New Mathematics and Natural Computation
Vol. 1, No. 1 (2005) 129–145
© World Scientific Publishing Company

GENE SELECTION USING LOGISTIC REGRESSIONS BASED ON AIC, BIC AND MDL CRITERIA

XIAOBO ZHOU

*Department of Electrical Engineering, Texas A&M University
College Station, TX 77843, USA*

XIAODONG WANG

*Department of Electrical Engineering, Columbia University
New York, NY 10027, USA
wangx@ee.columbia.edu*

EDWARD R. DOUGHERTY

*Department of Electrical Engineering, Texas A&M University
College Station, TX 77843, USA*

*Department of Pathology, University of Texas M.D. Anderson Cancer Center
Houston, TX 77030, USA
edward@ee.tamu.edu*

In microarray-based cancer classification, gene selection is an important issue owing to the large number of variables (gene expressions) and the small number of experimental conditions. Many gene-selection and classification methods have been proposed; however most of these treat gene selection and classification separately, and not under the same model. We propose a Bayesian approach to gene selection using the logistic regression model. The Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the minimum description length (MDL) principle are used in constructing the posterior distribution of the chosen genes. The same logistic regression model is then used for cancer classification. Fast implementation issues for these methods are discussed. The proposed methods are tested on several data sets including those arising from hereditary breast cancer, small round blue-cell tumors, lymphoma, and acute leukemia. The experimental results indicate that the proposed methods show high classification accuracies on these data sets. Some robustness and sensitivity properties of the proposed methods are also discussed. Finally, mixing logistic-regression based gene selection with other classification methods and mixing logistic-regression-based classification with other gene-selection methods are considered.

Keywords: Gene microarray; logistic regression; Bayesian gene selection; cancer classification.

1. Introduction

Given the thousands of genes and the small number of data samples involved in microarray-based classification, gene selection is a critical issue.[15] Methods based on various algorithms have been proposed in the context of gene classification: support vector machines,[8] genetic algorithms,[14] perceptrons,[12] Bayesian variable selection,[13,23] and the minimum description length principle for model selection.[10] The logistic regression model is an important model for binary data prediction, regression and classification, and it has been successfully applied to cancer classification;[3,16,17] however, gene selection and classification based on the same logistic regression model has not been addressed, most likely because no closed form expression for the posterior distribution of the selected genes exists for logistic regression. Note that such a closed form expression exists for linear probit regression.[13] Some variable selection schemes for the logistic regression model have been proposed in the statistics literature,[4,18] but they are not suitable for problems with large numbers of variables and small sample sizes.

In this paper, we propose to use the Akaike information criterion, the Bayesian information criterion and the minimum description length principle to construct the posterior distribution of the selected genes for the logistic regression model. A Gibbs sampler is employed to find the strongest genes based on such a posterior distribution. Since these methods have very high computational complexities, we also discuss some numerical techniques to speed up the computation. Furthermore, a gene pre-selection procedure is adopted to reduce the huge number of genes being considered for selection. After finding the strongest genes, we perform classification based on the strongest genes using the estimated logistic regressions. The proposed method is tested on several data sets including those from hereditary breast cancer, small round blue-cell tumors, lymphoma, and acute leukemia. The experimental results show that the proposed methods can effectively find important genes consistent with the biological considerations, and the classification accuracies are very high. Some robustness and sensitivity properties for the proposed methods are also discussed. Since gene selection and classification are separate (but related) tasks, we pair three Bayesian selection methods with four classifier methods using the heritary breast cancer data. In particular, we wish to see how using different models for gene selection and classification affects the results.

2. Problem Formulation

Assume we are interested in classifying whether a particular cancer is present or not. Let $\boldsymbol{y} = [y_1, \ldots, y_m]^T$ denote the class labels, where $y_i = 1$ indicates sample i has the cancer, and $y_i = 0$ indicates sample i does not have the cancer. Denote $x_1, \ldots, x_n$ as the expression levels of n genes. Let $x_{i,j}$ be the measurement of the expression level of the jth gene for the ith sample. Let $\boldsymbol{X} = (x_{i,j})_{m,n}$ denote the expression levels of all genes, i.e.

$$\boldsymbol{X} = \begin{bmatrix} \text{Gene 1} & \text{Gene 2} & \cdots & \text{Gene } n \\ x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix}. \tag{1}$$

Let $\boldsymbol{x}_i \triangleq [x_{i,1}, x_{i,2}, \ldots, x_{i,n}]$ denote the ith row of the above matrix. We model $\pi_i = P(y_i = 1|\boldsymbol{X})$ by using a logistic regression model given by

$$\log \frac{\pi_i}{1 - \pi_i} = \boldsymbol{x}_i\boldsymbol{\beta} \triangleq x_{i,1}\beta_1 + \cdots + x_{i,n}\beta_n, \quad i = 1, \ldots, m, \tag{2}$$

where $\boldsymbol{\beta} \triangleq [\beta_1, \ldots, \beta_n]^T$ contains the regression coefficients. Equivalently, (2) can be rewritten as $\pi_i = (1 + \exp(-\boldsymbol{x}_i\boldsymbol{\beta}))^{-1}$. The likelihood function of the logistic regression is

$$L(\boldsymbol{\beta}) \triangleq \prod_{i=1}^{m} \left[\frac{1}{1 + \exp(-\boldsymbol{x}_i\boldsymbol{\beta})}\right]^{y_i} \left[\frac{1}{1 + \exp(\boldsymbol{x}_i\boldsymbol{\beta})}\right]^{(1-y_i)}. \tag{3}$$

The corresponding log-likelihood function is given by

$$\log L(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} - \sum_{i=1}^{m} \log\bigl(1 + \exp(\boldsymbol{x}_i\boldsymbol{\beta})\bigr). \tag{4}$$

Then iterative methods such as the Newton-Raphson procedure can be adopted to obtain the maximum likelihood estimate of $\boldsymbol{\beta}$.[6]

Define $\boldsymbol{\gamma}$ as the $n \times 1$ indicator vector with the jth element γ_j such that $\gamma_j = 0$ if $\beta_j = 0$ (the variable is not selected) and $\gamma_j = 1$ if $\beta_j \neq 0$ (the variable is selected). The Bayesian variable selection is to estimate $\boldsymbol{\gamma}$ from the posterior distribution $p(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X})$. Given $\boldsymbol{\gamma}$, let $\boldsymbol{\beta}_\gamma$ consists of all nonzero elements of $\boldsymbol{\beta}$ and let $\boldsymbol{X}_\gamma$ be the columns of $\boldsymbol{X}$ corresponding to those $\boldsymbol{\gamma}$ that are equal to 1. Then (2) is rewritten as

$$\log \frac{\boldsymbol{\pi}}{1 - \boldsymbol{\pi}} = \boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma, \tag{5}$$

where $\log \frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}} \triangleq \left[\log \frac{\pi_1}{1-\pi_1}, \log \frac{\pi_2}{1-\pi_2}, \ldots, \log \frac{\pi_m}{1-\pi_m}\right]^T$. Now the problem is how to estimate $\boldsymbol{\gamma}$ and the corresponding $\boldsymbol{\beta}_\gamma$. Note that no closed form expression exists for the posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{X})$, and neither for $p(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X})$. In what follows, we propose to construct the posterior distribution $p(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X})$ based on information criteria such as the AIC, the BIC and the MDL.

3. Bayesian Gene Selection Based on AIC, BIC or MDL

The indicator vector $\boldsymbol{\gamma}$ can be modeled as a realization from any prior $p(\boldsymbol{\gamma})$ on the 2^n possible values of $\boldsymbol{\gamma}$ given by $p(\boldsymbol{\gamma}) = \prod_{i=1}^{n} \nu_i^{\gamma_i}(1-\nu_i)^{(1-\gamma_i)}$, where $\nu_i = P(\gamma_i = 1)$

is a prior probability to select the jth gene. This form is actually a Bernoulli distribution for selecting each gene. Now we define *a posterior* probability distribution for $p(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X})$ as

$$p(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X}) \propto \exp\{-S(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X})\} \prod_{i=1}^{n} \nu_i^{\gamma_i}(1-\nu_i)^{(1-\gamma_i)}. \tag{6}$$

We next specify the form of $S(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X})$ based on the AIC, BIC and MDL citeria as follows.

- The Akaike information criterion (AIC). The AIC was first introduced in Ref. 1 to measure a model fitting accuracy. It has been shown that an asymptotically unbiased estimator of an essential part of the relative entropy can be obtained as the negative maximum log-likelihood plus a penalty term equal to the dimension of the parameters in the employed model. For the logistic regression model, it is given by

  $$S(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X}) = -\log L(\boldsymbol{\beta}_\gamma) + n_\gamma, \tag{7}$$

  where $\boldsymbol{\beta}_\gamma$ is the maximum likelihood estimate of the logistic regression, and $n_\gamma \triangleq \sum_{j=1}^{n} \gamma_j$.
- The Bayesian information criterion (BIC). The model utility can also assessed by a Bayesian approach using a uniform prior for the candidate models.[20] The posterior distribution of the candidate models is proportional to the negative maximum log-likelihood plus a penalty term. For the logistic regression model, it is given by

  $$S(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X}) = -\log L(\boldsymbol{\beta}_\gamma) + \frac{n_\gamma}{2}\log m. \tag{8}$$
- The minimum description length (MDL) principle. The stochastic complexity based model selection was introduced in Ref. 19. It becomes a well known model selection criterion, i.e. the minimum description length principle. For the logistic regression model, it is approximated by Ref. 18

  $$S(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X}) \approx -\log L(\boldsymbol{\beta}_\gamma) + \frac{1}{2}\log|\boldsymbol{I}_n(\boldsymbol{\beta}_\gamma)| + \sum_{j=1}^{n_\gamma} \log|\boldsymbol{\beta}_\gamma|, \tag{9}$$

  $$\text{with } \boldsymbol{I}_n(\boldsymbol{\beta}_\gamma) = \boldsymbol{X}_\gamma^T \mathrm{diag}\{\pi_1(1-\pi_1), \ldots, \pi_m(1-\pi_m)\}\boldsymbol{X}_\gamma.$$

  Here $\boldsymbol{I}_n(\boldsymbol{\beta})$ is the Fisher information matrix for $\boldsymbol{\beta}_\gamma$.

Obviously, the above three criteria have a common form: $S(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X}) \triangleq -\log L(\boldsymbol{\beta}_\gamma) + g(\boldsymbol{\beta}, \boldsymbol{\gamma}, m)$, where $g(\boldsymbol{\beta}, \boldsymbol{\gamma}, m)$ is given in (7)–(9) respectively for the three different criteria.

Based on the posterior distribution (6), a Gibbs sampler can be employed to estimate all the parameters. We use the following Gibbs sampling algorithm to estimate $\{\boldsymbol{\gamma}, \boldsymbol{\beta}_\gamma\}$.

- Draw $\boldsymbol{\gamma}^{(t)}$ from $p(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X})$ in (6). Here we set as $\nu_j = 15/n$ based on the total number of samples m that we have. If ν_j is chosen as a larger value, then we have found that often times $(\boldsymbol{X_\gamma}^T \boldsymbol{X_\gamma})^{-1}$ (which will be used in estimating $\boldsymbol{\beta}_\gamma$) does not exist. In particular, we sample each $\gamma_j^{(t)}$ independently from

$$p(\gamma_j|\boldsymbol{y}, \boldsymbol{X}, \gamma_{i\neq j}) \propto \exp\{-S(\gamma|\boldsymbol{y}, \boldsymbol{X})\}\nu_i^{\gamma_i}(1-\nu_i)^{(1-\gamma_i)}, \quad j = 1, \ldots, n. \quad (10)$$

- Given the sample of $\boldsymbol{\gamma}^{(t)}$, obtain the maximum likelihood estimate of $\boldsymbol{\beta}_\gamma$ using the method given in Ref. 6.

In this study, 25,000 Gibbs iterations are implemented with the first 5,000 as burn-in period. We obtain the Monte Carlo samples as $\{\boldsymbol{\gamma}^{(t)},\ t = 1, \ldots, T\}$, where $T =$ 25,000. Finally, we count the number of times that each gene appears in $\{\boldsymbol{\gamma}^{(t)},\ t =$ $5{,}001, \ldots, 25{,}000\}$. We define the appearance frequency of a gene as the number of appearances of this gene divided by the total iteration (i.e. 20,000 here). The genes with the highest appearance frequencies play the strongest role in predicting the target gene. We will discuss some implementation issues in the next section.

Bayesian Estimation Using the Strongest Genes: Now assume the genes corresponding to non-zero $\boldsymbol{\gamma}$ are the strongest genes obtained by the above Bayesian variable-selection algorithm. We still use $\boldsymbol{x}_\gamma$ to denote the profiles of these strongest genes. After estimating $\boldsymbol{\beta}_\gamma$ using the maximum-likelihood estimation method, we predict the tested sample by

$$P(y = 1|\boldsymbol{x}_\gamma, \boldsymbol{\beta}_\gamma) = \frac{\exp\{\boldsymbol{x}_\gamma\boldsymbol{\beta}_\gamma\}}{1 + \exp\{\boldsymbol{x}_\gamma\boldsymbol{\beta}_\gamma\}}. \quad (11)$$

4. Fast Implementation

If there are 3,000 gene variables, then for each iteration we have to estimate $\boldsymbol{\beta}_\gamma$ 3,000 times because we need to sample γ_j for each gene according to (10). The computational complexity of the Bayesian gene selection algorithm in the previous section is very high. Hence, some fast algorithms must be developed to speed up the computation.

4.1. *Pre-selection method*

Suppose that the total number of genes is p, and we will only consider $n < p$ candidates in the Bayesian selection algorithm. We next discuss how to pre-select the n genes using an F-test. In pattern recognition, we usually adopt the following criterion: the smaller is the sum of squares within groups and the bigger is the sum of squares between groups, the better is the classification accuracy. Therefore, we can define a score using the above two statistics to pre-select genes, i.e the ratio of the between-group to within-group sum of squares:

$$R(j) \triangleq \frac{\sum_{i=1}^{m}\sum_{k=0}^{K-1} 1_{(y_i=k)}(\bar{x}_{k,j} - \bar{x}_j)^2}{\sum_{i=1}^{m}\sum_{k=0}^{K-1} 1_{(y_i=k)}(x_{i,j} - \bar{x}_{k,j})^2}, \quad 1 \le j \le p, \quad (12)$$

where K the number of classes; p is the total number of original genes (note that the number of genes n used in the Bayesian selection procedure is much smaller than p); $\bar{x}_j$ denotes the average expression level of gene j across all samples; and $\bar{x}_{k,j}$ denotes the average expression level of gene j across the samples belonging to class k where class k corresponds to $\{y_i = k\}$; and the indicator function 1_Ω is equal to one if event Ω is true and zero otherwise. We select a threshold Γ and keep those genes j such that $R(j) \geq \Gamma$. The pre-selection procedure yields n genes such that $R(j) \geq \Gamma$.

Computation of $p(\gamma_j|\boldsymbol{y}, \boldsymbol{X}, \gamma_{i\neq j})$ in (10)

Because γ_j only takes 0 or 1, we can re-consider $p(\gamma_j = 1|\boldsymbol{y}, \boldsymbol{X}, i \neq j)$ and $p(\gamma_j = 0|\boldsymbol{y}, \boldsymbol{X}, i \neq j)$. Let $\boldsymbol{\gamma}^1 = (\gamma_1, \ldots, \gamma_{j-1}, \gamma_j = 1, \gamma_{j+1}, \ldots, \gamma_n)$ and $\boldsymbol{\gamma}^0 = (\gamma_1, \ldots, \gamma_{j-1}, \gamma_j = 0, \gamma_{j+1}, \ldots, \gamma_n)$. According to (10),

$$p(\gamma_j = 1|\boldsymbol{y}, \boldsymbol{X}, \gamma_{i\neq j}) \propto \exp\{-S(\boldsymbol{\gamma}^1|\boldsymbol{y}, \boldsymbol{X})\}\nu_j$$

$$p(\gamma_j = 0|\boldsymbol{y}, \boldsymbol{X}, \gamma_{i\neq j}) \propto \exp\{-S(\boldsymbol{\gamma}^0|\boldsymbol{y}, \boldsymbol{X})\}(1 - \nu_j).$$

Since $p(\gamma_j = 1|\boldsymbol{y}, \boldsymbol{X}, \gamma_{i\neq j}) + p(\gamma_j = 0|\boldsymbol{y}, \boldsymbol{X}, \gamma_{i\neq j}) = 1$, some straightforward computation yields

$$p(\gamma_j = 1|\boldsymbol{y}, \boldsymbol{X}, \gamma_{i\neq j}) \propto \frac{1}{1+h}, \tag{13}$$

$$\text{with } h = \frac{1-\nu_j}{\nu_j} \exp\left[S(\boldsymbol{\gamma}^1|\boldsymbol{y}, \boldsymbol{X}) - S(\boldsymbol{\gamma}^0|\boldsymbol{y}, \boldsymbol{X})\right]. \tag{14}$$

If $\boldsymbol{\gamma} = \boldsymbol{\gamma}^0$ before γ_j is generated, meaning we have obtained $S(\boldsymbol{\gamma}^0|\boldsymbol{y}, \boldsymbol{X})$, then we only need to compute $S(\boldsymbol{\gamma}^1|\boldsymbol{y}, \boldsymbol{X})$, and vice versa. We summarize our fast Bayesian gene selection algorithm based on Gibbs sampling as follows.

Algorithm 1 [Fast Bayesian gene selection algorithm]

```
• Pre-select genes according to (12);
• Initialization: Randomly set initial parameters $\boldsymbol{\gamma}^{(0)}$;
• For $t = 1, 2, \ldots, 25{,}000$
    —Draw $\boldsymbol{\gamma}^{(t)}$. For $j = 1, \ldots, n$
        * Compute $S(\boldsymbol{\gamma}^{(t)}|\boldsymbol{y}, \boldsymbol{X})$.
        * Compute $p(\gamma_j = 1|\boldsymbol{y}, \boldsymbol{X}, \gamma^{(t)}_{i\neq j})$ according to (13).
        * Draw $\gamma^{(t)}_j$ from $p(\gamma_j = 1|\boldsymbol{y}, \boldsymbol{X}, \gamma^{(t)}_{i\neq j})$.
        * Estimate $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ using maximum likelihood method.
• Endfor
• Count the frequency of each gene appeared in $\boldsymbol{\gamma}^{(t)}, t = 5{,}001, \ldots, 25{,}000$.
```

5. Experimental Results

5.1. *Breast cancer data*

In our first experiment, we will focus on hereditary breast cancer data, which can be downloaded from the web page for the original paper.[9] In Ref. 9, cDNA microarrays are used in conjunction with classification algorithms to show the feasibility of using differences in global gene expression profiles to separate BRCA1 and BRCA2 mutation-positive breast cancers. 22 breast tumor samples from 21 patients were examined: 7 BRCA1, 8 BRCA2, and 7 sporadic. There are 3,226 genes for each tumor sample. We use our methods to classify BRCA1, BRCA2 and sporadic. The ratio data has been truncated from below at 0.1 and above at 20. The cross-validation (leave-one-out) method is employed to compute all classification errors in this paper.

Table 1 lists the strongest genes using the AIC criterion. Gene 1,008 (Clone ID: 897781, keratin 8) is the strongest gene. This is consistent with other references.[9,12,13] Keratin 8 is a member of the cytokeratin family of genes. Cytokeratins are frequently used to identify breast cancer metastases by immunohistochemistry. The gene TOB1 (Clone ID 823940) is also a key gene listed in Table 1.[12,13] The BIC and MDL criteria yield the same top three genes (Table 2).

Using the top 5, 10 and 15 genes for classification, it is seen that the classification error based 5 genes and 10 genes is zero in Table 4. There is one error using 15 genes, which is likely due to the small sample size. The conditional probabilities based the three criteria using top 10 genes are listed in Table 3. These are very close to the true label values (namely, 0 and 1).

Table 1. The top 20 important genes selected using AIC for breast cancer data ($\nu_i = 15/n$).

Gene No.	Frequency	Index No. (Clone ID)	Gene Description
1	0.1454	1008 (897781)	Keratin 8
2	0.1394	496 (376516)	Cell division cycle 4-like
3	0.1340	336 (823940)	Transducer of ERBB2, 1 (TOB1)
4	0.1331	2699 (44180)	Alpha-2-macroglobulin
5	0.1240	2761 (47884)	Macrophage migration inhibitory factor (glycosylation-inhibiting factor)
6	0.1167	742 (183200)	Fumarylacetoacetate
7	0.1044	2382 (21652)	Catenin (cadherin-associated protein), alpha 1 (102kD)
8	0.1003	2018 (139354)	ESTs
9	0.9025	157 (809981)	Glutathione peroxidase 4 (phospholipid hydroperoxidase)
10	0.0545	739 (214068)	GATA-binding protein 3
11	0.0483	1120 (841617)	Human mRNA for ornithine decarboxylase antizyme, ORF 1 and ORF 2
12	0.0473	2272 (309583)	ESTs
13	0.0472	1620 (137638)	ESTs
14	0.0463	1999 (247818)	ESTs
15	0.0433	1859 (307843)	ESTs
16	0.0426	439 (160793)	Discs, large (Drosophila) homolog 1
17	0.0424	2734 (46019)	Minichromosome maintenance deficient (*S. cerevisiae*) 7
18	0.0419	247 (725680)	Transcription factor AP-2 gamma
19	0.0414	3009 (366647)	Butyrate response factor 1 (EGF-response factor 1)
20	0.0405	2423 (26082)	Very low lipoprotein receptor

Table 2. The top 20 important genes selected using BIC and MDL for breast cancer data ($\nu_i = 15/n$).

Gene No.	BIC		MDL	
	Frequency	Index No. (Clone ID)	Frequency	Index No. (Clone ID)
1	0.1640	1008 (897781)	0.1427	1008 (897781)
2	0.1638	496 (376516)	0.1409	336 (823940)
3	0.1638	336 (823940)	0.1280	496 (376516)
4	0.1437	2382 (21652)	0.1175	742 (183200)
5	0.1255	2761 (47884)	0.0520	2761 (47884)
6	0.1253	2699 (44180)	0.1090	1120 (841617)
7	0.1241	742 (183200)	0.1051	2699 (44180)
8	0.1039	2018 (139354)	0.1007	2018 (139354)
9	0.0838	157 (809981)	0.0962	157 (809981)
10	0.0674	1120 (841617)	0.0520	2382 (21652)
11	0.0459	2272 (309583)	0.0397	1999 (247818)
12	0.0391	2734 (46019)	0.0385	2761 (47884)
13	0.0384	2423 (26082)	0.0370	2734 (46019)
14	0.0379	1443 (566887)	0.0366	2272 (309583)
15	0.0372	1228 (796137)	0.0357	3009 (366647)
16	0.0359	1628 (233365)	0.0352	1620 (137638)
17	0.0339	247 (725680)	0.0345	1446 (81331)
18	0.0338	1797 (144926)	0.0344	3013 (375922)
19	0.0335	523 (28012)	0.0340	1531 (711826)
20	0.0333	2833 (488801)	0.0338	585 (293104)

5.2. *Small round blue-cell tumors*

This experiment focuses on the small, round blue cell tumors (SRBCTs) of childhood, which include neuroblastoma (NB), rhabdomyosarcoma (RMS), non-hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS) in Ref. 11. We classify the rhabdomyosarcoma and neuroblastoma tumors. The data set for the two cancers is composed of 2,308 genes, and the sample consists of 35 tumors, 23 for RMS and 12 for NB. The ratio data has been truncated from below at 0.01.

Table 5 lists the strongest genes using the AIC criterion. Gene 2050 (Clone ID 295985) is the strongest for all methods. It is also an important gene in Ref. 11. A number of other previously noted genes also appear[11,22]: 246 (Clone ID 377461), 545 (Clone ID 1435862), 255 (clone ID 325182), 1389 (Clone ID 770394), 2144 (Clone ID 308231), 742 (Clone ID 812105), 867 (Clone ID 784593), 153 (Clone ID 383188), and 1601 (Clone ID 629896). Using the top 5, 10 and 15 genes for classification based on the three criteria, no error is found (Table 4).

5.3. *Lymphoma data*

The lymphoma data can be found in the original paper,[2] which consists of gene expressions from cDNA experiments involving three prevalent adult lymphoid malignancies: DLBCL, BCLL and Follicular Lymphoma (FL). We have analyzed

Table 3. The estimated probabilities of each sample for breast cancer data using AIC, BIC and MDL ($\nu_i = 15/n$).

Sample Index No.	True Label y	AIC $P(y = 1\|X)$	BIC $P(y = 1\|X)$	MDL $P(y = 1\|X)$
1	0	0.0000	0.0000	0.0021
2	0	0.0000	0.0000	0.0000
3	0	0.0000	0.0000	0.0000
4	0	0.0021	0.0000	0.0068
5	0	0.0000	0.0000	0.0000
6	0	0.0025	0.0000	0.0207
7	1	1.0000	1.0000	0.9971
8	1	0.9999	0.9973	1.0000
9	1	1.0000	1.0000	1.0000
10	1	1.0000	1.0000	1.0000
11	1	0.9982	1.0000	1.0000
12	1	0.9838	1.0000	1.0000
13	1	1.0000	1.0000	0.9998
14	1	1.0000	1.0000	1.0000
15	1	0.9996	1.0000	0.9409
16	1	1.0000	1.0000	1.0000
17	1	1.0000	1.0000	1.0000
18	0	0.0030	0.0001	0.0001
19	1	1.0000	1.0000	1.0000
20	1	1.0000	1.0000	1.0000
21	1	0.9742	1.0000	0.9960
22	1	0.9999	1.0000	1.0000
No of misclassification		0	0	0

Table 4. The number of misclassification using AIC, BIC and MDL ($\nu_i = 15/n$) with 5, 10 and 15 genes for the breast cancer data, the SRBCT data, the lymphoma data, and the leukemia data, respectively.

No. of Genes	Breast			SRBCT			Lymphoma			Leukemia		
	AIC	BIC	MDL	AIC	BIC	MDL	AIC	BIC	MDL	AIC	BIC	MDL
5	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0
15	1	1	1	0	0	0	1	0	0	0	1	0

the log relative-intensity ratios. To test the gene selection methods, we consider a subset of the data consisting of 45 DLBCL and 29 BCLL cases, with 9,216 genes (spots listed in the authors's web page^2).

Table 6 lists the strongest genes using the BIC criterion. It is seen that genes 4612, 5164, 1279, 3165, 103, 555, 2288, 5996, 4588, and 1836 are most important. Using the top 5, 10 and 15 genes for classification based on the three criteria, no error is found except one error for 15 genes using the AIC criterion (Table 4).

Table 5. The top 20 important genes selected using AIC for SRBCT data ($\nu_i = 15/n$).

Gene No.	Frequency	Index No. (Clone ID)	Gene Description
1	0.1168	2050 (295985)	ESTs
2	0.1162	246 (377461)	Caveolin 1, caveolae protein, 22kD
3	0.1156	545 (1435862)	Antigen identified by monoclonal antibodies 12E7, F21 and O13
4	0.1153	1662 (377048)	Homo sapiens incomplete cDNA for a mutated allele of a myosin class I, myh-1c
5	0.1145	842 (810057)	Cold shock domain protein A
6	0.1137	437 (448386)	No name
7	0.1128	255 (325182)	Cadherin 2, N-cadherin (neuronal)
8	0.1127	1389 (770394)	Fc fragment of IgG, receptor, transporter, alpha
9	0.1120	566 (357031)	Tumor necrosis factor, alpha-induced protein 6
10	0.1110	1873 (166195)	Ribonuclease/angiogenin inhibitor
11	0.1107	2144 (308231)	Homo sapiens incomplete cDNA for a mutated allele of a myosin class I, myh-1c
12	0.1105	742 (812105)	Transmembrane protein
13	0.1103	867 (784593)	ESTs
14	0.1100	1579 (204299)	Replication protein A3 (14kD)
15	0.1094	365 (1434905)	Homeo box B7
16	0.1091	976 (786084)	Chromobox homolog 1 (Drosophila HP1 beta)
17	0.1081	1954 (814260)	Follicular lymphoma variant translocation 1
18	0.1076	153 (383188)	Recoverin
19	0.1068	1601 (629896)	Microtubule-associated protein 1B
20	0.1060	823 (134748)	Glycine cleavage system protein H (aminomethyl carrier)

5.4. ***Acute leukemia data***

The leukemia data of Ref. 7 is publicly available at http://www-genome.wi.mit.edu/cgi-bin/cancer/publications/pub. The microarray data contains 7,129 human genes, sampled from 72 cases of cancer, of which 38 are of type B-cell ALL, 9 are of type T-cell ALL, and 25 of type AML. The data are preprocessed as recommended in Refs. 5 and 21: gene values are truncated from below at 100 and from above at 16,000; genes having the ratio of the maximum over the minimum less than 5 or the difference between the maximum and the minimum less than 500 are excluded; and finally the base-10 logarithm is applied to the 3,571 remaining genes. Here we consider the full 72-tumor sample, splitting it between ALL (47) and AML (25).

Table 7 lists the 20 strongest genes based on the MDL principle. The index number is the Clone ID in this data set. Genes 6345, 5402, 2056, 1144 and 1551 are the strongest. Genes 1144, 1120, 4535 and 3252 are also listed.[13] Using the top 5, 10 and 15 genes for classification based on the three criteria, no error is found except one error for 15 genes using the BIC criterion (Table 4).

Table 6. The top 20 important genes selected using BIC for lymphoma data ($\nu_i = 15/n$).

Gene No.	Frequency	Index No. (Spot No.)
1	0.1361	4612
2	0.1237	5164
3	0.1093	1279
4	0.1024	3165
5	0.0976	103
6	0.0863	555
7	0.0735	2288
8	0.0645	5996
9	0.0642	4588
10	0.0628	1836
11	0.0621	1734
12	0.0616	7497
13	0.0595	8049
14	0.0585	2286
15	0.0578	3130
16	0.0578	2286
17	0.0571	8320
18	0.0551	3434
19	0.0544	6421
20	0.0537	4613

5.5. *Sensitivity and robustness*

To check the sensitivity and robustness of our algorithms, we have added white Gaussian noise with different variances to the data and re-applied our algorithms to the contaminated data. The strongest genes are listed in Table 8. It is seen that genes 1008 (keratin 8) and 336 TOB1 remain very important for different noise levels. The results indicate that the proposed methods are not overly sensitive to the different noise levels.

To check the sensitivity to the prior distributions, we have re-run the algorithms for $\nu_i = 10/n$. According to Table 9, the selected genes are almost the same as before, thereby providing evidence of robustness relative to the prior setting.

Finally, we analyzed the proposed methods based on the natural-based log ratio of the breast cancer data.[9] Here, the important genes are quite different from the preceding results. From Table 10, it is seen gene 10 (phosphofructokinase, platelet) is the most important gene for all methods. It is also a key gene in Refs. 12 and 13. Whereas TOB1 is still listed among the 20 strongest genes, keratin 8 is not. Using the top ten genes for classification, no error is found based on any of the three criteria.

5.6. *Comparisons*

Various gene selection methods and classifiers for cancer classification have been proposed. In particular, there is strong evidence that Bayesian gene selection is

Table 7. The top 20 important genes selected using MDL for acute leukemia data ($\nu_i = 15/n$).

Gene No.	Frequency	Index No.	Gene Description
1	0.1219	6345	GLUL Glutamate-ammonia ligase (glutamine synthase)
2	0.1214	5402	MST1 Macrophage stimulating 1 (hepatocyte growth factor-like)
3	0.1202	2056	Heparin cofactor II (HCF2) gene, exons 1 through 5
4	0.1199	1144	GB DEF = Sialoprotein mRNA
5	0.1197	1551	RPL37 Ribosomal protein L37
6	0.1171	1903	Recombination activating protein (RAG-1) gene
7	0.1156	1120	G22P1 Thyroid autoantigen 70kD (Ku antigen)
8	0.1150	4328	MCP Membrane cofactor protein
9	0.1143	4142	RNH Ribonuclease/angiogenin inhibitor
10	0.1132	4781	Uridine phosphorylase
11	0.1131	1745	C-yes-1 mRNA
12	0.1124	6215	TNNT1 Troponin T1, skeletal, slow
13	0.1098	6797	GYPB Glycophorin B
14	0.1068	4535	SSR2 Signal sequence receptor, beta
15	0.1067	3320	Guanine nucleotide exchange factor p532 mRNA
16	0.1050	1208	Protein tyrosine kinase related mRNA sequence
17	0.1036	3258	Cystatin B gene
18	0.1010	3252	GB DEF = Dishevelled homolog (DVL) mRNA
19	0.0987	2242	INTEGRAL MEMBRANE PROTEIN E16
20	0.0968	6919	Skeletal beta-tropomyosin

effective.[13,24] Regarding classification, the linear probit (LProbit),[13] nonlinear probit (NLProbit),[24] and kNN classifiers[5,13,25] have proved effective. Using the breast-cancer data set, we will compare the performance of these classifiers when used in conjunction with the previously proposed Bayesian gene-selection methods and the logistic method developed in this paper. We summarize linear-probit-based and mutual-information-based gene selection, along with the corresponding classifiers:

- Probit gene selection and classification[13]: the relation between the class label y_i and the gene expression levels $\boldsymbol{x}_i$ is modeled by using a probit regression model which yields $P(y_i = 1|\boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i\boldsymbol{\beta}), \ \ i = 1, \ldots, m$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)^T$ is the vector of regression parameters and Φ is the standard normal cumulative distribution function. Gene selection based on probit regression is similar to that of logistic regression using Gibbs sampling. The difference is the posterior distribution of $p(\boldsymbol{\gamma}|\boldsymbol{z})$, see Ref. 13. After obtaining the strongest genes, we can estimate $P(y = 1|\boldsymbol{X})$ using Gibbs sampling for the probit regression classifier.

Table 8. The top 20 important genes selected using AIC for breast cancer data for different noise levels ($\nu_i = 15/n$).

Gene No.	$\sigma = 0.1$		$\sigma = 0.2$		$\sigma = 0.5$	
	Frequency	Index No. (Clone ID)	Frequency	Index No. (Clone ID)	Frequency	Index No. (Clone ID)
1	0.1427	336 (823940)	0.1334	1008 (897781)	0.1953	1008 (897781)
2	0.1409	1008 (897781)	0.1278	2018 (21652)	0.1301	496 (376516)
3	0.1280	496 (376516)	0.1252	336 (823940)	0.1266	336 (823940)
4	0.1175	742 (183200)	0.1199	496 (376516)	0.1164	2699 (44180)
5	0.1172	2382 (21652)	0.0974	2699 (44180)	0.1038	739 (214068)
6	0.1090	1120 (841617)	0.0956	742 (183200)	0.0993	157 (809981)
7	0.1051	2699 (44180)	0.0936	739 (214068)	0.0762	94 (191603)
8	0.1007	109 (810873)	0.0796	67 (50359)	0.0761	1446 (81331)
9	0.0962	157 (809981)	0.0794	2382 (21652)	0.0732	742 (183200)
10	0.0520	739 (214068)	0.0769	157 (809981)	0.0698	2382 (21652)
11	0.0397	1999 (247818)	0.0673	2732 (45840)	0.0686	883 (79898)
12	0.0385	2761 (47884)	0.0637	2272 (309583)	0.0682	2321 (240208)
13	0.0370	2734 (46019)	0.0627	1859 (244974)	0.0681	2027 (161195)
14	0.0366	2272 (309583)	0.0595	1200 (811930)	0.0663	489 (133178)
15	0.0357	3009 (366647)	0.0593	498 (667598)	0.0659	533 (345208)
16	0.0352	1620 (137638)	0.0589	118 (47542)	0.0640	1859 (244974)
17	0.0345	1446 (81331)	0.0579	94 (191603)	0.0635	1179 (788721)
18	0.0344	3013 (375922)	0.0573	1443 (566887)	0.0627	1851 (293977)
19	0.0340	1531 (711826)	0.0572	2761 (47884)	0.0616	842 (813280)
20	0.0338	585 (293104)	0.0561	1179 (788721)	0.0614	275 (242037)

- Mutual-information-based gene selection and nonlinear probit classifier[24]: given an initial set $V = \{X_1, X_2, \ldots, X_N\}$ with N random variables and the class variable C, the genes are selected according to the mutual information $I(C; X)$ maximization criterion.[24] The nonlinear probit classifier is defined as follows: the y_i and the gene expression levels are related through $P(y_i = 1|x_1, \ldots, x_n) = \Phi\left(\sum_{i=1}^{n} \alpha_i x_i + \sum_{k=1}^{2} \beta_k \phi_k(x_1, \ldots, x_n)\right)$, with $\phi_k(x_1, \ldots, x_n) \triangleq \exp\{-\lambda_k \|\boldsymbol{x} - \boldsymbol{\mu}_k\|^2\}$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_P)^T$ and $\boldsymbol{\beta} = [\beta_1, \beta_2]^T$ are regression parameters, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the centers of the two clusters obtained by using the fuzzy c means clustering algorithm, and the parameters λ_1 and λ_2 are empirically set as 2.0 and 4.0 respectively. The parameters can be estimated by using Gibbs sampling.

Table 11 lists the top 20 genes selected by probit regression and mutual information, respectively. Gene 1008 (Clone ID: 897781, keratin 8) is an important gene for both methods, but many genes are different. The reason is that the probit and logistic selection methods are subset-based, whereas the mutual-information-based method is a single-gene ranking method. The principles for gene selection are quite different. The misclassification numbers using the four classifiers (logit, LProbit, NLProbit, kNN) for three gene selection methods (logit, probit, MI) with five and

Table 9. The top 20 important genes selected using AIC, BIC and MDL for breast cancer data ($\nu_i = 10/n$).

Gene No.	AIC		BIC		MDL	
	Frequency	Index No. (Clone ID)	Frequency	Index No. (Clone ID)	Frequency	Index No. (Clone ID)
1	0.1409	336 (823940)	0.1624	496 (376516)	0.1573	1008 (897781)
2	0.1339	1008 (897781)	0.1560	336 (823940)	0.1538	336 (823940)
3	0.1300	496 (376516)	0.1422	1008 (897781)	0.1489	496 (376516)
4	0.1215	2382 (21652)	0.1117	2382 (21652)	0.1149	1120 (841617)
5	0.1127	2699 (44180)	0.1106	2699 (44180)	0.1119	2699 (44180)
6	0.1126	2018 (139354)	0.1103	1120 (841617)	0.1013	2018 (139354)
7	0.1063	742 (183200)	0.0994	2018 (139354)	0.01007	742 (183200)
8	0.1052	1120 (841617)	0.0945	742 (183200)	0.0979	2382 (21652)
9	0.0957	157 (809981)	0.0808	157 (809981)	0.0837	157 (809981)
10	0.0495	739 (214068)	0.0353	2761 (47884)	0.0386	2761 (47884)
11	0.0447	2761 (47884)	0.0345	1999 (247818)	0.0343	1999 (247818)
12	0.0437	2272 (309583)	0.0338	2734 (46019)	0.0331	1443 (566887)
13	0.0418	3009 (366647)	0.0314	10 (26184)	0.0318	739 (214068)
14	0.0411	1999 (247818)	0.0309	2272 (309583)	0.0314	2272 (309583)
15	0.0380	2734 (46019)	0.0298	739 (214068)	0.0299	1417 (825478)
16	0.0379	809 (810899)	0.0294	1797 (144926)	0.0298	809 (810899)
17	0.0369	1859 (307843)	0.0294	158 (204897)	0.0297	2734 (46019)
18	0.0367	94 (191603)	0.0288	489 (133178)	0.0297	1859 (307843)
19	0.0366	2833 (488801)	0.0285	3080 (280768)	0.0295	2833 (488801)
20	0.0364	1288 (564803)	0.0281	1443 (566887)	0.0284	10 (26184)

ten top genes are listed in Table 12. No error is found for all of the classifiers based on logit selection. Moreover, no error is found for the NLProbit classifier using all three gene selection methods.

6. Conclusion

This paper has investigated Bayesian gene selection using the logistic regression model where the posterior distribution of the selected genes is constructed using the Akaike information criterion, the Bayesian information criterion and the minimum description length principle. Once important genes are identified, the same logistic regression model is employed for cancer classification. Fast implementation issues for these methods are discussed. The proposed methods are tested on data sets arising from ereditary breast cancer data, small round blue-cell tumor data, lymphoma tumor data, and acute leukemia tumor data. The experimental results show that the proposed methods can effectively find some genes that are consistent with the existing biological knowledge, and the classification accuracies are very high. Some robustness and sensitivity properties have also been discussed. Finally, interaction of three Bayesian selection methods (including logistic) and four classifier methods (including logistic) has been considered.

Table 10. The top 20 important genes selected using AIC, BIC and MDL for breast cancer nature-based log ratio data ($\nu_i = 10/n$).

Gene No.	AIC		BIC		MDL	
	Frequency	Index No. (Clone ID)	Frequency	Index No. (Clone ID)	Frequency	Index No. (Clone ID)
1	0.1076	10 (26184)	0.1204	2222 (244227)	0.1167	2222 (244227)
2	0.1045	2300 (344352)	0.1066	2300 (344352)	0.1096	2300 (344352)
3	0.0986	408 (245198)	0.1032	408 (245198)	0.1036	10 (26184)
4	0.0983	2222 (244227)	0.1024	10 (26184)	0.0993	408 (245198)
5	0.0920	858 (783729)	0.0929	1059 (842894)	0.0980	1059 (842894)
6	0.0901	1059 (842894)	0.0895	858 (783729)	0.0915	858 (783729)
7	0.0876	560 (139540)	0.0820	560 (139540)	0.0813	560 (139540)
8	0.0838	336 (823940)	0.0785	2164 (143887)	0.0808	336 (823940)
9	0.0834	2164 (143887)	0.0761	336 (823940)	0.0760	2164 (143887)
10	0.0810	955 (950682)	0.0712	2226 (282980)	0.0759	2226 (282980)
11	0.0792	733 (134748)	0.0701	733 (134748)	0.0756	955 (950682)
12	0.0764	2226 (282980)	0.0697	955 (950682)	0.0683	733 (134748)
13	0.0705	742 (183200)	0.0657	742 (183200)	0.0655	742 (183200)
14	0.0693	2699 (44180)	0.0631	1443 (566887)	0.0589	1443 (566887)
15	0.0678	1443 (566887)	0.0539	2423 (26082)	0.0567	2699 (44180)
16	0.0676	2428 (26184)	0.0533	1999 (247818)	0.0563	1999 (247818)
17	0.0673	253 (28469)	0.0532	2428 (26184)	0.0512	2734 (46019)
18	0.0668	1999 (247818)	0.0519	2345 (141768)	0.0503	2423 (26082)
19	0.00652	2345 (141768)	0.0518	2699 (44180)	0.0501	2345 (141768)
20	0.00644	118 (47542)	0.0508	118 (47542)	0.0496	2428 (26184)

Table 11. The top 20 important genes selected using linear probit regression[13] and mutual-information[24] based gene selection method for breast cancer data.

Gene No.	Probit		MI	
	Frequency	Index No. (Clone ID)	Mutual information	Index No. (Clone ID)
1	0.0860	1008 (897781)	1.6165	556 (212198)
2	0.0840	336 (823940)	1.6018	2670 (42888)
3	0.0780	10 (26184)	1.4723	1008 (897781)
4	0.0750	1068 (840702)	1.3969	2893 (32790)
5	0.0710	496 (376516)	1.3890	1065 (843076)
6	0.0690	118 (47542)	1.3889	1999 (247818)
7	0.0660	3009 (366647)	1.3858	1345 (949932)
8	0.0660	585 (293104)	1.3837	1859 (307843)
9	0.0620	523 (28012)	1.3719	1443 (566887)
10	0.0610	556 (212198)	1.3527	2734 (46019)
11	0.0590	1999 (247818)	1.3518	3009 (366647)
12	0.0550	2423 (26082)	1.3231	1466 (767817)
13	0.0540	498 (667598)	1.3037	609 (246524)
14	0.0520	140 (30093)	1.3034	806 (46182)
15	0.0510	1277 (73531)	1.2987	272 (47681)
16	0.0500	955 (950682)	1.2915	2951 (291057)
17	0.0500	272 (47681)	1.2820	963 (897646)
18	0.0490	2734 (46019)	1.2730	2272 (309583)
19	0.0490	1859 (307843)	1.2670	2423 (26082)
20	0.0480	555 (548957)	1.2511	1179 (788721)

Table 12. The number of misclassification using logit, LProbit, NLProbit and KNN classifiers based on logit ($\nu_i = 15/n$), probit ($\nu_i = 15/n$) and mutual information-based gene selection methods with 5 and 10 genes for the breast cancer data, respectively.

Selection Method	Classifiers (5 genes)				Classifiers (10 genes)			
	Logit	LProbit	NLProbit	KNN	Logit	LProbit	NLProbit	KNN
Logit	0	0	0	0	0	0	0	0
Probit	2	1	0	1	2	0	0	1
MI	2	2	0	2	4	4	0	2

Acknowledgement

This research was supported by the National Human Genome Research Institute and the Translational Genomics Research Institute.

References

1. H. Akaike, A new look at statistical model indentification, *IEEE Trans. Automat. Control.* **19** (1974) 716–723.
2. A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown and L. M. Staudt, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature* **403** (2000) 503–511.
3. A. Antoniadis, S. Lambert-Lacroix and F. Leblanc, Effective dimension reduction methods for tumor classification using gene expression data, *Bioinformatics* **19** (2003) 563–570.
4. M.-H. Chen, J. G. Ibrahim and C. Yiannoutsos, Prior elicitation, variable selection, and Bayesian computation for logistic regression models, *Journal of the Royal Statistical Society, Series B* **61** (1999) 223–242.
5. S. Dudoit, J. Fridlyand and T. P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association* **97** (2002) 77–87.
6. W. H. Greene, *Econometric Analysis* (Prentice Hall, Saddle River, NJ, 1997).
7. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science* **286** (1999) 531–537.
8. I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene selection for cancer classification using support vector machines, *Machine Learning* **46** (2002) 389–422.
9. I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, M. Raffeld, Z. Yakhini, A. Ben-Dor, E. R. Dougherty, J. Kononen, L. Bubendorf, W. Fehrle, S. Pittaluga, S. Gruvberger, N. Loman, O. Johannsson, H. Olsson, B. Wilfond, G. Sauter, O.-P. Kallioniemi, A. Borg and J. Trent, Gene expression profiles in hereditary breast cancer, *The New England Journal of Medicine* **344** (2001) 539–548.
10. R. Jornsten and B. Yu, Simultaneous gene clustering and subset selection for classification via MDL, *Bioinformatics* **19** (2003) 1100–1109.

11. J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson and P. S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicince* **7** (2001) 673–679.
12. S. Kim, E. R. Dougherty, J. Barrea, Y. Chen, M. Bittner and J. M. Trent, Strong feature sets from small samples, *Computational Biology* **9** (2002) 127–146.
13. K. E. Lee, N. Sha, E. R. Dougherty M. Vannucci and B. K. Mallick, Gene selection: A Bayesian variable selection approach, *Bioinformatics* **19** (2003) 90–97.
14. L. Li, C. R. Weinberg, T. A. Darden and L. G. Pedersen, Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method, *Bioinformatics* **17** (2001) 1131–1142.
15. W. Li and Y. Yang, How many genes are needed for a discriminant microarray data analysis, in *Methods of Microarray Data Analysis*, eds. S. M. Lin and K. F. Johnson (Kluwer Academic, 2002), pp. 137–150.
16. D. V. Nguyen and D. M. Rocke, Tumor classification by partial least squares using microarray gene expression data, *Bioinformatics* **18** (2002) 39–50.
17. D. V. Nguyen and D. M. Rocke, Multi-class cancer classification via partial least squares with gene expression profiles, *Bioinformatics* **18** (2002) 1216–1226.
18. G. Qian and C. Field, Using MCMC for logistic regression model selection involving large number of candidate models, in *Proceeding of the 4th International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, November 27–December 1, Hong Kong (2000).
19. J. Rissanen, *Stochastic Complexity in Statistical Inquiry* (World Scientific Publishing Company, 1989).
20. G. Schwarz, Estimating the dimension of a model, *Ann. Statist.* **6** (1978) 461–464.
21. I. Tabus, J. Rissenan and J. Astola, Classification and feature gene selection using the normalized maximum likelihood for discrete regression, *Signal Processing* **83** (2003) 713–727.
22. X. Zhou, X. Wang and E. R. Dougherty, Binarization of microarray data based on a mixture model, *Molecular Cancer Therapeutics* (2003), in press.
23. X. Zhou, X. Wang and E. R. Dougherty, Missing value estimation based on linear and nonlinear regression with Bayesian gene selection, *Bioinformatics* (2003), in press.
24. X. Zhou, X. Wang and E. R. Dougherty, Nonlinear-probit gene classification using mutual-information and wavelet-based feature selection (2003), submitted.
25. E. Xing, M. Jordan and R. Karp, Feature selection for high dimensional genomic microarray data, in *Proc. 8th International Conferece on Machine Learning*, Williams College, Massachussets (2001).