations $\mathrm{MS} \longmapsto \sqrt{\mathrm{MS}}$ and $\mathrm{MVPT} \longmapsto \log(1 + MVPT)$ (when this latter statistic is multiplied by $M/2$ it is approximately Chi-square distributed). Similarly to the aging study we used a bootstrap resampling method to empirically compute FER contours in the dual criteria plane. These contours were superimposed on the multicriteria scattergram (see Fig. 11) to find the set of genes that are differentially expressed at a FER of prescribed level. Again we denote by G_1 the number of genes discovered in this first stage of the screening procedure. Stage 2 of the test consisted of retaining only those genes whose FDRCI's on differential foldchange $\{\mathrm{fc}_t(g)\}_{t=1}^{T}$ do not intersect $[-f_{\min}, f_{\min}]$ for any time point t (see Fig. 12). Specifically, the TG_1 level $(1-q)100\%$ simultaneous FDRCI intervals were computed as:

$$\overline{W_t}(g) - \overline{K_t}(g) - \mathrm{s}_t(g)/\sqrt{M/2}\,\mathcal{T}^{-1}_{2(M-1)}(1-q'/2) \leq \mathrm{fc}_t(g)$$
$$\leq \overline{W_t}(g) - \overline{K_t}(g) + \mathrm{s}_2(g)/\sqrt{M/2}\,\mathcal{T}^{-1}_{2(M-1)}(1-q'/2)$$

where q' is the adjusted confidence level

$$q' = 1 - (1 - qG_1/G)^{1/T},$$

and $\mathrm{s}_t(g)$ is the pooled variance estimate obtained from $\{W_{t,m}(g)\}_m$ and $\{K_{t,m}(g)\}_m$.

Figure 13 shows 9 of the differentially expressed gene profiles in (log2 scale) among the 15 genes selected by the two stage screening procedure at FDRCI level of significance $q = 0.1$ and minimum foldchange of $f_{\min} = 4.0$.

7. 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

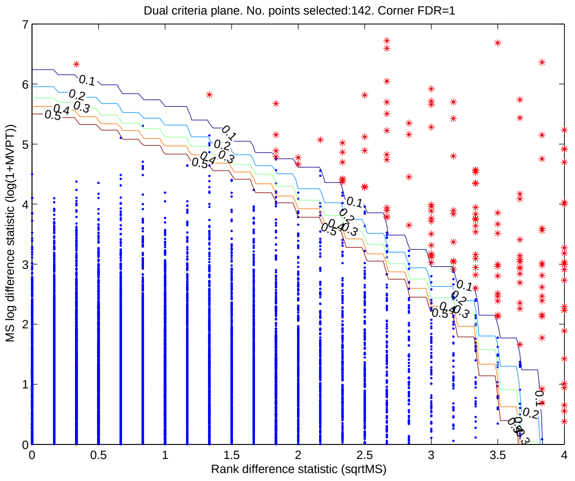

Figure 11. *Fitness of genes plotted in transformed dual criteria plane for detecting differentially expressed genes in knockout study. Points on the plane are the square root Mack-Skillings (MS) statistic and the log of 1 plus the multivariate paired T test (MVPT). Superimposed are the constant contours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

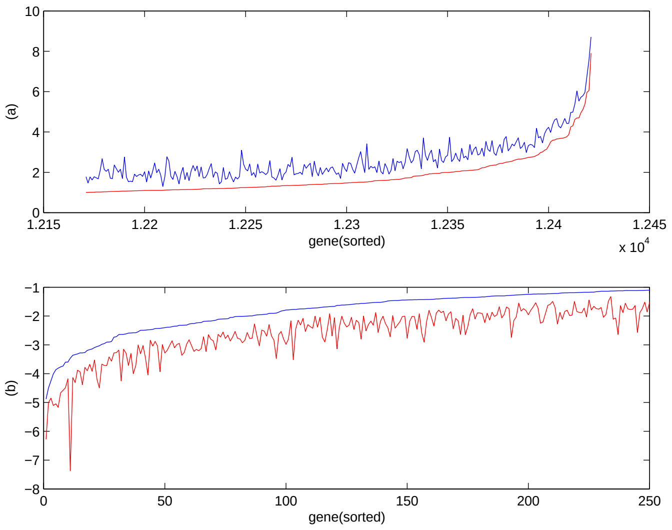

Figure 12. *(a) Segment of upper and lower curves specifying the 10% FDR confidence intervals (FDRCI) on the maximum foldchange $max_{t=1,2,3}\{\mathrm{fc}_t(g)\}$ sorted according to genes having largest lower endpoint (lower curve). (b) same as (a) except that FDRCI's are on the maximum foldchange sorted by largest upper endpoint values (upper curve). Only those genes whose FDRCI's do not intersect $[-f_{\min}, f_{\min}]$ pass the second stage of screening.*

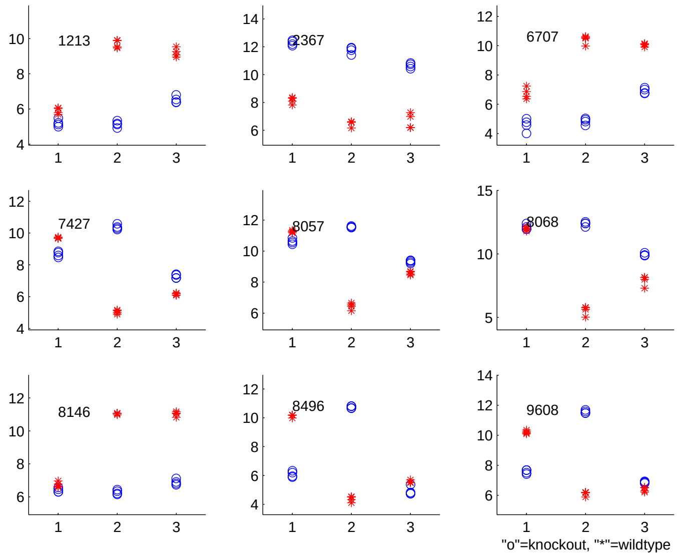

Figure 13. *Gene trajectories of 9 differentially expressed genes in Fig. 11 with FDRCI level of significance $q = 0.1$ and minimum foldchange of $f_{\min} = 4.0$. Knockout "o" and Wildtype "*" are as indicated.*

described one of our projects involving gene selection and ranking. To respond to the needs of our collaborators we had to develop a flexible multicriteria approach to gene selection and ranking. A single criterion would have much greater difficulty in capturing the variety of properties that our collaborators considered biologically significant. To account for statistical variation, we had to extend multicriteria optimization to a stochastic setting. To accomodate our collaborators minimum fold change requirements we had to incorporate simultaneous confidence intervals into our screening procedure. We continue to refine our methods to meet the changing requirements of interacting with a very rapidly changing field.

Acknowledgement

The authors would 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 collaboration.

REFERENCES

1. *Bioconductor: open source software for bioinformatics*. `www.bioconductor.org/\verb`.
2. *SMA microarray analysis package*. `www.stat.berkeley.edu/users/terry/zarray/Software/smacode.html`.
3. Affymetrix. *NetAffx User's Guide*, 2000. `www.netaffx.com/site/sitemap.jsp`.

4. 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.
5. 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.
6. 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.
7. Y. Benjamini and D. Yekutieli, “False discovery rate adjusted confidence intervals for selected parameters (preprint),” *J. Am. Statist. Assoc.*, vol. Submitted (2002), , 2002. `www.math.tau.ac.il/˜yekutiel/ci_jasa.pdf`.
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. F. C. Collins, M. Morgan, and A. Patrinos, “The Human Genome Project: lessons from large-scale biology,” *Science*, vol. 300, pp. 286–290, April 11 2003.
11. K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*, Wiley, New York, 2001.
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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`.
18. M. Hollander and D. A. Wolfe, *Nonparametric statistical methods (2nd Edition)*, Wiley, New York, 1991.
19. R. Ihaka and R. Gentleman, “R: A language for data analysis and graphics,” *Journal of Computational and Graphical Statistics*, vol. 5, no. 3, pp. 299–314, 1996.
20. 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.
21. 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.
22. 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.
23. M. Marra and *etal*, “The genome sequence of the SARS-associated coronavirus,” *Science Express*, vol. 10.1126, , May 1 2003. `www.scienceecpress.org`.
24. A. Mears and etal, “ms in preparation,”, 2003.
25. R. G. Miller, *Simultaneous Statistical Inference*, Springer-Verlag, NY, 1981.
26. D. F. Morrison, *Multivariate statistical methods*, McGraw Hill, New York, 1967.
27. National Human Genome Research Institute (NHGRI). *cDNA Microarrays*, 2001. `www.nhgri.nih.gov/DIR/Microarray`.
28. 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`.
29. 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.
30. 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.
31. K. Strimmer. *R Packages for Gene Expression Analysis*. `www.stat.uni-muenchen.de/˜strimmer/rexpress.html`.
32. J. Watson and A. Berry, *DNA: The secret of life*, Alfred A. Knopf, 2003.
33. P. Westfall and S. Young, *Resampling-Based Multiple Testing*, Wiley, NY, 1993.
34. 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.
35. S. Yosida and etal, “ms in preparation,”, 2003.