Gene Discovery Using Pareto Depth Sampling Distributions

G. Fleury[◇], A. Hero[†]

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

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

August 2, 2003

Abstract

Most methods for finding interesting gene expression profiles from gene microarray data are based on a single discriminant, e.g. the classical paired t test. Here a different approach is introduced based on gene ranking according to Pareto depth in multiple discriminants. The novelty of our approach, which is an extension of our previous work on Pareto front analysis (PFA), is that a gene's relative rank is determined according to the ordinal theory of multiple objective optimization. Furthermore, the distribution of each gene's rank, called Pareto depth, is determined by resampling over the microarray replicates. This distribution is called the Pareto depth sampling distribution (PDSD) and it is used to assess the stability of each ranking. We illustrate and compare the PDSD approach with both simulated and real gene microarray experiments.

Keywords: gene filtering, multi-objective optimization, false discovery rate, data depth analysis.

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 [24]. 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 Spring 2003, and almost two years ahead of schedule, the Human Genome Project was declared complete with 99% of the human genome sequenced at an unprecedented 99.9% accuracy [4]. In Spring 2003 the genome for the SARS corona virus (SARS-CoV) was sequenced and authenticated in less than 2 months time [20, 19]. These recent leaps in progress would not have been possible without significant advances in gene discovery technology. One such technology, which is the main focus of this paper, are gene microarrays and their associated signal extraction and processing algorithms [18, 16, 17, 5]. Specifically, we will present a new method for analyzing gene microarray data which we call the method of Pareto depth sampling distributions (PDSD).

The massive scale and variability of microarray gene data creates new and challenging problems of clustering and data mining: the so-called *gene filtering problem*. This problem has two subproblems called *gene screening* and *gene ranking*. Gene screening is concerned with determining a list of gene probes whose expression levels are statistically and biologically significant with respect to some p-value or familywise error rate. Gene ranking is concerned with finding a fixed number of genes that are rank ordered according to one or more statistical and biological criteria. These two subproblems are closely related, but this paper focusses on gene ranking using multiple criteria. Multicriteria methods of gene screening with familywise error constraints have been presented elsewhere [12] and will not be discussed further in this paper.

Multicriteria gene filtering seeks to find genes whose expression profiles strike an optimal compromise between maximizing (or minimizing) 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. In a well designed gene microarray experiment, multicriteria (or other) methods of screening 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.

It is thus clear that some sort of rank ordering of the selected genes would help guide the biologist to a cost effective solution of the validation problem. As a linear ordering of multiple criteria does not generally exist, an absolute ranking of the selected genes is generally impossible. However a partial ordering is possible when formulated as a multicriterion optimization problem. This partial ordering groups genes into successive Pareto fronts of the multicriteria scattergram (see Section 3). It is this partial ordering which was used in our previous work [8, 9, 11] to obtain relative rankings of gene expression levels based on microarray experiments. We called our multiobjective approach to gene ranking *Pareto front analysis* (PFA). As pointed out in [11] the PFA approach is related to John Tukey's notion of data depths and contours of depth in a multivariate sample [22, 6]. To highlight the contrast between PFA and the concept of data depths we will refer to the *Pareto depth* of a gene as the Pareto front on which the gene lies. It is to be noted that Pareto analysis has been adopted for many continuous and discrete optimization applications including evolutionary computing [21, 25].

Several variants of PFA were introduced in [8, 9, 11] including resistant PFA (RPFA), based on cross-validation, and posterior PFA (PPFA), based on Bayes posterior analysis, of gene rankings. These rankings were computed by rank ordering each gene's probability of lying on the first two or three Pareto fronts of the multicriteria scattergram. This paper introduces a more powerful PFA gene discovery tool, the aforementioned Pareto depth sampling distribution method, into the PFA toolbox. The PDSD method generates an empirical distribution of the depth of the front, the Pareto depth, on which each gene lies. This distribution is computed by implementing a resampling method similar to the bootstrap. From this distribution many different attributes of the Pareto depth can be determined and used for ranking the genes. The PDSD approach is more general than our previous cross-validation PFA approach that used a special attribute, the cumulative PDSD, to rank the genes. Using simulations we compare our PFA methods to the standard paired t test on the basis of correct discovery and false discovery rates. Our principal conclusion is that the PDSD approach, when formulated as a Pareto depth test, significantly outperforms previous PFA and paired t test methods.

Experience with our collaborators in the Dept. of Human Genetics at the University of Michigan has shown that the PFA methodology can discover important genes that elude standard analysis such as paired t test or other analysis of variance (ANOVA) methods. However, our objective here is to introduce and illustrate the PFA methodology and we do so using both controlled simulations and experimental data acquired from our collaborators. These datasets are representative of real world microarray experiments for studying the genetics of retinal aging and disease. We will report on more comprehensive comparisons, biological significance of genes discovered using PFA, and scientific significance of the experiments in future publications.

The paper is organized as follows. In Sec. 2 we give some background on genomics and briefly review gene microarray technology. In Sec. 3 we motivate and describe the PFA multicriterion ranking approach and introduce the concept of PDSD's. In Sec. 4 we report on quantitative comparisons between the Pareto depth test and other tests used for gene selection and ranking. Finally, in Sec. 5 we conclude with some general remarks.

2 Background

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 display a significant change in expression level between a pair of time points. 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 probes that are put in distinct locations on a small slide [14, 1, 7]. After hybridization of an unknown tissue sample to the microarray, the abundance of each probe present in the sample can be estimated from the measured levels of hybridization, called probe responses, of the sample to each probe.

Due to high response variability the study of differential gene expression between two or more populations or time points usually requires hybridizing several samples from each population. We assume that there are T populations each consisting of M_t samples, $t = 1, \dots, T$. For each of the $\sum_{t=1}^{T} M_t$ samples we assume an independent microarray hybridization experiment is performed yielding N gene probe responses extracted from the microarray. Define the measured response of the n-th probe on the m-th microarray acquired at time t

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

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) [10, 3]. Validation methods have been widely used and include [23, 15]: significance analysis of microarrays (SAM); bootstrapping cluster analysis; and leave-one-out cross-validation. 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. As contrasted to maximizing such *scalar* criteria, multicriteria gene filtering seeks to find the best compromise between maximizing or minimizing several criteria. This method is closely related to multi-objective optimization which has been used in many applications [21, 25].

2.1 Data Sets

Data from two microarray experiments are used in the sequel to illustrate our analysis. These data were collected by collaborators in Anand Swaroop's laboratory in the Dept. of Ophthalmology at the University of Michigan. We briefly describe these two experiments below.

2.2 Mouse Retinal Aging Study

The experiment consists of hybridizing 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) samples. DNA from each sample of retinal tissue was amplified and hybridized to the 12,422 probes on one of 24 Affymetrix U74 Mouse GeneChip microarrays. The data arrays from the GeneChips were processed by Affymetrix MAS5 software to yield log2 probe response data. Of interest to our biology collaborators is the effect of aging on retinal gene expression. For this purpose we compare two populations comprising the 8 tissue samples at the two extreme late development time points M2 and M21. Our objective was to find genes with a high level of differential expression between these points.

2.3 Human Retinal Aging Study

The experiment consists of hybridizing 16 retinal tissue samples taken from 8 young human donors and 8 old human donors. The ages of the young donors ranged from 16 to 21 years and the ages of the old donors ranged from 70 to 85 years old. The 16 tissue samples were hybridized to 16 Affymetrix Human GeneChip microarrays each containing $N = 12,642$ probes. Again MAS5 software was used to extract log2 probe response data and our objective was to find genes with a high level of differential expression between the young and old populations.

3 Gene Screening and Ranking

Consider the problem of finding a set of genes whose mean expression levels are significantly different between a pair of populations ($T = 2$). The measured probe responses from such genes should exhibit small within-population variability (intra-class dispersion) and large between-population variability (inter-class dispersion). Two natural measures of intra-class dispersion ξ_1 and inter-class dispersion ξ_2, respectively, are the (scaled) absolute difference between sample means:

$$\xi_2(n) = \frac{1}{\sqrt{\frac{1}{M_1} + \frac{1}{M_2}}} \left| \bar{y}_{1\cdot}(n) - \bar{y}_{2\cdot}(n) \right|, \qquad (1)$$

where,

$$\bar{y}_{t\cdot}(n) = \frac{1}{M_t} \sum_{m=1}^{M_t} y_{tm}(n)$$

and the pooled sample standard deviation:

$$\xi_1(n) = \sqrt{\frac{(M_1 - 1)\sigma_1^2(n) + (M_2 - 1)\sigma_2^2(n)}{(M_1 - 1) + (M_2 - 1)}} \qquad (2)$$

where,

$$\sigma_t^2(n) = \frac{1}{M_t - 1} \sum_{m=1}^{M_t} \left(y_{tm}(n) - \bar{y}_{t.}(n)\right)^2 .$$

The simple paired t-test [2] can be used to separate the populations by thresholding the ratio of the two dispersion measures:

$$T_{\mathrm{pt}}(n) = \frac{\xi_2(n)}{\xi_1(n)} \quad \overset{S}{\underset{\overline{S}}{\gtrless}} \quad \eta_T \tag{3}$$

where η_T is a user-specified threshold. Here S refers to selecting gene n while $\overline{S}$ refers to rejecting gene n. If the user wishes to constrain familywise error rate or false discovery rate then η_T is chosen as a function of the quantiles of the student-t density with $M_1 + M_2 - 2$ degrees of freedom. Alternatively, if the user wishes to select a fixed number p of genes for further study, e.g., by RT-PCR, then η_T is data-dependent. Specifically, in this latter case one reduces η_T until the number card$\{n : T_{\mathrm{pt}}(n) > \eta_T\}$ is equal to p, i.e. exactly p genes have $T_{\mathrm{pt}}(n)$ values that exceed η_T.

The test statistic $T_{\mathrm{pt}}(n)$ in (3) is a scalar criterion that could be used to rank the genes in decreasing order of T_{pt}, or, equivalently, in increasing p-value.

3.1 Multiple Objective Ranking

Multiple objective optimization captures the intrinsic compromises among possibly conflicting objectives. To illustrate, in the present context we consider the pair of criteria $\xi_2(n)$ (1) and $\xi_1(n)$ (2). A gene that maximizes ξ_2 and minimizes ξ_1 over all genes would be a very attractive gene indeed. Unfortunately, such an extreme of optimality is seldom attained with multiple criteria. A more common case is illustrated in Fig. 1.a. It should be obvious to the reader that gene A is “better” than gene C because both criteria are higher for A than for C. However it is not as straightforward to specify a preference between A, B and D. Multi-criteria ranking uses the “non-dominated” property as a way to establish such a preference relation. A and B are said to be non-dominated because improvement of one criterion in going from A to B corresponds to degradation of the other criterion. All the genes which are non-dominated constitute a curve which is called the Pareto front. A second Pareto front is obtained by stripping off points on the first front and computing the Pareto front of the remaining points. This process can be repeated to define a third front and so on. A gene that lies on the k-th Pareto front will be said to be at ”Pareto depth” k.

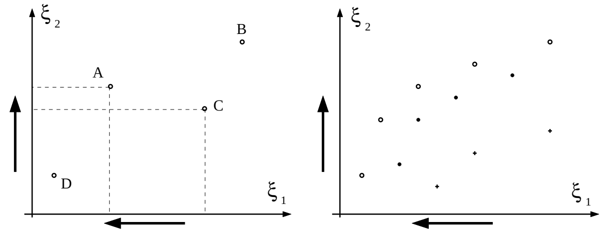

Figure 1: *a). A, B, C are non-dominated genes in the dual criteria plane where ξ_1 is to be minimized and ξ_2 is to be maximized. Genes A, B, and C are at Pareto depth 1 while gene C is at Pareto depth 2. b). Successive Pareto fronts in dual criteria plane (o : first Pareto front, * : second Pareto front, + : third Pareto front).*

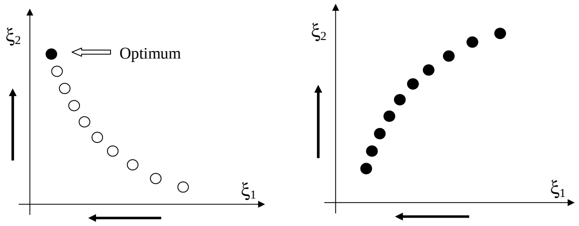

Figure 2: *a). Pareto front contains a single gene, b). Pareto front contains all genes.*

In rare cases the Pareto front consists of a single gene (see Fig. 2.a). At the opposite extreme, there are cases where the Pareto front consists of the entire set of genes (see Fig. 2.b). It can be shown that as the number of criteria increases the Pareto front becomes less and less discriminatory, e.g. for an infinite number of criteria it consists of the entire set of genes. In practical cases where only a few criteria are used there are multiple Pareto fronts each consisting of many genes. We illustrate in Fig. 3 where we show the scatterplot of the criteria $\{(\xi_1(n), \xi_2(n))\}_{n=1}^N$ defined in (1) and (2) for all genes probe responses extracted from microarrays in the mouse retina aging experiment. As in [11] we call this scatterplot the multicriterion scattergram. For this set of data, Fig. 3 shows the first Pareto front as lying on the left-upper boundary of the multicriterion scattergram.

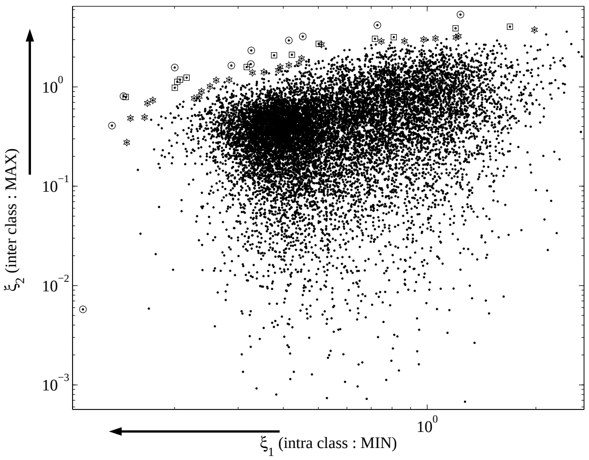

Figure 3: *The multicriterion scattergram for the mouse retina aging experiment. Each point in the scatterplot corresponds to the pair $(\xi_1(n), \xi_2(n))$ for a particular gene n. The first three Pareto fronts are indicated ($\circ$, $\Box$ and$\star$).*

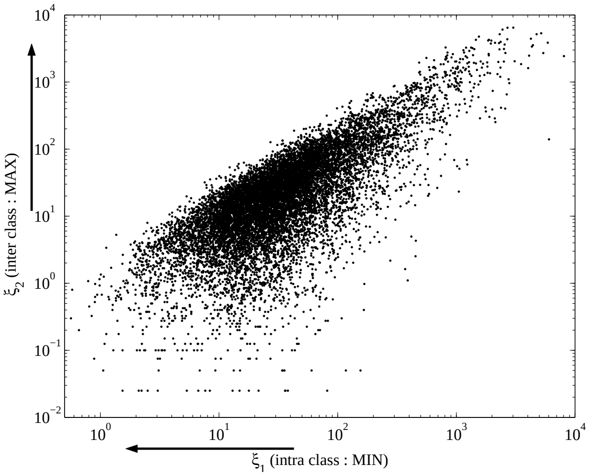

Figure 4: *The multicriterion scattergram for the human retina aging experiment. Each point in the scatterplot corresponds to the pair $(\xi_1(n), \xi_2(n))$ for a particular gene n.*

For comparison, in Fig. 4 we show the multicriterion scattergram for the human retina aging data set with the same pair of criteria ξ_1, ξ_2 defined in (1) and (2). As the upper left boundary of this scatterplot is much shallower and denser than the scatterplot in Fig. 3 the first Pareto front of the human data contains many more genes than the first Pareto front of the mouse data. Since they would not render well in this densely populated boundary region, the Pareto fronts are not indicated in Fig. 4.

3.2 Pareto Depth Sampling Distribution

Microarray data are strongly corrupted by biological variations and measurement variations. To account for this variation we applied a simple resampling procedure to robustify the Pareto analysis. This resampling is implemented as a bootstrap procedure and is equivalent to leave-one-out cross-validation [13]. Resampling proceeds as follows: for each time point a sample is omitted leaving 2^M sets of $(M-1)^2$ pairs to be tested (here we set $M_t = M$, corresponding to the two data sets presented above). For each of these resampled set of genes the Pareto fronts are computed. The most resistant genes are those which remain on the top Pareto fronts throughout the resampling process. To quantify the movement of a given gene across the Pareto fronts we introduce the Pareto depth sampling distribution (PDSD). For each gene this distribution corresponds to the empirical distribution of the 2^M Pareto front indexes visited during the resampling process:

$$\mathrm{Pdsd}_n(k) = \frac{1}{M_{\mathrm{resamp}}} \sum_{j=1}^{M_{\mathrm{resamp}}} 1_n(j,k), k = 1, \ldots, N$$

where $M_{\mathrm{resamp}} = 2^M$ is the number of resampling trials, and $1_n(j,k)$ is an indicator function of the event: "j-th resampling of n-th gene is on k-th Pareto front." If K is the total number of Pareto fronts in the scattergram $(\xi_1(n), \xi_2(n)\}_{n=1}^N$ then, by convention, we define $\mathrm{Pdsd}_n(k) = 0$ for $k > K$. As the PDSD is a probability distribution $\mathrm{Pdsd}_n(k) \geq 0$ and $\sum_k \mathrm{Pdsd}_n(k) = 1$.

Figure 5 corresponds to the (un-normalized) PDSDs over the first 40 Pareto depths for four different genes taken from the human data set under the dual criteria (ξ_1, ξ_2) of (1) and (2). The highly concentrated PDSD in the top-left panel indicates that this gene is very stable; it remains on the first front throughout the resampling process. At the opposite extreme, the highly dispersed PDSD on the bottom-right panel indicates a very unstable gene; its Pareto depth is highly sensitive to resampling.

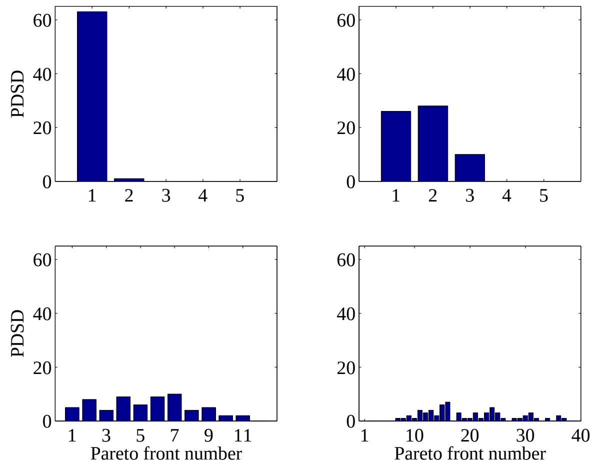

Figure 5: *Unormalized PDSDs for four different genes taken from human retina experiment. These PDSDs are indexed by the Pareto depth, which is equivalent to Pareto front number.*

The two other panels depict PDSD's of genes that lie within these two extremes. as the PDSD summarizes all of the empirical Pareto depth statistics it can be used to develop a wide array of gene ranking criteria. For example, in [8] we ranked genes in terms of the proportion of resampling trials for which a gene remained on one of the top 3 Pareto fronts. This ranking criterion is equivalent to the cumulative Pareto front test

$$T_{\text{cum}}(n)=\sum_{k=1}^{3}\text{Pdsd}_n(k)\ \begin{array}{c}S\\ >\\ <\\ \overline{S}\end{array}\ \eta_c. \qquad (4)$$

In this paper we investigate a different PDSD ranking statistic for pulling out genes that are both highly stable and have low Pareto depth. Genes with these attributes can be captured by requiring that their Pareto depth variance $\sigma^2(n)$ and Pareto depth mean squared $m^2(n)$ be small. Equivalently, we define the Pareto depth test

$$T_{\text{pd}}(n)=\sqrt{m^2(n)+\sigma^2(n)}\ \begin{array}{c}S\\ >\\ <\\ \overline{S}\end{array}\ \eta_d. \qquad (5)$$

Note that the test statistic $T_{\text{pd}}(n)$ is equivalently expressed as $T_{\text{pd}}(n)=\sum_k k^2\text{Pdsd}_n(k)$.

Figure 6 is the scatter plot of the pairs of moments $\{(m^2(n),\sigma^2(n))\}_{n=1}^{N}$ of the gene PDSDs for the human retina data. The best genes are those which have smallest mean and variance, i.e., the genes that lie on the lower left corner of the scatter plot. For a given threshold η_d the test (5) defines a quarter disk region in the plane of Fig. 6 centered at the origin $(0,0)$ with disk radius η_d. Genes whose moment pair $(m^2(n),\sigma^2(n))$ falls in this region will pass the test and be selected as having both low and stable Pareto depths. Bootstrap methods, implemented with random permutation and resampling, could be straightforwardly implemented to determine the p-values of this test. However, in this paper we will focus on constraining the number of discovered genes as opposed to the level of significance of the test. When the number of discovered genes is constrained to be 50, the top ranked 50 genes fall into the acceptance region of the test (5). Figure 7 shows a gray-coded image of the PDSDs for each of these top 50 genes for the human retina data. The figure indicates that the Pareto depths of these 50 genes are tightly concentrated in the range 1 to 6.

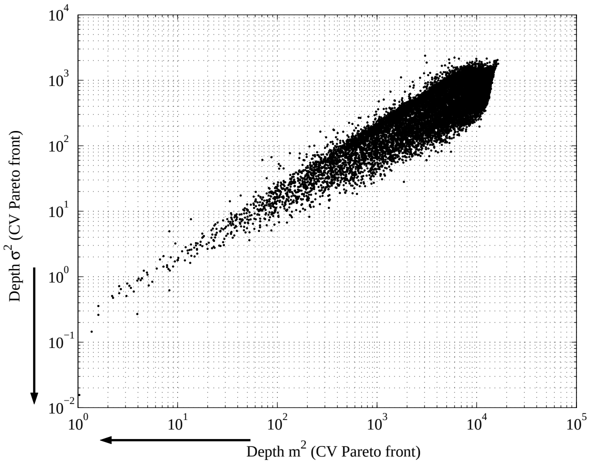

Figure 6: *The scatter plot of the square mean (horizontal axis) and the variance (vertical axis) of the Pareto depth of each gene for human retinal data. Here CV refers to our resampling method consisting of leave-one out cross-validation.*

For comparison, Fig. 8 shows the PDSDs obtained by applying exactly the same selection criterion (5) to the mouse aging experiment as we just presented for the human aging experiment. Notice that the PDSDs for the top 50 mouse genes are spread over 16 or more Pareto depths. This high spread is reflects the fact that there are fewer stable Pareto dominant genes in the mouse aging experiment as compared to the human aging experiment.

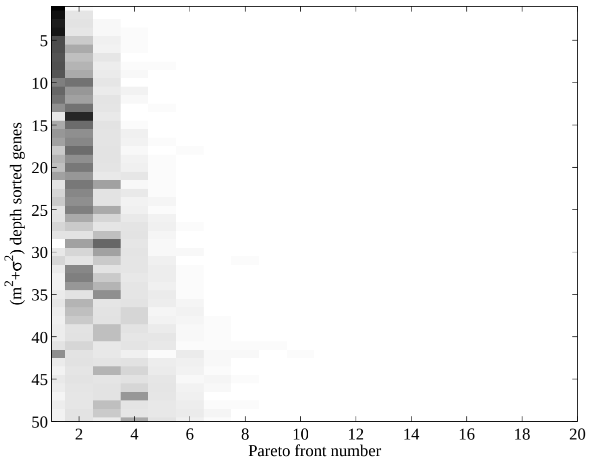

Figure 7: *The PDSDs of the 50 top human genes discovered using the test (5) applied to the scatter plot of Fig. 6 with threshold η_d determined such that exactly 50 genes fall into acceptance region. The magnitude of the PDSD is encoded in the false color range of black (PDSD=1) to white ($PDSD = 0$).*

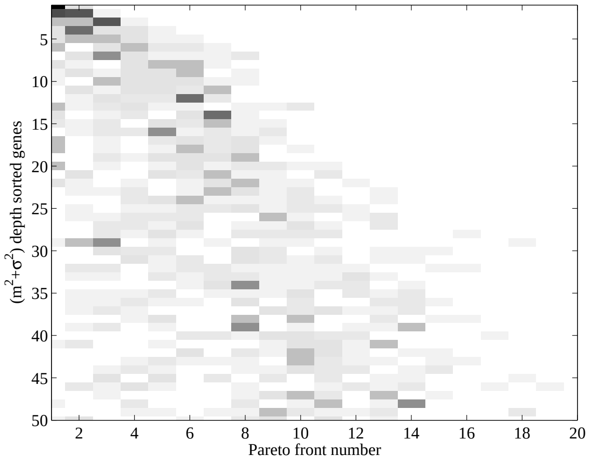

Figure 8: *The PDSDs of the 50 top mouse genes discovered using the test ((5). As compared to the human genes Fig. 7 there is much higher variability in the Pareto depths of the top 50.*

4 Experimental Comparisons

Here we compare the paired t test (3), the cumulative Pareto front test (4) used in [8], and the Pareto depth test (5) on the basis of their gene ranking performance for the retinal aging experimental data and for simulated data.

4.1 Experimental Data

Figures 9 and 10 show the number of genes discovered as a function of the paired t test threshold η_T for the experimental human and mouse data, respectively. The shapes of the curves in these two figures are substantially different. Indeed the distinctive plateau at the right tail of Fig. 10 is due to the existence of several mouse genes whose best scores $(\xi_1(n), \xi_2(n))$ are well detached from the scores of the rest of the genes. There are no such highly detached human genes as can be seen by comparing the multicriteria scattergrams of Figs. 3 and 4. Figures 11 and 12 show the number of genes discovered as a function of the inverse Pareto depth test threshold $1/\eta_d$ for the experimental human and mouse data, respectively. Again the shapes of the curves in these two figures are substantially different. As compared to the paired t-test figures, Figs. 9 and 10, the increase in the number of genes discovered by the Pareto depth test is much more gradual as $1/\eta_p$ decreases. This suggests that the Pareto depth test better discriminates between its highly ranked genes.

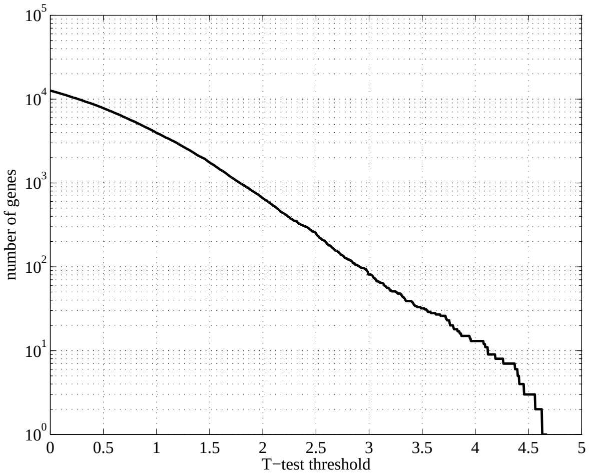

Figure 9: *Number of t-test-extracted genes as a function of threshold η_T for data in human retina aging study.*

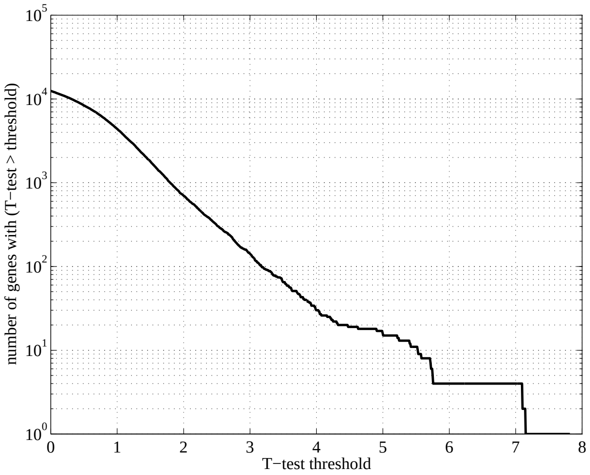

Figure 10: *Number of t-test-extracted genes as a function of threshold η_T for data in mouse retina aging study.*

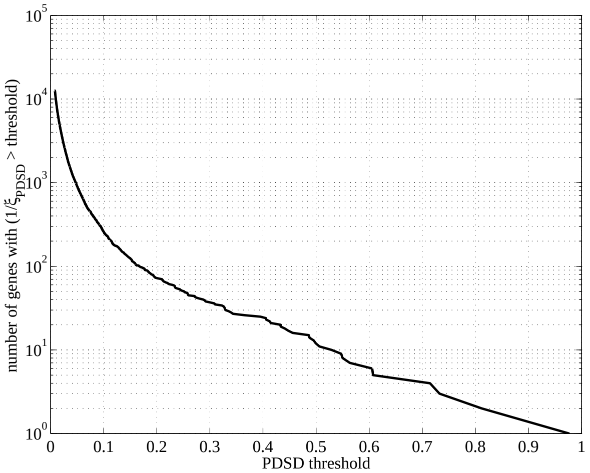

Figure 11: *Number of Pareto-depth-test-extracted genes as a function of inverse threshold $1/\eta_d$ for data in human retina aging study.*

4.2 Simulated Data

We performed a limited set of simulations to be able to compare estimated rankings to the "ground truth" true rankings. The simulations were designed to be representative of gene expressions in a typical gene microarray experiment. Three hundred ($N = 300$) different probe responses were simulated. Eight ($M = 8$) replicates of the n-th gene probe response were generated according to an i.i.d. Gaussian distribution with means and variances given by $(m_1(n), \sigma_1^2(n))$ and $(m_2(n), \sigma_2^2(n))$ for populations 1 and 2, respectively. The variances were made equal $\sigma_1^2(n) = \sigma_2^2(n) = \sigma^2(n)$ over both populations. The means and variances were set by the following formula:

$$\sigma(n) = \xi_2(n), \;\; m_1(n) = 0, \;\; m_2(n) = \xi_1(n)\xi_2(n)/2$$

where the values of $\xi_1(n), \xi_2(n)$ are indicated by the criteria structure illustrated in Fig. 13. The ground truth ranking of all genes is determined by this figure which can be viewed as the ensemble mean scattergram. We designate the 90 genes on the first 3 fronts of Fig. 13 (depth increasing along -45^o diagonal) as *ground-truth-optimal* genes.

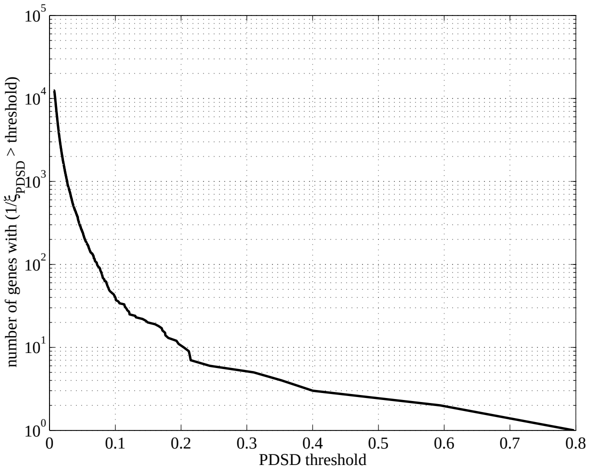

Figure 12: *Number of Pareto-depth-test-extracted genes as a function of inverse threshold $1/\eta_d$ for data in mouse retina aging study.*

Figure 14 shows a realization of the empirical scattergram obtained from sample mean and variance estimates derived from the replicates. Figure 15 shows the three first Pareto fronts and the boundaries of two acceptance regions for the paired t test applied to the empirical scattergram of Fig. 14. The first three Pareto fronts do not capture all of the ground-truth-optimal genes but they have a very low (0%) false discovery rate (proportion of genes found which are not ground-truth-optimal). The solid line boundary of the paired t test corresponds to a threshold η_T which discovers the 90 genes with highest $T_{\rm pt}(n)$ value. Use of this acceptance region would result in discovery of more ground-truth-optimal genes than discovered by the first three Pareto fronts, but with a false discovery rate of approximately 15%. The dashed line

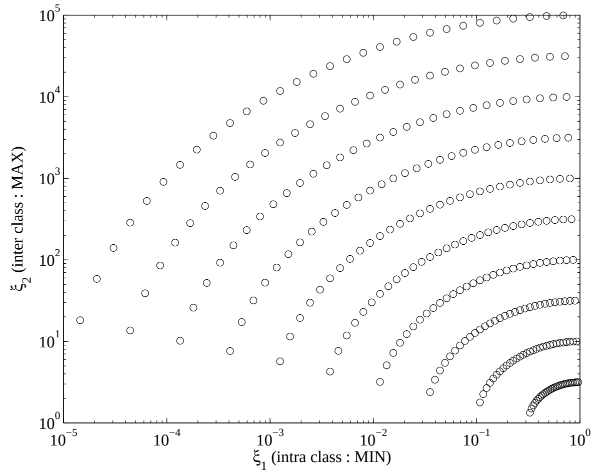

Figure 13: *Ensemble mean scattergram (ground truth) for simulation study. There are 30 genes represented in each of the 10 semicircles. Ground truth Pareto optimal genes lie on the outermost front of lowest curvature.*

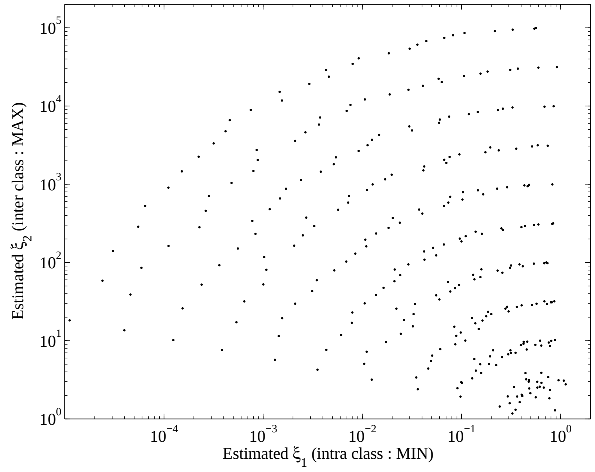

Figure 14: *Empirical scattergram constructed from estimating sample mean and variance from the $M = 8$ i.i.d. samples of each gene.*

boundary corresponds to a paired t test threshold η_T which would lead to discovery of all of the 90 ground-truth-optimal genes. However, the false discovery rate of this acceptance region is quite high ($> 40\%$). In Fig. 16 is another depiction of the two acceptance regions of the paired t test. It is clear from this example that neither the paired t-test nor the cumulative Pareto front test (4) succeed in extracting all the ground-truth-optimal genes with low false discovery rate. The next question we address is: would a Pareto depth test do better than the simple three front Pareto test and paired t tests illustrated here?

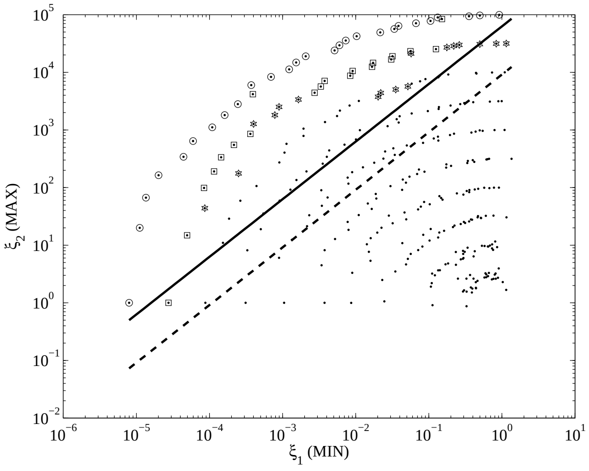

Figure 15: *Three first Pareto fronts (○, □ and $*$) and boundaries of paired t test acceptance ragions for the scattergram of Fig. 14.*

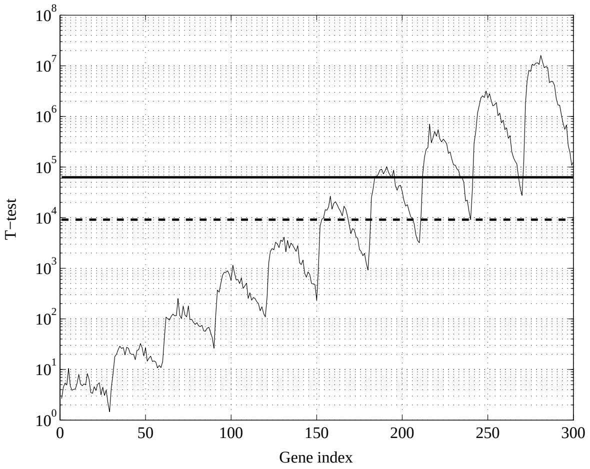

Figure 16: *Paired t test statistic and thresholds corresponding to boundaries in Fig. 15. Genes are ordered from right to left by scanning successive fronts in the ensemble mean scattergram of Fig. 13.*

To quantify the tradeoffs between the paired t test and the Pareto tests for extracting the ground-truth-optimal genes we performed a representative simulation study to compute the average correct discovery rates and the average false discovery rates as a function of the number M of replicates. All tests were implemented with a data dependent threshold which selected the 90 top genes as ranked by the respective test statistics. For the range of M studied this threshold setting gave the paired t test a nearly constant correct discovery rate of approximately 88%. The cumulative Pareto front test (4) and the Pareto depth test (5) were implemented for comparison.

In Figs. 17 and 18 we plot the correct discovery rate and the false discovery rate, respectively, for the paired t-test and the cumulative Pareto front test. From the figures it is clear that the cumulative Pareto front test has better performance than the paired t test for large M. However, it suffers from lower correct discovery rate than the paired t test for small M. In Figs. 19 and 20 the same error rates are compared for the paired t-test and the Pareto depth test. The Pareto depth test performed significantly better (higher correct discovery rate and lower false discovery rate) than the paired t test for all M.

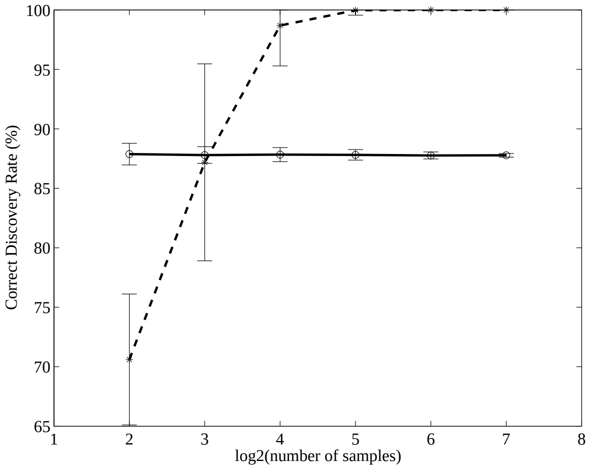

Figure 17: *Correct discovery rate as a function of the number of replicates for paired t-test (solid) versus cumulative Pareto front test (dashed). Both tests have data-dependent thresholds that select the* 90 *top ranked genes according to their respective test statistics.)*

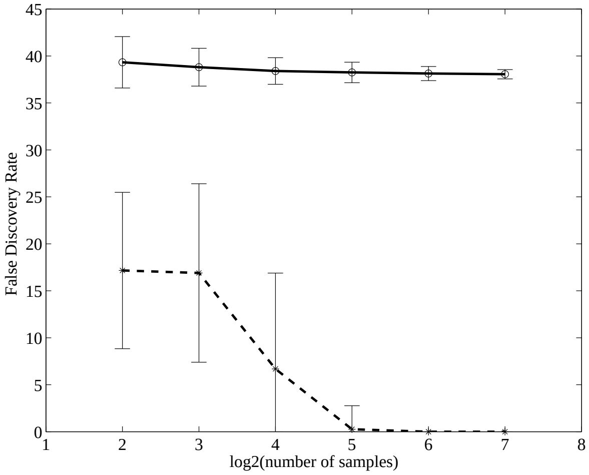

Figure 18: *False discovery rate as a function of the number of replicates for paired t-test (solid) versus cumulative Pareto front test (dashed). Both tests have data-dependent thresholds that select the* 90 *top ranked genes according to their respective test statistics.*

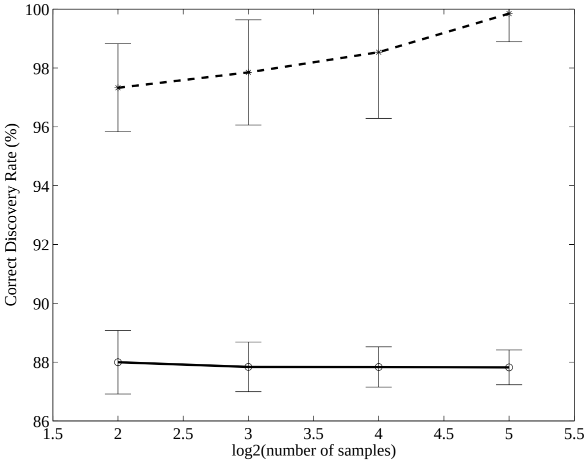

Figure 19: *Correct discovery rate as a function of the number of replicates for paired t-test (solid) versus Pareto depth test (dashed). Both tests have data-dependent thresholds that select the* 90 *top ranked genes according to their respective test statistics.*

The alert reader will realize that our definition of ground-truth-optimal genes favors the Pareto methods of gene ranking and selection as compared to the paired t methods. Our definition of ground-truth-optimality was motivated by our several years of experience helping molecular biologists discover biologically interesting genes, in particular genes

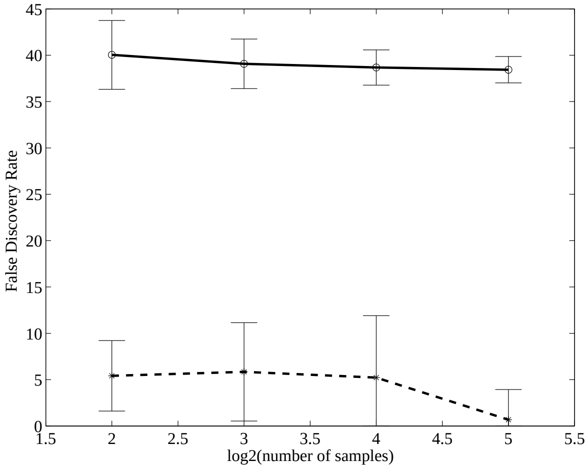

Figure 20: *False discovery rate as a function of the number of replicates for paired t-test (solid) versus Pareto depth test (dashed). Both tests have data-dependent thresholds that select the* 90 *top ranked genes according to their respective test statistics.*

with weak but interesting transcription factors. A more comprehensive study would compare the performance of Pareto to paired t approaches when the ground-truth-optimal genes are defined differently. Due to space limitations we do not present the results of this study here.

5 Conclusion

DNA microarray technology allows one to evaluate the expression profile of thousands of genes simultaneously. However, to take full advantage of these powerful tools, we need to find new methods to handle large amounts of data and information without becoming overwhelmed by the potentially large number of candidate genes. This paper has presented a new method of Pareto analysis that can identify and rank genes that have both stable and low Pareto depths relative to the remaining genes. Additional genes discovered using this algorithm are now being validated by RT-PCR methods. Many signal processing challenges remain due to the increasingly high dimensionality of genetic data sets. The developed method has been implemented in matlab and C and is sufficiently fast to be part of an interactive tool for gene screening, ranking, and clustering.

Acknowledgements

The authors would like to thank Prof. A. Swaroop, Dr. S. Yosida, R. Farjo and A. Mears in the Departement of Human Genetics at University of Michigan for their collaboration.

References

[1] 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.

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

[3] 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.

[4] 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.

[5] J. DeRisi, V. Iyer, and P. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, no. 5338, pp. 680–686, Oct 24 1997.

[6] D. Donoho and M. Gasko, "Breakdown properties of location estimates based on halfspace depth and projected outlyingness," *Annals of Statistics*, vol. 4, pp. 1803–1827, 1992.

[7] P. Fitch and B. Sokhansanj, "Genomic engineering: moving beyond DNA sequence to function," *IEEE Proceedings*, vol. 88, no. 12, pp. 1949–1971, Dec 2000.

[8] 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.

[9] 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.

[10] 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.

[11] A. Hero and G. Fleury, "Pareto-optimal methods for gene analysis," *Journ. of VLSI Signal Processing, Special Issue on Genomic Signal Processing*, vol. to appear, , 2003. `www.eecs.umich.edu/~hero/bioinfo.html`.

[12] A. Hero, G. Fleury, and S. Cerbourg, "Multicriteria gene screening for microarray experiments," *EURASIP Journ. of Applied Signal Processing*, p. in revision, 2003.

[13] J. U. Hjorth, *Computer intensive statistical methods*, CHapman and Hall, London, 1994.

[14] 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.

[15] 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, 2000. `citeseer.nj.nec.com/414709.html`.

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

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

[18] D. Lockhart, H. Dong, M. Byrne, M. Follettie, M. Gallo, M. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. Brown, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nat. Biotechnol.*, vol. 14, no. 13, pp. 1675–80, 1996.

[19] M. Marra and *etal*, "The genome sequence of the SARS-associated coronavirus," *Science Express*, vol. 10.1126, , May 1 2003. `www.scienceecpress.org`.

[20] 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`.

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

[22] J. Tukey, *Exploratory Data Analysis*, Wiley, NY NY, 1977.

[23] 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.

[24] J. Watson and A. Berry, *DNA: The secret of life*, Alfred A. Knopf, 2003.

[25] 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.