

Gene Selection Criterion for Discriminant Microarray Data Analysis Based on Extreme Value Distributions

Wentian Li*

Center for Genomics and Human Genetics
North Shore LIJ Research Institute

Ivo Grosse†

Cold Spring Harbor Laboratory

ABSTRACT

An important issue commonly encountered in the analysis of microarray data is to decide which and how many genes should be selected for further studies. For discriminant microarray data analyses based on statistical models, such as the logistic regression model, this gene selection can be accomplished by a comparison of the maximum likelihood of the model given the real data, $\hat{L}(D|M)$, and the expected maximum likelihood of the model given an ensemble of surrogate data, $\hat{L}(D_0|M)$. Typically, the computational burden for obtaining $\hat{L}(D_0|M)$ is immense, often exceeding the limits of available resources by orders of magnitude. Here, we propose an approach that circumvents such heavy computations by mapping the simulation problem to an extreme value problem, which can be easily solved by numerical simulation. We choose three classification problems from two publicly available microarray datasets to illustrate that approach.

Categories and Subject Descriptors

J.3 [Life and medical sciences]: Biology and genetics

General Terms

Performance

Keywords

microarray, classification, logistic regression, extreme values

*Corresponding author: Center for Genomics and Human Genetics, North Shore LIJ Research Institute, 350 Community Drive, Manhasset, NY 11030, USA. FAX: 516-562-1153. email: wli@nslj-genetics.org

†Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. email: grosse@cshl.edu

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

RECOMB'03 April 10-13, 2003, Berlin, Germany.

Copyright 2003 ACM 1-58113-635-8/03/0004 ...\$5.00.

1. INTRODUCTION

Discriminant microarray data analysis can be understood as a comparison and classification of the expression levels of samples from one group versus another group, such as diseased tissues versus normal tissues, or one subtype of cancer versus another subtype. Discriminant analysis or classification can be carried out on a whole set of genes or on individual genes, and it has become increasingly clear that, for many classification tasks based on microarray data, it is not necessary to consider many genes simultaneously. In many cases it has been shown that a few genes are sufficient for classifying two groups of samples [1, 6, 21, 24, 27, 28, 30, 32, 39], and in some cases as few as one or two genes are sufficient for a perfect classification [24, 35, 39]. Based on this understanding, a simple approach is to examine one gene at a time, rank them according to their classification ability, and select only the top ranking genes for further studies, including new confirmation experiments [36, 33].

Two single-gene discriminant methods that were first applied to the analysis of microarray data are the fold-change method (see, e.g., [5]) and the *t*-test. As repeatedly pointed out [4, 7, 10, 20, 29, 31, 37], the fold-change method is not rigorous from the statistical point of view, because it considers neither the variances nor the sample sizes of the data. For example, a two-fold increase obtained from narrowly distributed data with 1000 samples is statistically much more significant than the same increase obtained from broadly distributed data with 10 samples. The *t*-test overcomes this shortcoming by including the variance and sample size information in making a statistical conclusion. However, the *t* distribution is obtained by assuming that the random variables are sampled from a normal (gaussian) distribution. The assumption of normally distributed random variables is usually satisfied by using the logarithm of the spot intensity in a microarray experiment as a measure of the expression level (see, e.g., [13]).

There are alternative discriminant methods that do not require normal distributions. Out of the four linear discriminant analysis methods—Fisher's linear discriminant analysis, logistic regression (LR), Rosenblatt's perceptron, and support vector machine (SVM)—LR and SVM do not rely on the assumption of normally distributed random variables [19], and hence they are more robust when the actual data are not normally distributed, including the presence of outliers. Another difference between *t*-tests and LRs is that *t*-tests compare two group averages, whereas LRs check each

individual sample for consistent differential expressions [23]. In the following we focus on LR, which has already been used in discriminant microarray data analyses [12, 24, 25, 30, 34, 38].

Two different schemes in the LR framework to decide whether the discriminant ability of a gene is significant are (i) cross-validation and (ii) resampling. In cross-validation, the set of samples is divided into two parts, where the first part is used for fitting the model parameters, and the second part is used for assessing the classification performance [1, 11]. The shortcoming of this method is that not all samples are used in the learning process, which is not optimal for datasets with a small number of samples. Often, the number of samples is small in microarray datasets, partly due to the fact that microarray chips are expensive. In resampling, the sample labels are randomly permuted, and the entire analysis is re-run for the shuffled dataset, multiple times. Comparing likelihoods of single-gene LRs of the real data with those of the shuffled data provides a way to select top-ranking genes [25]. One problem of the resampling scheme is that the calculation of the LR likelihoods for ten-thousands of genes is computationally intensive, and that repeating that calculation for, say, 10^4 sets of shuffled data is prohibitive.

Here we propose an alternative gene selection criterion that circumvents such heavy computations by performing some of the calculations analytically. Our approach is based on the recognition that we are only interested in the extreme values in the following sense: in order to define a threshold for gene selection, we compare the maximum likelihood of genes in the real data with the maximum likelihood of the top-ranking gene in the surrogate data. Whereas traditional approaches require the calculation of all single-gene likelihoods in the surrogate data for each of the surrogate datasets, we propose to compute (the expected value of) the likelihood of the top-ranking gene from the null data directly. We note in passing that the well known extreme value distributions, such as the Gumbel, Fréchet, or Weibull distributions [18], do not provide a solution to our problem, because these generic distributions are the limiting distributions for infinitely large sample sizes, whereas we are interested in the extreme value distribution for a *finite* sample size given by a finite number of genes.

2. METHODS

2.1 Logistic regression of microarray data

First, we introduce the following notation. Let the samples be indexed by i , and let the genes be indexed by j . Denote the total number of samples by N , the total number of genes by p , the expression level by x , e.g., $x = \log(\text{spot intensity})$, and the sample label value by y , e.g., $y = 0$ or $y = 1$ for a binary classification problem. Then, the single-gene LR model M_j of gene j is defined by the conditional probabilities of the sample label y_i given the expression levels x_{ij} :

$$M_j : \quad P(y_i = 1|x_{ij}) = \frac{1}{1 + e^{-a_j - b_j x_{ij}}} \quad (1)$$

for $i = 1, 2, \dots, N$ and $j = 1, 2, \dots, p$. Here, a_j and b_j are parameters to be estimated from all samples $i = 1, 2, \dots, N$. The data-fitting performance of M_j is measured by the maximum likelihood

$$\hat{L}_j(D|M) = \max_{a_j, b_j} \prod_{i=1}^N [P(y_i = 1|x_{ij})]^{y_i} [1 - P(y_i = 1|x_{ij})]^{1-y_i}, \quad (2)$$

where D denotes the data, and M denotes the LR model in Eq. (1).

2.2 Maximum likelihood for the shuffled-label data

Denote by D_0 a dataset with shuffled sample labels, and by $\hat{L}_j(D_0|M)$ the maximum likelihood under the single-gene LR model M_j . For a particular realization of the shuffled data, we define by

$$l \equiv \max_j \hat{L}_j(D_0|M) \quad (3)$$

the maximum value of the maximum likelihoods among all genes. Note the two maximizations: the first is over the parameter values for a given gene j , and the second is over all genes j . When D_0 is repeatedly generated, those maximum values l vary from realization to realization, and our goal is to characterize the distribution of l , e.g. by computing the expected value, the median, or the standard deviation of l .

Toward the calculation of the expected value of l , we use the fact that, under the assumption that the null model M_0 is the true model of the data, the asymptotic distribution of the (logarithm, and multiplied by a factor of 2) ratio of two maximum likelihoods is a χ^2 distribution with df degrees of freedom, where $df = d(M) - d(M_0)$ is the difference of the number of parameters in models M and M_0 [8], i.e.,

$$2 \log \hat{L}_j(D_0|M) = 2 \log \hat{L}_j(D_0|M_0) + t(\chi_{df}^2), \quad (4)$$

where $t(\chi_{df}^2)$ denotes a random variable sampled from a χ^2 distribution with df degrees of freedom.

Denote by M_0 the model that is the same for all genes, i.e., $P(y_i = 1|x_{ij}) = c$ for all $j = 1, 2, \dots, p$. The maximum likelihood estimate of c is the percentage of samples that are labeled as 1, i.e., $\hat{c} \equiv N_1/N$. The maximum likelihood under M_0 is

$$\hat{L}(D_0|M_0) = \hat{c}^{N_1} (1 - \hat{c})^{N - N_1}, \quad (5)$$

and its logarithm is related to the entropy

$$H \equiv -\frac{N_1}{N} \log \frac{N_1}{N} - \frac{N - N_1}{N} \log \frac{N - N_1}{N} \quad (6)$$

by $\log \hat{L}(D_0|M_0) = -NH$. Note that $\hat{L}(D|M_0) = \hat{L}(D_0|M_0)$, because the percentage N_1/N of samples with sample label $y = 1$ is the same in D and D_0 .

Applying the LR model to the shuffled-label surrogate data, we obtain for the best single-gene maximum (log) likelihood

$$\begin{aligned} \log l &= \max_j \log \hat{L}_j(D_0|M) \\ &= \max_j (\log \hat{L}_j(D_0|M_0) + t(\chi_{df}^2)/2) \\ &= -NH + \frac{1}{2} \max(t_1, t_2, \dots, t_p), \end{aligned} \quad (7)$$

where t_1, t_2, \dots, t_p are p random variables sampled from a χ^2 distribution with df degrees of freedom. In this example, the single-gene LR model contains two parameters, and M_0 contains one parameter, so $df = 2 - 1 = 1$.

2.3 Extreme value distribution

Define $T_p \equiv \max(t_1, t_2, \dots, t_p)$, where t_1, t_2, \dots, t_p are statistically independent and identically distributed (i.i.d.) random variables. Then, the cumulative distribution function of T_p , $P(T_p < x)$, is related to the cumulative distribution function of t , $Q(t < x)$, by

$$\begin{aligned} P(T_p < x) &= Q(t_1 < x)Q(t_2 < x) \cdots Q(t_p < x) \\ &= [Q(t < x)]^p \\ &= \left[\int_0^x q(t) dt \right]^p, \end{aligned} \quad (8)$$

where $q(t)$ denotes the probability density function of t . If $q(t)$ is uniform or exponential, the integral on the r.h.s. of Eq. (8) can be computed analytically and expressed in closed form, allowing a closed-form expression for the probability density function of T_p , $f(T_p) = dP(T_p < x)/dx$ [3, 9, 15]. In those cases the expected value of T_p , $E[T_p]$, as well as its standard deviation, $\sigma[T_p]$, can be calculated analytically.

However, if the random variables t are sampled from a χ_{df}^2 distribution with $df = 1$ degree of freedom, we do not have a closed form expression of $f(T_p)$ or $E[T_p]$. Hence, we perform numerical simulations to obtain approximations of $E[T_p]$ and $\sigma[T_p]$ for p ranging from 10^3 to 1.5×10^5 .

2.4 Gene selection by the extreme value distribution

The gene selection criterion can be set by requiring the maximum likelihood of gene j (maximized over the parameter values in the LR model) to be greater than the average of the greatest maximum likelihood from the shuffled data. Here, ‘‘the average of the greatest’’ means that for each shuffled dataset we choose the greatest maximum likelihood among the maximum likelihoods of all genes j , and then we take the average over many shuffled datasets. Then, one may select all genes k that satisfy

$$2 \log \hat{L}_k(D|M) > 2E[\max_j \log \hat{L}_j(D_0|M)] = -2NH + E[T_p]. \quad (9)$$

If a more stringent criterion is required, one may add one standard deviation to the r.h.s. of Eq. (9), i.e., one may select all genes k that satisfy

$$2 \log \hat{L}_k(D|M) > -2NH + E[T_p] + \sigma[T_p]. \quad (10)$$

Eq. (10) is an arbitrary way to modify criterion (9), and we note in passing that there are other arbitrary options to tighten or relax the criterion, such as using the second greatest, the third greatest, or the n -th greatest likelihood instead of the greatest, using a multiple of $E[\max_j \log \hat{L}_j(D_0|M)]$, such as $(1 \pm \alpha)(-2NH + E[T_p])$, and using several standard deviations, such as $-2NH + E[T_p] \pm \beta\sigma[T_p]$.

3. RESULTS

3.1 Numerical simulation of the extreme value distribution

We perform numerical simulations in order to obtain the expected value $E[T_p]$, the median $M[T_p]$, and the standard deviation $\sigma[T_p]$ of T_p as a function of p . For each value of p ranging from 1 to 1.5×10^5 we generate 10^4 samples of p random variates sampled from the $\chi_{df=1}^2$ distribution. In Fig. 1 we show $E[T_p]$, $M[T_p]$, and $\sigma[T_p]$ versus $\log p$. Note that we choose a logarithmic scale for the abscissa p , which denotes the number of genes in our problem. We find from Fig. 1 that the general trends of both $E[T_p]$ versus $\log p$ and $M[T_p]$ versus $\log p$ are approximately linear. This is in agreement with the asymptotic result $E[T_p] = \gamma + \log p$, with $\gamma = 0.5772\dots$ denoting the Euler constant, for random variables sampled from an exponential distribution [14]. We note that $\chi_{df=2}^2$ distribution is in fact an exponential distribution, but this result does not lead to an analytic solution of $E[T_p]$ for $\chi_{df=1}^2$ distribution.

The functional dependence of $E[T_p]$ on p is not exactly logarithmic. We find from Fig. 1 that there is a systematic deviation from the logarithmic trend for $p < 100$. For $p > 1.5 \times 10^5$ we cannot predict whether the logarithmic trend continues asymptotically, but the extrapolation to these ranges of p is irrelevant for our problem, because in typical applications we need to analyze less than 10^5 genes. For the range of $10^3 < p < 1.5 \times 10^5$ we obtain the regression line $E[T_p] \approx -1.14 + 1.89 \log p$.

From Fig. 1 we also find that, for all studied values of p , the median is smaller than the average, but that the difference between the two is small. The fact that $M[T_p] < E[T_p]$ indicates that the distribution of T_p is asymmetric with a longer right tail. We also find from Fig. 1 that the standard deviation of T_p , $\sigma[T_p]$, increases very slowly with p , reaching approximately $\sigma[T_p] \approx 2.43$ for values of $p \approx 10^4$. Again, it is interesting to compare this behavior to the asymptotic result $\sigma[T_p] = \pi/\sqrt{6} \approx 1.28$ obtained for the case that the random variables t are sampled from an exponential distribution ($\chi_{df=2}^2$ is equivalent to an exponential distribution).

3.2 Application to microarray datasets

We use two microarray datasets to illustrate the proposed criterion for deciding how many top-ranking genes should be selected: (i) the leukemia subtype data from the Whitehead Institute [16], and (ii) the colon cancer data from Princeton University [2]. Fig. 2(A) shows the rank-ordered distribution of the maximum likelihoods for all single-gene LR models for the discrimination of acute lymphoblastic leukemia (ALL) from acute myeloid leukemia (AML). The sample size is 72, which combines both the training and testing sets, as designated in Ref. [16]. The ALL-AML classification problem is thoroughly discussed in [26], and it is well known to be an easy classification problem [24, 27, 30, 35]. We find that there are 405 genes in the ALL-AML dataset with a max-log-likelihood that exceeds (the expected value of) the greatest of the maximum likelihood in the shuffled datasets, based on the extreme value calculation presented in Eq. (9).

Extreme values of p chi-square distributed samples

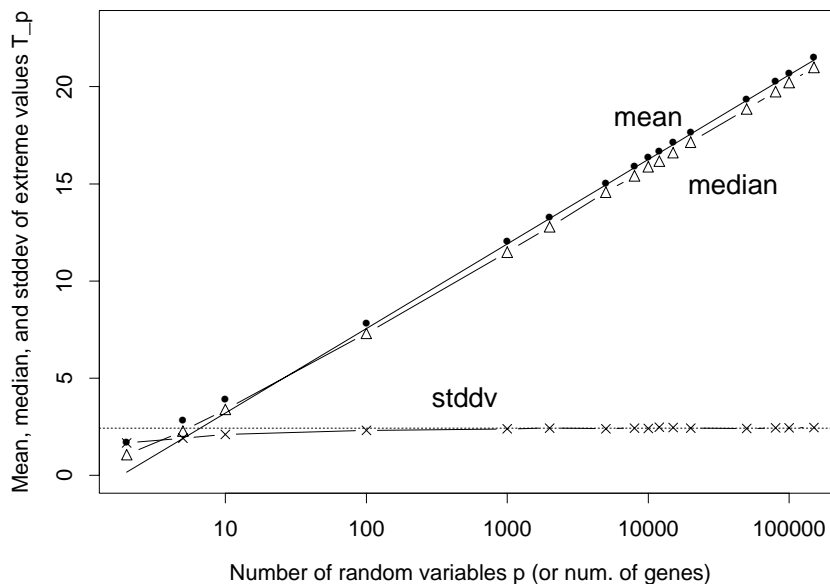


Figure 1: Numerical simulation of the extreme values $T_p = \max(t_1, t_2, \dots, t_p)$ of p random variables t_1, t_2, \dots, t_p sampled from a χ^2 distribution with $df = 1$ degree of freedom. We plot the expected value $E[T_p]$, the median $M[T_p]$, and the standard deviation $\sigma[T_p]$ versus $\log p$ for p ranging from 1 to 1.5×10^5 . For $p > 10^3$ both $E[T_p]$ and $M[T_p]$ show an approximately logarithmic growth with p . For p ranging from 10^3 to 1.5×10^5 we obtain the linear regression $E[T_p] \approx -1.14 + 1.89 \log p$. We also find that, at $p \approx 10^3$, $\sigma[T_p]$ settles at approximately $\sigma[T_p] \approx 2.43$, indicated by the horizontal dashed line.

If we raise the selection threshold from $E[T_p]$ to $E[T_p] + \sigma[T_p]$, according to Eq. (10), then the number of genes is reduced to 310. We note in passing that these two numbers, 405 and 310, are substantially smaller than 1100, which is the number of genes that are “more highly correlated with the AML-ALL class distinction than would be expected by chance,” as reported by [16] using the “neighborhood analysis”.

As pointed out in Ref. [17], the ALL samples of the leukemia datasets are still a heterogeneous dataset, with sources from B-cells and T-cells being different from each other. Fig. 2(B) shows the rank-ordered distribution of the maximum likelihoods using single-gene LR models for the B-cell versus T-cell classification, with a reduced sample size of 47. The criterion of Eq. (9) selects 114 genes, and adding one standard deviation, according to Eq. (10), leads to 89 genes. These findings are in agreement with the observation in Ref. [17] that there are differentially expressed genes in B-cells and T-cells, and with a similar observation in [22] based on cluster analysis.

Fig. 2(C) shows the rank-ordered distribution of the maximum likelihoods using single-gene LR models for the colon cancer versus normal tissue dataset [2]. This dataset consists of 47 samples, and only the data for 2000 genes that have the “highest minimal intensity across the samples” are available [2]. In contrast to the two examples presented above, we find in this example that only 49 and 27 genes are selected by criteria Eq. (9) and Eq. (10), respectively. One possible explanation why these numbers are so small is that the pre-processing methods applied to limit the number of genes to 2000 might have removed some differentially

expressed genes. Another possible explanation is that the colon cancer versus normal samples in this dataset are just harder to classify.

4. CONCLUSIONS

One ubiquitous question arising in discriminant analyses of microarray data is to choose an appropriate number of genes to be selected for further studies. While a too conservative estimate of the number of relevant genes causes some loss of information, a too liberal estimate of the number of relevant genes causes the increase of noise in the resulting dataset. Finding the optimal number of genes, which maximizes the signal to noise ratio in subsequent studies, is a difficult goal, and many approaches have been proposed to accomplish that goal.

Some of the most successful approaches are based on re-sampling schemes, but the price for their reliable output is an enormous computational cost. Often, that cost exceeds by far the available resources, i.e., the available computing power is not sufficient for solving the computational problem in a practical time frame. In order to avoid that computational burden, we propose an approach that compares the maximum likelihoods of all single-gene LR models given the real data with the top-ranking maximum likelihoods of all single-gene LR models given the ensemble of surrogate data where the sample labels are randomly permuted [25].

A naive implementation of that approach would still require a huge load of computations, because the LR param-

leukemia: ALL vs AML(N=72)

ALL T vs B-cell(N=47)

colon cancer vs normal(N=62)

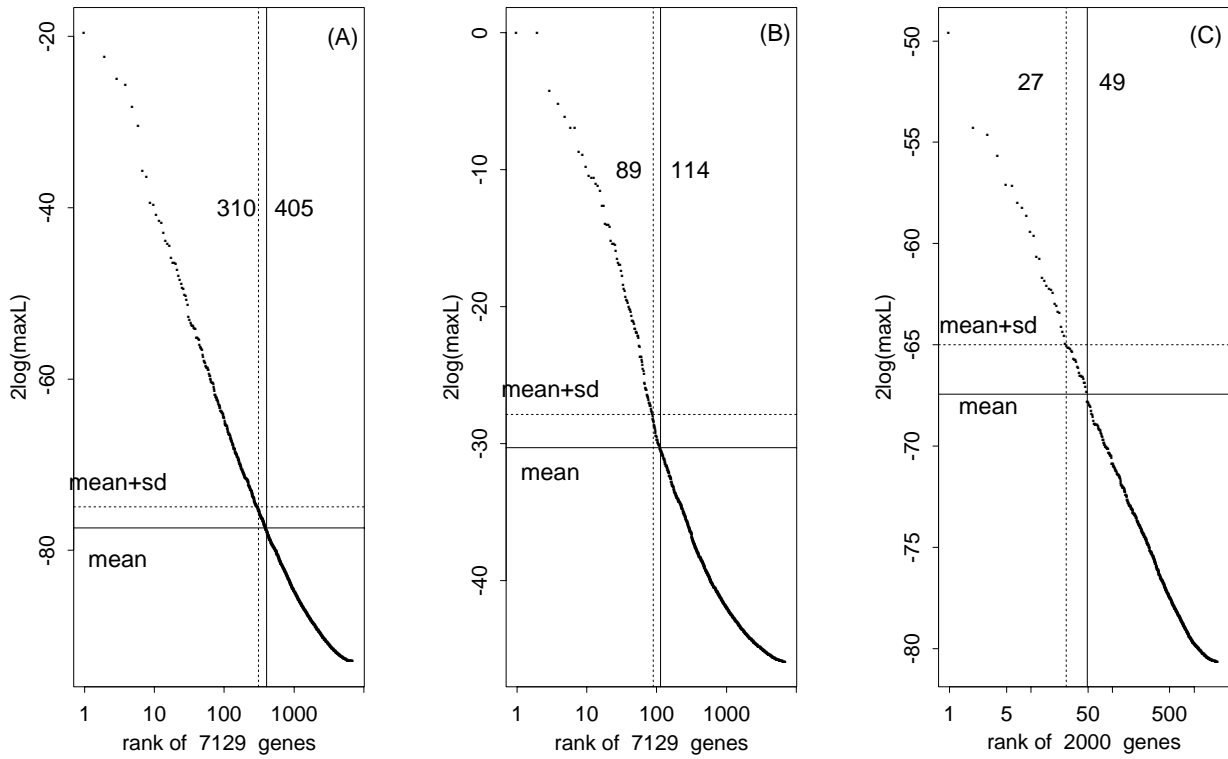


Figure 2: Maximum-likelihoods of single-gene LR models ranked from the gene with the highest maximum likelihood (rank 1) to the gene with the lowest maximum likelihood (rank p). The three classification tasks are: (A) two leukemia subtypes: ALL versus AML; (B) tissue source of ALL samples: T-cells versus B-cells; and (C) colon cancer versus normal tissues. In all three cases we indicate by solid lines the number of differentially expressed genes predicted by Eq. (9), and we indicate by dashed lines the number of differentially expressed genes predicted by Eq. (10).

eters would have to be recomputed for each gene and for each simulation run. To be specific, let us consider the example of analyzing a single dataset with $N = 50$ samples and $p = 25,000$ genes. Computing the LR parameters and maximum likelihoods for such a dataset takes of the order of a few CPU hours. The naive implementation of the resampling scheme would require of the order of 10^5 CPU hours for computing the LR parameters and maximum likelihoods for 10^4 surrogate data sets with randomly permuted labels, which corresponds to a running time of the order of a half year on a 100-CPU cluster. One way of circumventing those extensive computations is to compute the expected value of the top-ranking maximum likelihood of all single-gene LR models directly from the null model. We show that this shortcut can be accomplished by two simple steps: (i) use a theoretical expression for the top-ranking maximum likelihood of the shuffled data, and (ii) use simple and computationally inexpensive numerical simulations to compute the mean value and standard deviation of the resulting extreme value distribution. Further studies are needed to test the accuracy of the approximation solution provided in this paper.

For illustrative purposes we applied the proposed approach to estimate the number of relevant genes in two publicly available microarray datasets, and we found that—at least in case of the leukemia subtype data from the Whitehead Institute—the number of differentially expressed genes estimated by Eq. (9) or Eq. (10) is substantially smaller than the corresponding estimates from Ref. [16]. This finding suggests that the proposed gene selection criterion is more conservative and hence more focused on potentially useful genes than traditional approaches.

5. ACKNOWLEDGMENTS

We would like to thank Yaning Yang, Stephan Beirer, Hanspeter Herzel, Dirk Holste, and Armin Schmitt for valuable discussions. WL acknowledges partial support from NIH contract N01-AR12256, and IG is supported by a CSHL Association fellowship and NIH grant R01-HG01696.

6. REFERENCES

- [1] C Ambrose, GJ McLachlan (2002), “Selection bias in gene extraction on the basis of microarray gene-expression data”, *Proceedings of the National Academy of Sciences*, 99(10):6562-6566.
- [2] U Alon, N Barkai, DA Notterman, K Gish, S Ybarra, D Mack, AJ Levine (1999), “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays”, *Proceedings of National Academy of Sciences*, 96(12):6745-6750.
- [3] N Balakrishnan, AC Cohen (1991), *Order Statistics and Inference* (Academic Press).
- [4] P Baldi, AD Long (2001), “A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inference of gene changes”, *Bioinformatics*, 17(6):509-519.
- [5] Y Chen, E Dougherty, ML Bittner (1997), “Ratio-based decisions and the quantitative analysis of cDNA micro-array images”, *Journal of Biomedical Optics*, 2:364-374.
- [6] F Chiaromonte, J Martinelli (2002), “Dimension reduction strategies for analyzing global gene expression data with a response”, *Mathematical Biosciences*, 176(1):123-144.
- [7] JM Claverie (1999), “Computational methods for the identification of differential and coordinated gene expression”, *Human Molecular Genetics*, 8:1821-1832.
- [8] DR Cox, DV Hinkley (1974), *Theoretical Statistics* (Chapman & Hall).
- [9] HA David (1981), *Order Statistics*, 2nd ed (Wiley).
- [10] S Draghici (2002), “Statistical intelligence: effective analysis of high-density microarray data”, *Drug Discovery Today*, 7(11):S55-S63.
- [11] S Dudoit, J Fridlyand, TP Speed (2002), “Comparison of discrimination methods for the classification of tumors using gene expression data”, *Journal of the American Statistical Association*, 97(457):77-87.
- [12] PH Eilers, JM Boer, GJ van Ommen, HC van Houwelingen (2001), “Classification of microarray data with penalized logistic regression”, *Proceedings of SPIE*, 4266:187-198.
- [13] M Eisen, P Spellman, P Brown, D Botstein (1998), “Cluster analysis and display of genome-wide expression patterns”, *Proceedings of the National Academy of Sciences*, 95:14863-14868.
- [14] WJ Ewens, G Grant (2001), *Statistical Methods in Bioinformatics: An introduction* (Springer-Verlag).
- [15] JD Gibbons (1971), *Nonparametric Statistical Inference* (McGraw-Hill),
- [16] TR Golub, DK Sonim, P Tamayo, C Huard, M Gassenbeek, JP Mesirov, H Coller, ML Loh, JR Downing, MA Caligiuri, CD Bloomfield, ES Lander (1999), “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring”, *Science*, 286:531-536.
- [17] G Grant, E Manduchi, C Stoeckert Jr. (2002), “Using non-parametric methods in the context of multiple testing to determine differentially expressed genes”, in *Methods of Microarray Data Analysis: Papers from CAMDA '00*, eds. SM Lin, KF Johnson (Kluwer Academic), pp.37-55.
- [18] EJ Gumbel (1960), *Statistics of Extremes* (Columbia University Press).
- [19] T Hastie, R Tibshirani, J Friedman (2001), *The Elements of Statistical Learning: Data mining, inference, and prediction* (Springer).
- [20] T Ideker, V Thorsson, AF Siegel, LE Hood (2000), “Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data”, *Journal of Computational Biology*, 7(6):805-817.
- [21] H Li, F Hong (2001), “Cluster-Rasch models for microarray gene expression data”, *Genome Biology*, 2(8):research0031.
- [22] L Li, LG Pedersen, TA Darden, CR Weinberg (2002), “Computational analysis of leukemia microarray expression data using the GA/KNN method”, in *Methods of Microarray Data Analysis: Papers from*

- CAMDA'00, eds. SM Lin, KF Johnson (Kluwer Academic), pp.81-95.
- [23] W Li (2003), "Subtle differences in measuring the degree of differential expression in microarray data", preprint.
- [24] W Li, Y Yang (2002), "How many genes are needed for a discriminant microarray data analysis?", in *Methods of Microarray Data Analysis: Papers from CAMDA'00*, eds. SM Lin, KF Johnson (Kluwer Academic), pp. 137-150.
- [25] W Li, Y Yang (2002), "Zipf's law in importance of genes for cancer classification using microarray data", *Journal of Theoretical Biology*, 219:539-551.
- [26] SM Lin, KF Johnson, eds. (2002), *Methods of Microarray Data Analysis: Papers from CAMDA'00* (Kluwer Academic).
- [27] J Lu, S Hardy, WL Tao, S Muse, B Weir, S Spruill (2002), "Classical statistical approaches to molecular classification of cancer from gene expression profiling", in *Methods of Microarray Data Analysis: Papers from CAMDA'00*, eds. SM Lin, KF Johnson (Kluwer Academic), pp. 97-107.
- [28] F Model, P Adorjan, A Olek, C Piepenbrock (2001), "Feature selection for DNA methylation based cancer classification", *Bioinformatics*, 17(Suppl 1):S157-S164.
- [29] DM Mutch, A Berger, R Mansourian, A Rytz, MA Roberts (2002), "The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data", *BMC Bioinformatics*, 3:17.
- [30] DV Nguyen, DM Rocke (2002), "Tumor classification by partial squares using microarray gene expression data", *Bioinformatics*, 18:39-50.
- [31] W Pan (2002), "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments", *Bioinformatics*, 18(4):546-554.
- [32] PJ Park, M Pagano, M Bonetti (2001), "A nonparametric scoring algorithm for identifying informative genes from microarray data", *Biocomputing 2001: Proceedings of the Pacific Symposium* (World Scientific), pp.52-63.
- [33] M Sanchez-Carbayo, N Socci, JJ Lozano, W Li, T Belbin, M Preystowski, A Ortiz, G Childs, C Cordon-Cardo (2003), "Gene discovery in bladder cancer progression using cDNA microarrays", submitted to *American Journal of Pathology*.
- [34] SK Shevade, SS Keerth (2002), "A simple and efficient algorithm for gene selection using sparse logistic regression", Technical Report CD-02-22, Control Division, Department of Mechanical Engineering, National University of Singapore.
- [35] JN Siedow (2001), "Making sense of microarrays" (meeting report), *Genome Biology*, 2(2):reports4003.
- [36] GK Smyth, YH Yang, T Speed (2003), "Statistical issues in cDNA microarray data analysis", in *Functional Genomics: Methods and Protocols*, eds. MJ Brownstein and AB Khodursky (Methods in Molecular Biology Series, Vol 224) (Humana Press), in press.
- [37] JG Thomas, JM Olson, SJ Tapscott, LP Zhao (2001), "An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles", *Genome Research*, 11:1227-1236.
- [38] LJ Van't Veer et al (2002), "Gene expression profiling predicts clinical outcome of breast cancer", *Nature*, 415:530-536.
- [39] M Xiong, WJ Li, J Zhao, L Jin, E Boerwinkle (2001), "Feature (gene) selection in gene expression-based tumor classification", *Molecular Genetics and Metabolism*, 73(3):239-247.