10

HOW MANY GENES ARE NEEDED FOR A DISCRIMINANT MICROARRAY DATA ANALYSIS

Wentian Li and Yaning Yang
Laboratory of Statistical Genetics, Boc 192, The Rockefeller Univeristy, 1230 York Avenue, New York, NY 10021, USA.

Abstract: The analysis of the leukemia data from Whitehead/MIT group is a discriminant analysis (also called a supervised learning). Among thousands of genes whose expression levels are measured, not all are needed for discriminant analysis: a gene may either not contribute to the separation of two types of tissues/cancers, or it may be redundant because it is highly correlated with other genes. There are two theoretical frameworks in which variable selection (or gene selection in our case) can be addressed. The first is model selection, and the second is model averaging. We have carried out model selection using Akaike information criterion and Bayesian information criterion with logistic regression (discrimination, prediction, or classification) to determine the number of genes that provide the best model. These model selection criteria set upper limits of 22~25 and 12~13 genes for this data set with 38 samples, and the best model consists of only one (no.4847, zyxin) or two genes. We have also carried out model averaging over the best single-gene logistic predictors using three different weights: maximized likelihood, prediction rate on training set, and equal weight. We have observed that the performance of most of these weighted predictors on the testing set is gradually reduced as more genes are included, but a clear cutoff that separates good and bad prediction performance is not found.

Key words: model/variable selection, Akaike/Bayesian information criterion, maximum likelihood, logistic regression, discriminant analysis

INTRODUCTION

There are two types of microarray experiments. In the first type, samples are not categorized, but gene expression levels are followed with time. The goal is to find genes whose expression levels vary together. In the second type, samples are labelled as either normal or diseased tissues. The goal is to find genes whose expression levels can distinguish different labels. Using terminology from machine learning, the data analysis of the first type of experiment is unsupervised, whereas that of the second type is supervised (see, e.g., [Haghighi et al, 1999] for this particular application, and [Ripley, 1996] for general discussions). The prototype of the first type of analysis is cluster analysis, whereas that of the second type is discriminant analysis.

We focus on the data analysis of the second type, using the leukemia data from Whitehead/MIT's group [Golub et al, 1999]. The question we are addressing is how many expressed genes are needed for a discriminant analysis. In this data set, 7129 gene expressions are measured on 38 sample points. In the original analysis [Golub et al, 1999], 50 out of more than 7000 gene expressions are used for discriminant analysis. We ask whether 50 genes are still too large a number, considering that the sample size is only 38. This seemingly simple question may receive two answers from two different perspectives: model selection (see, e.g., [Burnham and Anderson, 1998]) and model averaging (see, e.g., [Geisser, 1993]). We will address the two separately below.

MODEL SELECTION

The goal of model selection is to pick the model among many possible models that achieves the best balance between data-fitting and model complexity. A perfect data-fitting performance by a complicated model with many parameters can well be an example of overfitting. On the other hand, a simplistic model with few parameters that fits the data poorly is an example of underfitting. There are at least two proposals for achieving the balance between data-fitting and model complexity: Akaike information criterion (AIC) [Akaike, 1974; Parzen et al, 1998; Burnham and Anderson, 1998] and Bayesian information criterion (BIC) [Schwarz, 1976; Raftery, 1995]. These two quantities are defined as

$$AIC=-2log(\hat{L})+2K \quad \text{and} \quad BIC=-2log(\hat{L})+log(N)K \tag{1}$$

where $\hat{L}$ is the maximized likelihood [Edwards, 1972], K the number of free parameters in the model, and N the sample size. High-order terms of $O(1/N)$ (for AIC) and $O(1) + O(1/\sqrt{N}) + O(1/N)$ (for BIC) are ignored here for simplicity. The model with the lowest AIC or BIC is considered to be the preferred model.

Table 1. Logistic regression results (sample size N=38, log(N)=3.637568) K: number of free parameters in the logistic regression (LR) (number of genes included plus 1); ΔAIC (ΔBIC) is the AIC (BIC) value relative to that of the best model) single-gene LR using g4847, zyxin); p_{train} (p_{test}) is the success (prediction) rate on the training (testing) set; "top 37" is the LR using 37 best genes by their single-variable LR performance. Notes (a) Since this model/predictor fits the data perfectly, L should be 1, and $-2\log(\hat{L})$ should be 0. In a real optimization procedure, the actual value may depend on the number of iterative steps. (b) This LR is selected by a stepwise variable selection (by either AIC or BIC) from the starting group of top 22 genes (top22 → AIC/BIC → 2). (c) Similar to (b), but the starting group of genes in the LR containing the top 10 genes (top10 → AIC/BIC → 2). (d) This null model 1 uses 11/38 ≈ 0.29 as the probability for AML for any sample data point. Since 0.29 < 0.5, any sample will be predicted as ALL type, which is correct 27 times in the training set, and 20 times in the testing set. The $-2\log(\hat{L})$ is equal to $-2\log[(27/38)^{27}(11/38)^{11}]$. (e) The $-2\log(\hat{L})$, p_{train}, p_{test} for the null model 0 (random guess model) is expected to be:$-2\log(0.5^N)= 2N\log(2)$, 0.5, and 0.5. A more complete gene description and the GenBank accession number of most genes listed here can be found in the Appendix.

Model	K	$-2\ln \hat{L}$	AIC	ΔAIC	BIC	ΔBIC	P_{train}	P_{test}
1 g4847(zyxin)	2	~0[a]	4	0	7.28	0	38/38	31/34
2 g1882(CST3 cystatin C)	2	6.97	10.97	6.97	14.25	6.97	36/38	32/34
3 g3320(c4 synthase)	2	10.91	14.91	10.91	18.19	10.92	35/38	27/34
4 g5039(leptin receptor)	2	11.36	15.36	11.36	18.63	11.36	36/38	22/34
5 g6218(ELA2 elastatse 2)	2	12.46	15.46	12.46	18.73	11.46	34/38	22/34
6 g2020(FAH…)	2	12.10	16.10	12.10	19.38	12.10	36/38	25/34
7 g1834 (CD33 antigen)	2	12.23	16.23	12.23	19.50	12.23	35/38	31/34
8 g760 (cystatin A)	2	13.10	17.10	13.10	20.38	13.10	35/38	32/34
9 g1745 (LYN v-yes-1…)	2	13.15	17.15	13.15	20.43	13.15	33/38	28/34
10 g5772 (c-myb)	2	14.72	18.72	14.72	22.00	14.72	35/38	27/34
100 g2833 (AF1q)	2	27.22	31.22	27.22	34.49	27.22	30/38	28/34
200 g3312 (p.k. ATR)	2	30.84	34.84	30.84	38.12	30.84	29/38	21/34
g1834+g2267 [b]	3	0.004	6.00	2.00	10.92	3.64	38/38	22/34
g5039+g5772 [c]	3	0.008	6.01	2.01	10.92	3.65	38/38	26/34
Top2(g4847+g1882)	3	0.029	6.03	2.03	10.94	3.67	38/38	32/34
Top5	6	0.011	12.01	8.01	21.84	14.56	38/38	24/34
Top10	11	0.002	22.00	18.00	40.02	32.74	38/38	31/34
Top22	23	0.001	46.00	42.00	83.67	76.39	38/38	27/34
Top37	38	0.001	76.00	72.00	138.2	130.9	38/38	21/34
null 1 (proportion)	1	45.73[d]	47.73	43.73	49.73	42.09	27/38[d]	20/34[d]
null 0 (random guess)	0	52.68[e]	52.68	48.68	52.68	45.40	19/38[e]	17/34[e]

Selecting genes relevant to a discriminant microarray data analysis may become an issue of model selection. But we have to be clear in what context this is the case. Model selection adjusts the number of parameters in the model, not the number of variables. Nevertheless, if we plan to combine variables (additively or any other functional form), each variable has one or perhaps more coefficients. Removing a variable removes the corresponding coefficients. In this context, variable selection is a special case of model selection. We illustrate this by the logistic regression/discrimination for our data set:

$$\Pr(AML) = \frac{1}{1 + \exp(-a_0 - \sum_{j \in \text{top genes}} a_j x_j)} \qquad (2)$$

where x_j is the (log, normalized) gene expression level of gene j among the top performing genes, and AML (acute myeloid leukemia) is one type of leukemia (ALL, the acute lymphoblastic leukemia, is another type). Using a linear combination of all 7129 genes in logistic discrimination requires 7130 parameters, whereas using one gene requires only 2 parameters. AIC/BIC defined in Eq.(1) will then compare multiple-gene models with single-gene models, and determine which scenario is better.

Our results are summarized in *Table 1,* where we list the type of model (which variables are additively combined in the logistic regression), number of parameters in the model, -2log of the maximum likelihood, AIC/BIC (absolute and relative), success rate in training set, and prediction rate in the testing set. The most striking result from this data set is that many genes are strong predictors for the leukemia class, consistent with the observation in [Golub et al, 1999]. This observation is confirmed by the following facts in *Table 1*:

•Logistic regressions using the top 2, 5, 10, 22,and 37 genes all fit the data almost perfectly (as measured by the -2 log($\hat{L}$) value: the closer is $\hat{L}$ to 1, the nearer to perfect fit) (note that the model is saturated when the number of variables used in the logistic regression is 37, since the number of parameters is then equal to the number of sample points).

•Stepwise variable selection leads to only two genes (note that stepwise variable selection fails to find the best model, a single-gene model, because it is a local minimization/maximization procedure).

•The No.10 best-performing gene is only slightly worse than the No.1 performing gene: 35 vs. 38 correct predictions on the training set (though the No.100 and No.200 best-performing genes predict only 30 and 29 correctly on the training set).

This situation of strong prediction or easy classification is in contrast with the epidemiology and pedigree data used in human complex (non-

Mendelian) diseases, where strong predictors are rather rare [Li et al, 2000; Li and Nyholt, 2001].

The last two rows in *Table 1* are predictors/classifiers that do not use any gene expression levels. These are called null models. The first null model (1) uses the proportion of AML in all training samples as Prob(AML) (11/38 in our data set) as the guessing probability for AML. The second null model (0) uses half-half probabilities. It is interesting to note that to beat both null models, the number of genes in logistic regression can not exceed certain upper limits: for null model 1, since we require AIC/BIC to be smaller (where p is the number of genes used in logistic regression, and $-2\log(\hat{L})$ is assumed to be zero for the best-case scenario):

$$0 + 2(p+1) \approx AIC < AIC(null\ 1) = 47.728$$
$$0+3.637568\ (p+1) \approx BIC < BIC(null\ 1) = 49.365, \qquad (3)$$

it sets $p < 22.86 \approx 22$, and $p < 12.57 \approx 12$, respectively. Similarly, to beat null model 0, we require:

$$0+2(p+1) \approx AIC < AIC(null\ 0) = 52.679$$
$$0+3.637568\ (p+1) \approx BIC < BIC(null\ 0) = 52.679 \qquad (4)$$

which are $p < 25.34 \approx 25$, and $p < 13.48 \approx 13$, respectively. Any logistic regression using more than these number of genes (22 ~ 25, 12 ~ 13), will have a larger value of AIC or BIC than the two null models. Since by the model selection criterion, only models with smaller values of AIC/BIC are selected, we should not include more than these number of genes in our logistic regression. Note that these specific upper limits are directly related to our particular sample size (i.e., 38). For larger sample sizes, the upper limit for the number of genes will be larger.

MODEL AVERAGING

In the model selection framework, we can not use a logistic regression with too many genes because it may not improve the data-fitting performance enough to compensate for the increase of model complexity. This conclusion is correct when one model (e.g. a logistic regression that uses one gene) is compared to other alternative models (e.g. a logistic regression that uses, say, 10 genes). Nevertheless, it is possible to average/combine many different models each involving one gene. The restriction on model complexity during the model selection process does not apply to model averaging. Model averaging has also been discussed under similar names such as "committee machines", "boosting weak predictors", "mixture of experts", "stacked generalization" (see, e.g., [Ripley, 1996]).

Without guidance from the model selection framework on the number of models (number of genes) to be included, we have tried several empirical approaches. We first examine whether there is a gap in data-fitting performance among top genes. Genes that do not fit the data should not be considered in model averaging. For this purpose, Fig.1 shows the -2 log($\hat{L}$) for the top 1000 genes on the training set, as a function of the (log) rank. A linear fitting of -2log($\hat{L}$) on log(rank) seems to fit the points well, and it is hard to see "better-than-average" genes except the first gene. This is not surprising since g4847 discriminates the 38 training sample points perfectly. Note that the linear trend in Fig.1 is equivalent to a power-law function in likelihood vs. rank plot ($\hat{L} \sim 1/r^{2.56}$). Such power-law is similar to the power-law rank-frequency plots observed in many social and natural data, also known as Zipf's law [Zipf, 1949; Li, 1997-2001]. In any case, there is no discernible gap in Fig.1 that separates relevant and irrelevant genes (more discussions on Zipf's plots similar to Fig.1, but obtained from other microarray data can be found in [Li, 2001]).

We then check how the prediction rate on the training set correlates with that on the testing set. Fig.2 shows the error rates on both training (x-axis) and testing sets (y-axis) for the top 500 performing genes. The left plot in Fig.2 shows the mean square error, and the right plot shows the prediction errors. The left plot contains both information on the success rate of prediction and that of confidence of prediction, whereas the right plot contains only information on the success rate of prediction. Points along the diagonal line in Fig.2 exhibit similar error rates in the training and testing set, and thus are reasonable predictors. On the other hand, points well above the diagonal line indicate overtraining. The most reasonable predictor based on both the training and testing set is g1882 (on the other hand, the best predictor based on the training set only is g4847).

Finally we examine the model averaging performance with various numbers of models included, each being a single-gene logistic regression:

$$\Pr(AML) = \sum_{j \in \text{top genes}} w_j \left(\frac{1}{1+\exp(-a_j - b_j x_j)} \right) \quad (5)$$

Although the terms to be summed in Eq.(5) can be any model, here we only examine the simplest case where the model to be averaged is the single-gene logistic regression model.

Figure 1. Data-fitting performance as measured by -2log of the maximized likelihood (the smaller, the better fit) for the top 1000 genes. The x-axis is the rank of the gene by the maximum likelihood of its single-gene logistic regression. A linear regression line of -2log( ) vs log (rank) is also shown, with the slope equal to 5.12.

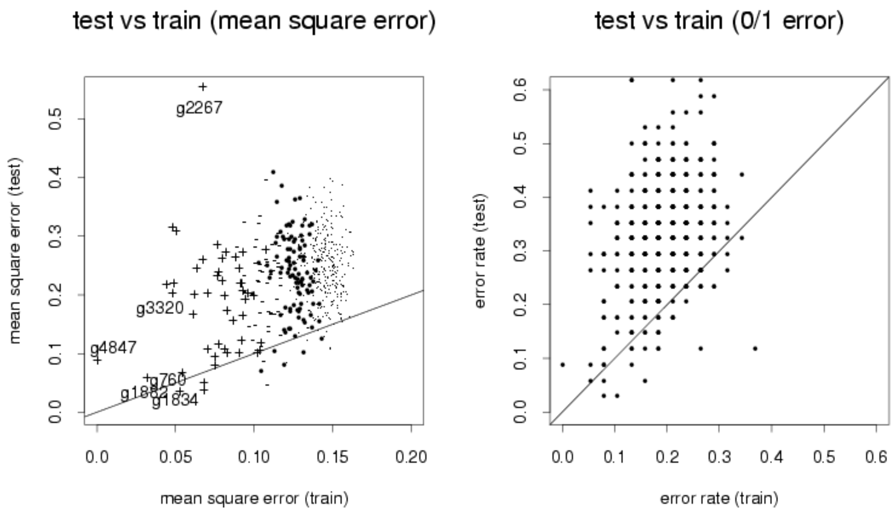

Figure 2. Error rates on the testing set (y-axis) are compared to those on the training set (x-axis) for the top 500 genes (based on the performance on the training set). Each point represents one single-gene logistic regression predictor. The left plot uses the mean squared error ("+" for the first top 100 genes, "-" for the rank 51-100 genes, "*" for the rank 101-200 genes, and "." For the rank 201-500 genes, and the right plot uses the prediction error.

Figure 3. Model averaging performance with three different weighting schemes: (1) weight being proportional to the prediction rate on the training set; (2) equal weight; and (3) weight being proportional to the maximum likelihood obtained on the training set. The mean square error is plotted against the number of models being averaged. Each model is a single-gene logistic regression. The mean square error on both the training set (bottom) and the testing set (top) are shown.

We have chosen three weighting schemes: the first is proportional to the prediction rate on the training set ($w_j \propto p_{j,train}$), the second is the equal weight ($w_j \propto 1$), and the last one is proportional to the maximum likelihood as obtained from the training set ($w_j \propto \hat{L}_j$). Fig.3-4 shows the behaviour of all three weighting schemes on both the training and the testing sets, using either the mean square error or the prediction error, up to 200 genes.

The maximum likelihood weight is equivalent to the Akaike weight ($w_j \propto \exp(-AIC_j/2)$) [Parzen et al, 1998] and Bayesian weight ($w_j \propto \exp(-BIC_j/2)$) [Raftery, 1995], since all models being averaged in Eq.(5) have the same number of parameters and same sample size (assuming high-order terms are ignored). Using Bayesian weight in model averaging is essentially to derive the posterior predictive distribution [Gelman, et al, 1995]: "posterior" because data in a training set is used, and "predictive" because unknown new data in the testing set is to be predicted.

For our data set, this weighting scheme is nevertheless uninteresting: the mean square error only decreases slightly with the number of models used, while the error in prediction rate is unchanged. The reason for this is very simple: the best model (the single-gene logistic regression using the g4847) discriminates the training set perfectly. Its likelihood is much higher than any other models. As a result, the weight of other models is negligible, and the model averaging essentially remains as one model. The equal weight can be potentially incorrect since models that do not fit the data should not contribute as equally as models that fit the data better. In our data set, however, the equal weight scheme is actually similar to the weighting scheme that uses the prediction rate, because many genes (up to 200) exhibit similar prediction rates on the training set.

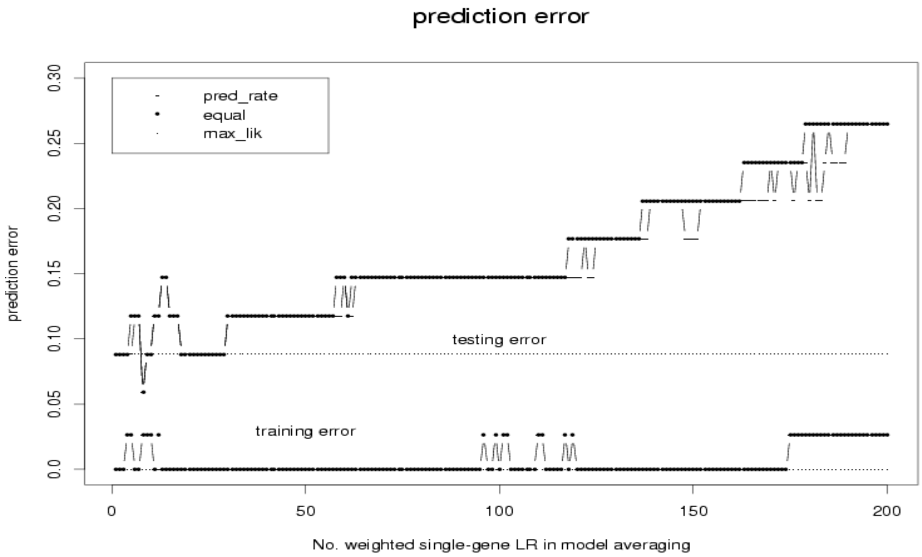

Figure 4. Similar to Fig.3, but the prediction error rate is plotted against the number of models being averaged (each model is a single-gene logistic regression).

It is difficult from Fig.3-4 to determine a cutoff on the number of models (number of genes) to be included. Fig.3-4, however, clearly shows that a lower number of models (genes) is typically better (except the maximum likelihood weight, which is insensitive to the number of genes included). Fig.3, in particular, indicates that it is possible to perform better (in term of mean square error) on the testing set than using just one model if the number of models is less than around 25. Of course, this result is obtained from the specific testing set we have at hand. A more conclusive result may require a combining of the training and testing set, or requires more sample points than are currently available.

Due to the weight in a model averaging, the apparent number of terms (models, genes) included does not reflect on the true number of genes involved. For this reason, we may introduce a quantity called “effective number of genes (terms, models)". The leading term in a model averaging contributes a number of 1 to this quantity, but the contribution from the term *j* is equal to w_j/w_1. For example, with the maximum likelihood weight, the relative weights of the first ten terms are 1, 0.031, 0.0043, 0.0034, 0.0032, 0.0024, 0.0022, 0.0014, 0.0014, and 0.0006. The effective number of terms when the top ten genes are included is 1.05, a far less number than 10.

The discrimination used in [Golub et al, 1999] is a model averaging instead of a model selection. The number of genes used is 50. To quote from [Golub, et al, 1999], "the number was well within the total number of genes strongly correlated with the class distinction, seemed likely to be large enough to be robust against noise, and was small enough to be readily applied in a clinical setting." Our Fig.3-4 shows that although the prediction rate on the training set stays close to 100% even as the number of models (genes) is increased, the prediction rate on the testing set decreases with more genes. The only exception is the maximum likelihood weight, where the prediction rate is almost unchanged due to the dominance of the best gene. It seems that we should not increase the number of models (genes) in model averaging arbitrarily.

SUMMARY

It is important to emphasize what we have done/concluded and we have not. We have addressed the number of genes to be included in a predictor *when these genes are combined in a logistic regression model or any other model*. The answer to this question is found from the perspective of model selection. If gene expressions from different genes are not combined, but examined individually, how many genes should be claimed to be relevant to the discrimination of the two labels is a different issue, and usually involves a subjective decision. For example, in a t-test, the hypothesis that the expression of a particular gene is different in two types of samples is tested. Whether this hypothesis is accepted, or more accurately, whether the null hypothesis is rejected, depends on the significance level. Different choices of the significance level lead to various number of genes claimed to be "relevant".

In a model averaging framework, the question on the number of genes to be included in the discrimination is somehow ill phrased. Because of the concept of the weight, inclusion of one gene is not equal to an inclusion of another gene with less effect. There is no limit on the number of genes that could be included in a model averaging because adding expression values from unimportant gene does not greatly affect the nature of the predictor. In this context, the more natural question is '*what is the value of the effective* *number of genes'*. It is very much possible that even if the expression of all 7129 genes is included in the predictor by a model averaging, the effective number of genes can be still as small as less than 2.

Another topic we did not address is the correlation between genes and the related concept of gene clusters. In this paper, the variable being selected is the individual gene, the model being averaged is logistic regression using an individual gene, etc. Nevertheless, all these analyses can be easily extended to situations where not only genes, but also clusters of genes, are targets of selection or averaging. In [Hastie et al., 2001], for example, the variable to be selected includes both the original genes and clusters at various levels obtained from a hierarchical cluster analysis.

We also did not address the possibility that the incorrect predictions are due to the sample outlier [Li and Wong, 2001] or sample heterogeneity. *Table 1* shows that the best fitting model on the training set is the single-gene logistic regression using g4847 (zyxin). With this predictor, there are three incorrect predictions on the testing set: samples 66(AML), 67(T-cell ALL) and 68(B-cell ALL). The best predictor on the testing set is the single-gene logistic regression using g1882 (CST3 cystatin C), or that using g760 (cystatin A). With the g1882 predictor, there are only two incorrect predictions, on samples 61(AML) and 67(T-cell ALL); with g760, samples 61 and 66 (both AML) are incorrectly predicted. We notice that sample 67 is the only T-cell ALL in the testing set, so it is not impossible that it behaves different from the B-cell ALL's (see [Grant et al., 2001]).

In conclusion, due to the small sample size and the presence of strong predictors, we believe that the number of genes used in a discriminant analysis in this data set can be much smaller than 50 as stated in [Golub et al., 1999]. Although we can not give a definite answer as to the exact number of genes to be used, one proposal is to use only one or two genes, and other exploratory data analyses indicate an upper limit of 10-20 genes for this data set. A more extensive study of the same issue using other microarray data sets can be found elsewhere (Li et al., in preparation).

ACKNOWLEDGEMENTS

W.Li's work was supported by NIH grant K01HG00024 and Y. Yang's work was supported by the grant MH44292.

APPENDIX

Table 2. Gene names and GenBankAccession numbers for the genes in Table 1*:*

Gene #	description	GenBank A#
4847	Zyxin	X95735_at
1882	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	M27891_at
3320	Leukotriene C4 synthase (LTC4S) gene	U50136_rna1_at
5039	LEPR Leptin receptor	Y12670_at
6218	ELA2 Elastatse 2, neutrophil	M27783_s_at
2020	FAH Fumarylacetoacetate	M55150_at
1834	CD33 antigen (differentiation antigen)	M23197_at
760	Cystatin A	D88422_at
1745	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog	M16038_at
5772	C-myb gene extracted from Human, complete primary CDS, and five complete alternatively spliced CDS	U22376_cds2_s_at
2267	CDC25A Cell division cycle 25A	81933_at

REFERENCES

Akaike, H (1974), “A new look at the statistical model identification", IEEE Transactions on Automatic Control, 19:716-723.

Burnham, KP, Anderson, DR (1998), *Model Selection and Inference* (Springer).

Edwards, AWF (1972), *Likelihood* (John Hopkins Univ Press).

Geisser, S (1993), *Predictive Inference*: *An Introduction* (Chapman & Hall).

Gelman, A, Carlin, JB, Stern, HS, Rubin, DB (1995), *Bayesian Data Analysis* (Chapman & Hall).

Golub, TR, Slonim, DK, Tamayo, P, Huard, C, Gaasenbeek, M, Mesirov, JP, Coller, H, Loh, ML, Downing, JR, Caligiuri, MA, Bloomfield, CD, Lander, ES (1999),"Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science, 286:531-537.

Grant GR, Manduchi E, Stoeckert Jr. CJ (2001), “Using non-parametric methods in the context of multiple testing to identify differentially expressed genes”, this volume.

Haghighi, F, Banerjee, P, Li, W (1999), "Application of artificial neural networks in whole-genome analysis of complex diseases" (meeting abstract), Cold Spring Harbor Meeting on Genome Sequencing & Biology, page 75.

Hastie, T, Tibshirani, R, Botstein, D, Brown, P (2001), "Supervised harvesting of expression trees", Genome Biology, 2(1):research0003.1-0003.12.

Li C, Wong WH (2001), "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection", Proceedings of National Academy of Sciences, 98:31-36.

Li, W (1997-2001), A page on Zipf's law (URL: linkage.rockefeller.edu/wli/zipf/).

Li, W (2001), "Zipf's law in importance of genes for cancer classification using microarray data", submitted.

Li, W, Nyholt, D (2001), "Marker selection by Akaike information criterion and Bayesian information criterion", Genetic Epidemiology, in press.

Li, W, Sherriff, A, Liu, X (2000), "Assessing risk factors of complex diseases by Akaike information criterion and Bayesian information criterion" (meeting abstract), American Journal of Human Genetics, 67 (supp 2), page 222.

Parzen, E, Tanabe, K, Kitagawa, G (1998), eds. *Selected Papers of Hirotugu Akaike* (Springer).

Raftery, AE (1995), "Bayesian model selection in social research", in Sociological Methodology, ed. PV Marsden (Blackwells),pages 185-195.

Ripley, BD (1996), Pattern Recognition and Neural Networks (Cambridge University Press).

Schwarz, G (1976),"Estimating the dimension of a model", Annals of Statistics, 6:461-464.

Zipf, GF (1949), *Human Behavior and the Principle of Least Effect* (Addison-Wesley).