BIOINFORMATICS

Vol. 19 no. 1 2003
Pages 90–97

Gene selection: a Bayesian variable selection approach

Kyeong Eun Lee[1], Naijun Sha[4], Edward R. Dougherty[2, 3], Marina Vannucci[1] and Bani K. Mallick[1,]*

[1]Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA, [2]Department of Electrical Engineering, Texas A&M University, College Station, TX 77840, USA, [3]Department of Pathology, University of Texas, M. D. Anderson Cancer Center, USA and [4]Mathematical Sciences Department, University of Texas at El Paso, USA

Received on January 24, 2002; revised on April 29, 2002; accepted on June 17, 2002

ABSTRACT

Selection of significant genes via expression patterns is an important problem in microarray experiments. Owing to small sample size and the large number of variables (genes), the selection process can be unstable. This paper proposes a hierarchical Bayesian model for gene (variable) selection. We employ latent variables to specialize the model to a regression setting and uses a Bayesian mixture prior to perform the variable selection. We control the size of the model by assigning a prior distribution over the dimension (number of significant genes) of the model. The posterior distributions of the parameters are not in explicit form and we need to use a combination of truncated sampling and Markov Chain Monte Carlo (MCMC) based computation techniques to simulate the parameters from the posteriors. The Bayesian model is flexible enough to identify significant genes as well as to perform future predictions. The method is applied to cancer classification via cDNA microarrays where the genes BRCA1 and BRCA2 are associated with a hereditary disposition to breast cancer, and the method is used to identify a set of significant genes. The method is also applied successfully to the leukemia data.

Supplementary information: http://stat.tamu.edu/people/faculty/bmallick.html

Contact: bmallick@stat.tanu.edu

INTRODUCTION

cDNA microarrays provide expression measurements for thousands of genes at once (Duggan *et al.*, 1999; Schena *et al.*, 1995). A key goal of gene selection via different expression patterns is to identify the responsible genes for certain events (say, certain diseases or certain types of tumors). The inherent power of expression data to separate sample types was first clearly demonstrated by clustering samples on the basis of gene expression patterns. Microarray problems can be classified as unsupervised, when only the expression data are available, and supervised, when a response measurement is available for each sample. In unsupervised problems the goal is mainly to identify distinct sets of genes with similar expressions, suggesting that they may be biologically related. Supervised and unsupervised problems also focus on finding sets of genes that, for example, relate to different kinds of diseases, so that future tissue samples can be correctly classified. Traditional statistical methods for clustering and classification have been extensively applied to microarray data, see Eisen *et al.* (1998); Alizadeh *et al.* (2000) for clustering and Golub *et al.* (1999) and Hedenfalk *et al.* (2001) for classification. In the supervised case a Bayesian approach to dimension reduction with a probit model has been applied by West *et al.* (2000) where unlike selecting actual genes, the singular-value decomposition is applied to the design matrix to reduce the dimension of the problem. In this paper we mainly want to identify (select) the important genes which are significantly more influential than the others for the classification process and therefore focus on variable selection.

In many published studies, the number of selected genes is large, for instance, 495 genes (Khan *et al.*, 1998) and 2000 genes (Alon *et al.*, 1999). Even in studies that obtained smaller numbers of genes, the numbers are excessive for the small number of sample points (microarrays), for instance 50 genes (Golub *et al.*, 1999) or 96 genes (Khan *et al.*, 2001). A large number of genes in conjunction with a small sample size is not advisable because it can create an unreliable selection process (Dougherty, 2001). Dudoit *et al.* (2000) have proposed a method for the identification of singly differentially expressed genes by considering a univariate

*To whom correspondence should be addressed.

testing problem for each gene and then correcting for multiple testing using adjusted p-values. Tusher *et al.* (2001) have proposed *Significance Analysis of Microarray* (*SAM*), which assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements. Given the genes with scores greater than an adjustable threshold, SAM uses permutations of the repeated measurements to estimate the percentage of genes identified by chance. Hastie *et al.* (2000) have suggested gene shaving, a new class of clustering methods that tries to identify subsets of genes with coherent expression patterns with large variation across conditions. Kim *et al.* (2002) have proposed analytically designing low-dimension linear classifiers from a probability distribution resulting from spreading the mass of the sample points to make classification more difficult, while maintaining sample geometry. The algorithm is parameterized by the variance of the spreading distribution. By increasing the spread, the algorithm finds gene sets whose classification accuracy remains strong relative to greater spreading of the sample.

In this paper we suggest a model-based approach to the variable selection problem. Rather than fixing the dimension (the number of selected genes), we assign a prior distribution over it. The approach creates additional flexibility by allowing the imposition of constraint, such as not allowing the dimension to be too big (say not more than 8) by using this prior. The prior can work as a penalty to create constraint. We use a Markov Chain Monte Carlo (MCMC; Gilks *et al.*, 1996) based stochastic search algorithm that discovers important genes. Given that the model space is very large, as with p genes (say $p = 3000$) we have 2^p models, exhaustive computation over this model space is not possible. MCMC based stochastic search algorithms, less greedy and more efficient than most other existing algorithms are therefore successfully implemented to identify significant genes. We consider the model for binary events only, assuming there are only two categories of events; extension to multi-category data is done recently by Sha (2002).

We will consider a data set from Hedenfalk *et al.* (2001) comparing the expression profiles of hereditary breast cancers where we want to identify genes that can discriminate between BRCA1 and BRCA2 breast cancers. The idea is to identify a small number of genes (by penalizing the dimension) having the greatest discriminating power, thereby allowing researchers to quickly focus on the most promising candidates for diagnostics and therapeutics. This same data set was used in Kim *et al.* (2002), and we will point out some similarities between the resulting gene lists. We also use our method on a larger leukemia data set from Golub *et al.* (1999) to show its predictive power over an independent test set.

MODEL FOR GENE SELECTION

We consider binary responses as $Y_i = 1$ indicates that the tumor sample i is BRCA1 and $Y_i = 0$ indicates that it is BRCA2, for $i = 1, \ldots n$. For each sample we measure expression levels for a set of genes, so X_{ij} is the measurement of the expression level of the jth gene for the ith sample where $j = 1, \ldots p$.

$$\begin{bmatrix} & \text{Gene 1} & \text{Gene 2} & \cdots & \text{Gene } p \\ Y_1 & X_{11} & X_{12} & \cdots & X_{1p} \\ Y_2 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_n & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

We assume that Y_i has an independent binary distribution so that $Y_i = 1$ with probability p_i, independently of the other Y_j, $j \neq i$. Then we relate the gene expression level with the response using a probit regression model which yields

$$Pr(Y_i = 1|\beta) = \Phi(X_i^{'}\beta)$$

where X_i is the ith row of the matrix X (vector of gene expression levels of the ith sample), β is the vector of regression parameters (β_j is the regression parameter corresponding to the jth gene) and Φ is the normal cumulative distribution function.

Albert and Chib (1993) introduced n independent latent variables $Z_1, \cdots, Z_n$ with $Z_i \sim N(X_i'\beta, 1)$ and where

$$Y_i = \begin{cases} 1 & Z_i > 0 \\ 0 & Z_i < 0. \end{cases}$$

The latent variable has a linear model form as $Z_i = X_i'\beta + \epsilon_i$ where $\epsilon_i \sim N(0, 1)$.

In order to perform the variable selection we define γ to be a $p \times 1$ vector of indicator variables with ith element γ_i such that $\gamma_i = 0$ if $\beta_i = 0$ (the gene is not selected) and $\gamma_i = 1$ if $\beta_i \neq 0$ (the gene is selected). Given γ, let β_γ consist of all nonzero elements of β and let X_γ be the columns of X corresponding to those elements of γ that are equal to 1. To complete the hierarchical model we need make the following prior assumptions:

(1) Given γ, the prior for β_γ is $\beta_\gamma \sim N(0, c(X_\gamma' X_\gamma)^{-1})$ where c is a positive scale factor specified by the user. Smith and Kohn (1996) suggested to choose c between 10 to 100 for linear model problems after extensive testing. We will fix it to a large value as $c = 100$, in that for this value the prior of β_γ, given γ, contains very little information about β_γ compared to the likelihood.

(2) The γ_i are assumed to be *a priori* independent with $Pr(\gamma_i = 1) = \pi_i, 0 \leq \pi_i \leq 1$, for $i = 1, \ldots, p$. We will choose the values π_i to be small, therefore restricting the number of genes in the model. Prior knowledge that some

genes are more important than others can be implemented easily by assigning larger or smaller values of π in a scale of importance from 0 to 1. This prior allows also to model interactions among the genes. Small values of π will indirectly prevent the number of selected genes in the model from being too large.

COMPUTATION

The posterior distribution is not available in explicit form and we therefore use the MCMC method (Gilks *et al.*, 1996) specifically Gibbs sampling (Gelfand and Smith, 1990) to generate the parameters from the posterior distributions.

Our unknowns are (Z, β, γ) and in order to implement the Gibbs sampling we need to simulate from the complete conditional distributions. Rather than drawing from the complete conditionals we will modify the algorithm and draw γ from the marginal distribution (integrating β out) which will speed up the computations. It can be shown that this modified Gibbs sampler still leaves the target posterior distribution invariant. So our computation scheme will be as follows:

(i) Draw from $\gamma|Z$, the marginalized conditional distribution obtained after integrating β out (this is conditionally independent of Y). Now

$$p(Z|\gamma) \propto \int_\beta p(Z|\beta_\gamma)p(\beta_\gamma|\gamma)d\beta_\gamma \propto \exp[-1/2(Z'Z - \frac{c}{1+c}Z'X_\gamma(X'_\gamma X_\gamma)^{-1}X'_\gamma Z)]$$

where β_γ is integrated out as normal integral. The proof is a simple application of Bayesian linear model theory (Lindley and Smith, 1972; Denison *et al.*, 2002) and provided as supplementary information in our web site http://stat.tamu.edu/people/faculty/bmallick.html.

Then the conditional distribution is

$$p(\gamma|Z) \propto p(Z|\gamma)p(\gamma) \propto \exp[-1/2(Z'Z - \frac{c}{1+c} \times Z'X_\gamma(X'_\gamma X_\gamma)^{-1}X'_\gamma Z)]\Pi_{i=1}^{p}\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}$$

Rather than drawing γ as a vector, it is better to draw component wise from $p(\gamma_i|Z, \gamma_{j\neq i})$ which is

$$p(\gamma_i|Z, \gamma_{j\neq i}) \propto p(Z|\gamma)p(\gamma_i) \propto \exp[-1/2(Z'Z - \frac{c}{1+c}Z'X_\gamma(X'_\gamma X_\gamma)^{-1}X'_\gamma Z)]\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}.$$

(ii) Draw β from $\beta_\gamma|\gamma, Z$, (which is conditional independent of Y). The conditional distribution is $p(\beta|\gamma, Z) \sim N(V_\gamma X'_\gamma Z_\gamma, V_\gamma)$ where $V_\gamma = \frac{c}{1+c}(X'_\gamma X_\gamma)^{-1}$ and index γ means all the elements only corresponding to $\gamma = 1$.

(iii) The full conditional distribution of Z_i is as follows:

$Z_i\|\beta, Y_i = 1 \propto N(X'_i\beta, 1)$	truncated at the left by 0
$Z_i\|\beta, Y_i = 0 \propto N(X'_i\beta, 1)$	truncated at the right by 0

The distribution of Z is a truncated normal, so its random number generating algorithm follows Robert's (1995) optimal exponential accept–reject algorithm.

After suitable burn-in period (usually 10 000) we obtain the MCMC samples at the tth iteration as $\{\beta^{(t)}, Z^{(t)}, \gamma^{(t)},\ t = 1, \ldots, m\}$. We can use this samples from the posterior distributions for posterior inference and prediction.

The Algorithm

Start with initial values $[\gamma^{(0)}, Z^{(0)}, \beta^{(0)}]$

At the tth iteration

(i) Draw $\gamma^{(t)}$ from $p(\gamma|Z^{(t-1)})$.

(ii) Draw $Z^{(t)}$ from $p(Z|\beta^{(t-1)}, \gamma^{(t)})$

(iii) Draw $\beta^{(t)}$ from $p(\beta|Z^{(t)}, \gamma^{(t)})$.

Increase t until the required number of iterations.

Stop

For decision making we can calculate the relative number of times that each gene appeared in the sample (number of times the corresponding γ is 1). This will give us an estimate of the posterior probability of inclusion of the single gene as a measure of the relative importance of the gene for classification purpose.

Also we can obtain the predictive classification of a new observation Y_{new}, conditioning on the expression levels as

$$P(Y_{new} = 1|X) = \int_\gamma \int_Z \int_\beta p(Y_{new} = 1|X, Z, \beta, \gamma) \times p(z, \beta, \gamma|Y)dZd\beta d\gamma. \quad (1)$$

The Monte–Carlo estimate of this probability will be

$$\hat{P}(Y_{new} = 1|X) = \frac{1}{m}\sum_{t=1}^{m} p(Y_{new} = 1|X, Z^{(t)}, \beta^{(t)}, \gamma^{(t)})$$

and it can be easily evaluated using normal CDF.

APPLICATION OF GENE SELECTION TO HEREDITARY BREAST CANCER DATA

We apply the proposed strategy for discovering significant genes to a published data set (Hedenfalk *et al.*, 2001) on breast tumors from patients carrying mutations in the predisposing genes, BRCA1 or BRCA2 or from patients not expected to carry a hereditary predisposing mutation. Pathological and genetic differences appear to imply different but overlapping functions for BRCA1 and BRCA2. In Hedenfalk *et al.* (2001), cDNA microarrays were used in conjunction with classification algorithms

Table 1. Breast cancer data: strongly significant genes for the classification of BRCA1 versus BRCA2 or sporadic

Frequency* (%)	Image clone ID	Gene description
8.6	897781	keratin 8
8.4	823940	TOB1
7.8	26184	'phosphofructokinase, platelet'
7.5	840702	SELENOPHOSPHATE SYNTHETASE; Human selenium donor protein
7.1	376516	cell division cycle 4-like
6.9	47542	small nuclear ribonucleoprotein D1 polypeptide (16 kD)
6.6	366647	butyrate response factor 1 (EGF-response factor 1)
6.6	293104	phytanoyl-CoA hydroxylase (Refsum disease)
6.2	28012	O-linked *N*-acetylglucosamine (GlcNAc) transferase
6.1	212198	'tumor protein p53-binding protein, 2'
5.9	247818	ESTs
5.5	26082	very low density lipoprotein receptor
5.4	667598	PC4 and SFRS1 interacting protein 1
5.2	30093	RAN binding protein 1
5.1	73531	nitrogen fixation cluster-like
5	950682	'phosphofructokinase, platelet'
5	47681	'splicing factor, arginine/serine-rich (transformer 2 Drosophila homolog)'
4.9	46019	minichromosome maintenance deficient (*S. cerevisiae*) 7
4.9	307843	ESTs
4.8	548957	'general transcription factor II, i, pseudogene 1'
4.7	788721	KIAA0090 protein
4.7	843076	signal transducing adaptor molecule (SH3 domain and ITAM motif)
4.7	204897	'phospholipase C, gamma 2 (phosphatidylinositol-specific)'
4.7	812227	'solute carrier family 9 (sodium/hydrogen exchanger), isoform 1'
4.6	566887	heterochromatin-like protein 1
4.6	563598	'gamma-aminobutyric acid (GABA) A receptor, pi'
4.5	324210	sigma receptor (SR31747 binding protein 1)

* Percentage of times the genes appeared in the posterior samples.

to show the feasibility of using differences in global gene expression profiles to separate BRCA1 and BRCA2 mutation-positive breast cancers. They examined 22 breast tumor samples from 21 breast cancer patients, and all patients except one were women. Fifteen women had hereditary breast cancer, 7 tumors with BRCA1 and 8 tumors with BRCA2. 3226 genes were used for each breast tumor sample. We use our method to classify BRCA1 versus the others (BRCA2 and sporadic).

We used a two-sample t-statistics to identify the starting values, say the 5 most significant genes. We then ran the MCMC sampler, in particular, the Gibbs sampling approach fixing $\pi_i = 0.005$ for all $i = 1, 2, \ldots, p$. The chain moved quite frequently and we used 50 000 iterations after a 10 000 burn-in period. Table 1 lists the most significant genes as those with the largest frequencies.

We note that the three leading genes in Table 1 appear among the six strongest genes in an analogous list in Kim *et al.* (2002). This has occurred even though the rating in the latter paper is based upon the ability of a gene to contribute to a linear classifier, which is quite different than the criterion here. The leading gene in Table 1 is keratin 8 (KRT8), which also leads the list of strong genes in Kim *et al.* (2002). It is a member of the cytokeratin family of genes. Cytokeratins are frequently used to identify breast cancer metastases by immunohistochemistry, and cytokeratin 8 abundance has been shown to correlate well with node-positive disease (Brotherick *et al.*, 1998). The gene TOB1 is second in Table 1, and appeared fifth in Kim *et al.* (2002). It interacts with the oncogene receptor ERBB2, and is found to be more highly expressed in BRCA2 and sporadic cancers, which are likewise more likely to harbor ERBB2 gene amplifications. TOB1 has an anti-proliferative activity that is apparently antagonized by ERBB2 (Matsuda *et al.*, 1996). We note that the gene for the receptor was not on the arrays, so that the gene-selection algorithm was blinded to its input. Lastly, the third gene in Table 1 appears as the sixth gene in the list of Kim *et al.* (2002).

We check the model adequacy in two ways. (i) Cross validation approach: we excluded a single data point (leave-one-out cross validation) and predicted the probability of $Y = 1$ for that point using Equation (1). We compared this with the observed response and most of the cases obtained almost perfect fitting: 0 classification errors (number of misclassified observations). (ii) Deviance: Deviance calculation is a criterion-based method measuring the goodness of fit (McCullagh and Nelder, 1989). Lower deviance means better fit. We calculated the probabilities and the deviance measures for the different models in Table 2, showing their adequacy:

Model 1 : Using all strong significant genes.
Model 2 : Using genes with frequencies more than 5%.
Model 3 : Using genes with frequencies more than 6%.
Model 4 : Using genes with frequencies more than 7%.

We compared our cross validation results with other popular classification algorithms including feed forward neural networks, k-nearest neighbors, support vector machines (SVM). Results are in Table 3. All other methods have used 51 genes (which we think is too many with respect to a sample size of 22) which may produce instability in the classification process. Our procedure has used a much less number of genes though the results are competitive to any other method.

Table 2. Crossvalidated classification probabilities and deviances of the 4 models for the breast cancer data set

Y	Model 1 $Pr(Y=1\|X)$	Model 2 $Pr(Y=1\|X)$	Model 3 $Pr(Y=1\|X)$	Model 4 $Pr(Y=1\|X)$
1	1	1	0.9993	0.9998
1	1	1	1	0.9969
1	1	1	0.9999	1
1	1	1	0.9999	0.8605
1	1	1	0.9999	0.7766
1	1	1	0.9998	1
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0.0002
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0.0002
0	0	0	0.0018	0.0867
0	0	0	0.0005	0.007
0	0	0	0	0
0	0	0	0	0.2864
1	1	1	1	1
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
Deviance	$1.2683e-12$	$3.1464e-7$	0.0071	1.6843
Number of misclassifications	0	0	0	1

Table 3. Cross-validation errors of different models for the breast cancer data set

	Model	Cross-validation error*
1	Feed-forward neural networks (3 hidden neurons, 1 hidden layer)	1.5 (Average error)
2	Gaussian kernel	1
3	Epanechnikov kernel	1
4	Moving window kernel	2
5	Probabilistic neural network ($r = 0.01$)	3
6	kNN ($k = 1$)	4
7	SVM linear	4
8	Perceptron	5
9	SVM Nonlinear	6

$*$ Number of misclassified samples.
Feature Selection: 51 Features used in the paper 'Gene-expression profiles in hereditary breast cancer' (Hedenfalk *et al.*, 2001).

Sensitivity analysis

We have checked the sensitivity (stability) of our analysis to the measurement of expression levels (as they are always subject to measurement errors) by adding Gaussian noise to the expression values. We re-analyzed the data contaminated by different levels of Gaussian noise to obtain the newly selected genes and have reproduced the results in Table 4. The table shows that the analysis is quite stable, as it is selecting almost similar genes under different noise levels over the expression values. Among the seven leading genes in Table 1, the following appear across all noise conditions in Table 4: Keratin 8, 'phosphofructokinase platelet,' Selenophosphate synthetase, and butyrate response factor 1. TOB1 is only omitted at the highest noise level.

To check the prior sensitivity we have re-run our algorithm for several choices of c between 10 and 100 and the results are not that sensitive towards the choice of c. We suggest to fix a large value of c (say 100) as it is almost a non-informative prior. The number of selected genes are very sensitive towards the choice of π. On average the number of genes selected will be $m \times \pi$ where m is the total number of genes. For our case $m = 3226$ and sample size is only 23. With this small sample size we do not want to select too many genes (not more than 23) and we can restrain the number of selected genes by choosing π to be small. For example if we want to keep the number of selected genes around 23 the choice of π should be 0.007. This way the Bayesian method penalizes the number of selected genes through the help of the prior specifications. We re-analyzed the data for several choices of π from 0.001 to 0.1 which selects different number

Table 4. Breast cancer data: sensitivity analysis

Error $\sim N(0, 0.1^2)$		Error $\sim N(0, 0.2^2)$		Error $\sim N(0, 0.5^2)$	
Frequency	Image clone ID	Frequency	Image clone ID	Frequency	Image clone ID
10.5	840702*	10.5	26184*	12.7	840702*
9.5	897781*	10.1	840702*	10.4	26184*
8.6	247818*	9.6	897781*	9.0	293104*
8.3	26184*	7.6	566887	9.0	897781*
7.7	212198*	7.6	293104*	8.0	247818*
7.5	307843*	7.3	46019*	7.8	307843*
7.4	47681*	7.3	212198*	7.8	566887*
6.8	293104*	6.9	247818*	7.4	548957*
6.3	823940*	6.8	564803	7.1	46019*
5.7	566887*	6.3	788721*	7.1	810899
5.7	28012*	6.0	366647*	6.6	46182
5.6	376516*	5.9	307843*	6.5	47681*
5.5	46019*	5.9	73531*	6.4	366647*
5.4	548957*	5.8	825478*	6.4	28012*
5.3	26082*	5.8	28012*	6.3	843076*
5.3	46182	5.4	376516*	6.0	26082*
5.3	30093*	5.3	204897*	5.9	788721*
5.1	366647*	5.2	26082*	5.8	667598*
5.0	248531	5.2	248531	5.7	212198*
4.9	246524	5.2	47681*	5.6	73531*
4.9	204897*	5.1	667598*	5.5	30093*
4.7	139540	5.0	810899*	5.3	825478*
4.7	47542*	5.0	823940*	5.3	246524
4.4	32790	4.8	843076*	5.1	564803
4.4	134748	4.8	46182	5.0	248531
4.3	810899*	4.7	246524	4.9	897646
4.2	667598*	4.7	324210*	4.8	950682

∗ Selected genes which were already in the original analysis.

of genes in different cases, but the identification of the frequently arising genes remained the same.

A heat map of the identified genes is provided in Figure 1 on a website containing the figures associated with this paper http://stat.tamu.edu/people/faculty/bmallick.html.

APPLICATION TO LEUKEMIA DATA

We have applied our method to a larger data set containing a test set where we can perform our predictive validation. The leukemia data set is described in Golub *et al.* (1999). Bone marrow or peripheral blood samples are taken from 72 patients with either myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL). Following the experimental setup of the original paper, the data is split into a training set consisting of 38 samples of which 27 are ALL and 11 are AML, and a test set of 34 samples, 20 ALL and 14 AML. The data set contains expression levels for 7129 human genes produced by Affymetrix high-density oligonucleotide microarrays. The scores in the dataset represent the intensity of gene expression after being rescaled to make overall intensities. Golub *et al.* (1999) used a predictor trained using their weighted voting scheme on the training samples, and correctly classified all samples for which a prediction is made, 29 of the 34, declining to predict for the other five.

We performed our analysis with the same choices of the hyper-parameters as in the first example and report here our results. In Table 5 we provide the genes which appeared more frequently in our posterior samples. There are several genes, including the top one, that also belong to the set of 50 genes used by Golub *et al.* (these genes are reported with asterisks in Table 5).

We used the genes that appeared more than 2.5% of times (appeared to be the top 5 genes) to perform predictions on the test data. The prediction results are reported in Table 6. Only one observation is misclassified (the observation number 29). These results appear to improve predictions made by Golub *et al.* (1999) and use only 5 genes rather than 50. Figure 2 on the web site shows the heat map of the 28 genes identified by our method.

A heat map of the identified genes is provided in Figure 2 on the previously mentioned website containing the figures associated with this paper.

Table 5. Leukemia data: strongly significant genes for classification

Percentage	Frequency ID	Gene description
7.72	1882	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)*
4.85	760	CYSTATIN A
2.83	2288	DF D component of complement (adipsin)*
2.78	4847	Zyxin*
2.68	1144	1SPTAN1 Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)'
2.42	1120	SNRPN Small nuclear ribonucleoprotein polypeptide N
2.11	4535	RETINOBLASTOMA BINDING PROTEIN P48
1.98	6218	'ELA2 Elastatse 2, neutrophil'
1.95	6200	Interleukin 8 (IL8) gene *
1.95	1834	CD33 CD33 antigen (differentiation antigen)*
1.88	1630	Inducible protein mRNA*
1.79	5772	C-myb gene extracted from Human (c-myb) gene*
1.69	1745	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog*
1.67	804	Macmarcks
1.61	2354	CCND3 Cyclin D3 *
1.49	3252	'GLUTATHIONE S-TRANSFERASE, MICROSOMAL'
1.35	6201	INTERLEUKIN-8 PRECURSOR*
1.35	1685	Terminal transferase mRNA
1.31	6041	APLP2 Amyloid beta (A4) precursor-like protein 2
1.31	1779	MPO Myeloperoxidase
1.27	6855	TCF3 Transcription factor 3
1.26	173	'PRKCD Protein kinase C, delta'
1.2	2642	MB-1 gene*
1.2	1829	PPGB Protective protein for beta-galactosidase
1.19	4107	PLECKSTRIN
1.18	697	'KIAA0235 gene, partial cds'
1.17	229	KIAA0102 gene

Table 6. Leukemia data: prediction on the test set using genes with frequencies higher than 2.5%.

Y	$Pr(Y\|X_{test})$	Y	$Pr(Y\|X_{test})$
1	1.0000	1	0.2503
1	1.0000	1	1.0000
1	1.0000	1	1.0000
1	0.9972	1	0.9999
1	1.0000	1	1.0000
1	1.0000	1	1.0000
1	1.0000		
1	1.0000		
1	1.0000		
1	1.0000		
1	1.0000		
0	0.0000		
0	0.0000		
0	0.0000		
0	0.0000		
0	0.0000		
1	0.9963		
1	1.0000		
0	0.0000		
0	0.0000		
1	1.0000		
0	0.0000		
0	0.1143		
0	0.0000		
0	0.0000		
0	0.0000		
0	0.0000		
0	0.0612		

DISCUSSION

We have proposed a Bayesian model for variable selection with binary data and used it to identify important genes using their expression levels. We have used a hierarchical probit model and MCMC based stochastic search techniques to obtain the posterior samples.

Here we have fixed the π value but we can extend our model assuming π is is an unknown model parameter and assigning to it a conjugate beta prior.

We have assumed that the genes are independent though in our framework we can easily extend the model to the dependent case. For example: knowledge of the fact that if the ith gene is expressed the jth gene will be expressed too can be included in our model through the prior distribution on γ. Rather than taking all the γ_i as independently distributed we can use a Markov model whose transition matrices will be defined as $p(\gamma_j = 1|\gamma_i = 1)$ or so. This type of problem will be handled in future research.

In this paper we have considered binary data. Extension to more than two categories using multinomial models can be found in Albert and Chib (1993) and development of a variable selection model in that setup is in Sha (2002).

SUPPLEMENTARY DATA

Supplementary data are available on *Bioinformatics* online.

ACKNOWLEDGEMENTS

Marina Vannucci is partially supported by NSF Career award DMS-0093208. The research of Bani Mallick is supported by National Cancer Institute (CA-57030) grant and a grant from the Center for Environmental and Rural Health (CERH) of Texas A&M.

REFERENCES

Albert,J. and Chib,S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.*, **88**, 669–679.

Alizadeh,A., Eisen,M., Davis,R.E., Chi Ma, Lossos,I., Rosenwald,A., Boldrick,J., Sabet,H., Tran,T., Yu,X. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D. and Lavine,A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Brotherick,I., Robson,C.N., Browell,D.A., Shenfine,J., White,M.D., Cunliffe,W.J., Shenton,B.K., Egan,M., Webb,L.A., Lunt,L.G., Young,J.R. and Higgs,M.J. (1998) Cytokeratin expression in breast cancer: phenotypic changes associated with disease progression. *Cytometry*, **32**, 301–308.

Denison,D., Holmes,C., Mallick,B. and Smith,A.F.M (2002) *Bayesian methods for nonlinear classification and regression*. Wiley, London.

Dougherty,E.R. (2001) Small sample issues for microarray-based classification. *Comparative and Functional Genomics*, **2**, 28–34.

Dudoit,Y., Yang,H., Callow,M. and Speed,T. (2000) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Technical report, 578*.

Duggan,D.J., Bittner,M.L., Chen,Y., Meltzer,P.S. and Trent,J.M. (1999) Expression profiling using cdna micrarrays. *Nature Genet.*, **21**, 10–14.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Gelfand,A. and Smith,A.F.M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, **85**, 398–409.

George,E. and McCulloch,R. (1993) Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.*, **88**, 881–889.

Gilks,W., Richardson,S. and Spiegelhalter,D. (1996) *Markov Chain Monte Carlo in practise*. Chapman and Hall, London.

Golub,T.R., Slonim,D., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J., Coller,H., Loh,M., Downing,J., Caligiuri,M., Bloomfield,C. and Lender,E. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hastie,T., Tibshirani,R., Eisen,M., Alizadeh,A., Levy,R., Studt,L., Chan,W., Botstein,D. and Brown,P. (2000) Gene Shaving: a new class of clustering methods for expression arrays. *Technical Report*. Stanford University.

Hedenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner,M., Simon,R., Meltzer,P., Gusterson,B., Esteller,M., Raffeld,M. *et al.* (2001) Gene expression profiles in hereditary breast cancer. *The New England Journal of Medicine*, **344**, 539–548.

Khan,J., Simon,R., Bittner,M., Chen,Y., Leighton,S.B., Pohida,P.D., Jiang,Y., Gooden,G.C., Trent,J.M. and Meltzer,P.S. (1998) Gene Expression Profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.*, **58**, 5009–5013.

Khan,J., Wei,J., Ringner,M., Saal,L., Ladanyi,M., Westermann,F., Berthold,F., Schwab,M., Antinescu,C., Peterson,C. and Meltzer,P. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.

Kim,S., Dougherty,E.R., Barrera,J., Chen,Y., Bittner,M. and Trent,J.M. (2002) Strong feature sets from small samples C. *J. Comput. Biol.*, **9**, 127–146.

Lindley,D.V. and Smith,A.F.M. (1972) Bayes estimates for the linear models (with discussion). *J. R. Statist. Soc.*, **B34**, 1–41.

Matsuda,S., Kawamura-Tsuzuku,J., Ohsugi,M., Yoshida,M., Emi,M., Nakamura,Y., Onda,M., Yoshida,Y., Nishiyama,A. and Yamamoto,T. (1996) Tob, a novel protein that interacts with p185erbB2, is associated with anti-proliferative activity. *Oncogene*, **12**, 705–713.

McCullagh,P. and Nelder,J. (1989) *Generalized Linear Model*. Chapman and Hall, London.

Metropolis,N., Rosenbluth,A.W., Rosenbluth,M.N., Teller,A.H. and Teller,E. (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1091.

Robert,C. (1995) Simulation of truncated normal variables. *Statistics and Computing*, **5**, 121–125.

Schena,M., Shalon,D., Davis,R. and Brown,P. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Sha,N. (2002) *Bolstering CART and Bayesian variable selection methods for classification*, Ph.D. thesis, Department of Statistics, Texas A&M University.

Smith,M. and Kohn,R. (1997) Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, **75**, 317–344.

Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci.*, **98**, 5116–5121.

West,M., Nevins,J.R., Marks,J.R., Spang,R. and Zuzan,H. (2000) Bayesian regression analysis in the 'Large p, small n' paradigm with application in DNA micrarray studies. *Technical Report*. Duke University.