BIOINFORMATICS

Vol. 00 no. 00 2003
Pages 1–9

Biomarker Discovery in Microarray Gene Expression Data with Gaussian Processes

Wei Chu [a], *Zoubin Ghahramani* [a], *Francesco Falciani* [b] *and David L. Wild* [c]

[a]*Gatsby Computational Neuroscience Unit, University College London, UK,* [b]*School of Biosciences, University of Birmingham, UK,* [c]*Keck Graduate Institute of Applied Life Sciences, Claremont, CA 91711, USA*

ABSTRACT

Motivation: In clinical practice, pathological phenotypes are often labelled with ordinal scales rather than binary, e.g. the Gleason grading system for tumor cell differentiation. However, in the literature of microarray analysis, these ordinal labels have been rarely treated in a principled way. This paper describes a gene selection algorithm based on Gaussian processes to discover consistent gene expression patterns associated with ordinal clinical phenotypes. The technique of automatic relevance determination is applied to represent the significance level of the genes in a Bayesian inference framework.

Results: The usefulness of the proposed algorithm for ordinal labels is demonstrated by the gene expression signature associated with the Gleason score for prostate cancer data. Our results demonstrate how multi-gene markers that may be initially developed with a diagnostic or prognostic application in mind are also useful as an investigative tool to reveal associations between specific molecular and cellular events and features of tumor physiology. Our algorithm can also be applied to microarray data with binary labels with results comparable to other methods in the literature.

Availability: The source code was written in ANSI C, which is accessible at www.gatsby.ucl.ac.uk/∼chuwei/code/gpgenes.tar.

Contact: wild@kgi.edu

1 INTRODUCTION

Microarray technologies now enable the simultaneous interrogation of the expression level of thousands of genes to obtain a quantitative assessment of their differential activity in a given tissue or cell. The development of these technologies has also motivated interest in their use in clinical trials and diagnosis. For instance, a key aim of many investigators is to identify genomic factors that are prognostic for survival or relapse-free survival and which predict those patients who respond to treatment. Typically, such experiments investigate on the order of dozens of samples from different patients. The samples are usually labelled with some information about the disease. Many studies have attempted to find subsets of genes that distinguish well between samples with different labels. A minimal subset of these relevant genes, often referred to as "biomarkers", may be useful in segregating patients in diagnosis, prognosis and for appropriate therapeutic selection in clinical management.

The increasing use of gene expression profiles in these types of study requires computational methods of high accuracy for solving feature selection and classification problems associated with these data. Although the cases of binary labels, e.g. healthy/diseased, have been extensively studied in the literature [Alon et al., 1999, Furey et al., 2000, Golub et al., 1999, Guyon et al., 2002, Li et al., 2002, Shevade and Keerthi, 2003], the observed or measured labels are often ordinal in routine clinical practice, such as the TNM system for staging prostate cancer and the Gleason grading system for tumor cell differentiation. These ordinal scales are discrete and finite, differing from continuous variables, and metric distances between the adjacent ordinal scales are not defined. In contrast to the labels of multiple classes, ordinal scales are rank-ordered, e.g. "low", "medium" and "high". The learning task of predicting ordinal variables is known as *ordinal regression*. Interestingly, the popular binary label is a special case of the ordinal variable with only two ranks. Singh et al. [2002] studied gene expression patterns that are correlated with the Gleason score and built an expression-based model to predict patients' clinical outcome. However, the ordinal nature of the Gleason score has not previously been treated in a principled way.

In this paper, we propose a feature selection algorithm based on Gaussian processes [Williams and Barber, 1998] to identify biomarkers for tasks with ordinal (or binary) labels. The important advantage of Gaussian process models is the explicit probabilistic framework that can efficiently take into account the uncertainty in microarray data. The *automatic relevance determination* (ARD) parameters [1] can be embedded into the

[1] The techniques of *automatic relevance determination* were originally proposed by MacKay [1994] and Neal [1996] in the context of Bayesian neural networks as a hierarchical prior over the weights.

covariance function, which represents the correlation between samples, to control the contribution from individual features. After Bayesian inference, the optimal values of the ARD parameters can be used as the indicator of the relevance level of a particular gene. A relatively large ARD parameter indicates that the associated gene is more correlated with the sample labels, while a gene weighted with a very small ARD parameter implies that this gene is irrelevant. Genes can then be sorted downwards from relevant to irrelevant according to the optimal values of these ARD parameters. A forward selection procedure can be further employed to determine the minimal set of relevant genes as biomarkers. We apply this ARD technique to publicly available microarray gene expression data sets. The usefulness of these biomarkers are validated by reference to the biological literature.

The paper is organized as follows. In Section 2, we describe the Gaussian processes model for ordinal regression and then present our algorithm in detail. The experimental results on three publicly accessible data sets are reported and discussed in Section 3. We conclude in Section 4.

2 METHODOLOGY

Consider a gene expression data set $\mathcal{D}$ composed of n samples from different patients. Each sample is represented by the expression level of the d genes, denoted as a column vector $x_i \in \mathcal{R}^d$, and labelled by an ordinal scale $y_i \in \mathcal{Y}$. These labels are denoted as consecutive integers $\mathcal{Y} = \{1, 2, \ldots, r\}$ that keep the known ordering information.

2.1 Bayesian framework

The main idea is to assume an unobservable latent function $f(x_i) \in \mathcal{R}$ associated with a sample x_i in a Gaussian process, and the label y_i dependent on the latent function $f(x_i)$ by modelling the ordinal scales as intervals on the real line [Chu and Ghahramani, 2004].

2.1.1 Prior Probability The values of the latent function $\{f(x_i)\}$ are assumed to be the realizations of random variables in a zero-mean Gaussian process. The covariance between the function values corresponding to the inputs x_i and x_j can be defined as

$$\mathrm{Cov}[f(x_i), f(x_j)] = \mathcal{K}(x_i, x_j) = \sum_{\ell=1}^{d} \kappa_\ell x_i^\ell x_j^\ell \qquad (1)$$

where $\kappa_\ell > 0$. x_i^ℓ denotes the ℓ-th gene expression level of the i-th sample and κ_ℓ is the ARD variable for the ℓ-th gene that controls the contribution of this gene in the modelling. The prior probability of these latent function values $\{f(x_i)\}$ is a multivariate Gaussian

$$\mathcal{P}(\boldsymbol{f}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{f}^T \Sigma^{-1} \boldsymbol{f}\right) \qquad (2)$$

where $\boldsymbol{f} = [f(x_1), f(x_2), \ldots, f(x_n)]^T$ and Σ is the $n \times n$ covariance matrix whose ij-th element is defined as in (1).

2.1.2 Ordinal Likelihood The likelihood $\mathcal{P}(\mathcal{D}|\boldsymbol{f})$ is the joint probability of observing the sample labels given the the latent function values. The likelihood can be evaluated as a product of the likelihood function on individual observations:

$$\mathcal{P}(\mathcal{D}|\boldsymbol{f}) = \textstyle\prod_{i=1}^{n} \mathcal{P}(y_i|f(x_i)).$$

A standard likelihood function for ordinal labels is obtained from the difference of two cumulative normals

$$\mathcal{P}(y_i|f(x_i)) = \Phi\left(z_1^i\right) - \Phi\left(z_2^i\right) \qquad (3)$$

where $z_1^i = \frac{b_{y_i} - f(x_i)}{\sigma}$, $z_2^i = \frac{b_{y_i - 1} - f(x_i)}{\sigma}$, and $\Phi(z) = \int_{-\infty}^{z} \mathcal{N}(\gamma; 0, 1) d\gamma$. The noise level $\sigma > 0$ is unknown and reflects the measurement noise in the microarray experiments. $b_0 = -\infty$ and $b_r = +\infty$ are defined as auxiliary variables, and we impose the inequality $b_1 < b_2 < \ldots < b_{r-1}$ on these thresholds. The role of the thresholds is to divide the real line into r contiguous intervals; these intervals map the real function value $f(x_i)$ into the discrete variable y_i while enforcing the ordinal constraints. As a special case with $r = 2$, the ordinal likelihood function (3) becomes the *probit* function for binary classification. [2]

2.1.3 Model Evidence The Bayesian framework described above is conditional on the model parameters including the ARD parameters κ_ℓ in the covariance function (1), the threshold parameters $\{b_1, b_2, \ldots, b_{r-1}\}$ and the noise level σ in the likelihood function (3). All these parameters can be collected into $\boldsymbol{\theta}$, which is the model parameter vector. The quantity $\mathcal{P}(\mathcal{D}) = \int \mathcal{P}(\mathcal{D}|\boldsymbol{f})\mathcal{P}(\boldsymbol{f}) d\boldsymbol{f}$, more exactly $\mathcal{P}(\mathcal{D}|\boldsymbol{\theta})$, is known as the *evidence* for $\boldsymbol{\theta}$, a yardstick for model selection. The optimal values of the model parameters $\boldsymbol{\theta}$ can be inferred by maximizing the evidence $\mathcal{P}(\mathcal{D}|\boldsymbol{\theta})$.[3]

A popular idea for computing the evidence is to approximate the posterior distribution $\mathcal{P}(\boldsymbol{f}|\mathcal{D}) \propto \mathcal{P}(\mathcal{D}|\boldsymbol{f})\mathcal{P}(\boldsymbol{f})$ as a Gaussian by applying the Laplace approximation at the maximum a posteriori (MAP) estimate of $\boldsymbol{f}$, and then the evidence can be calculated by an explicit formula. The MAP estimate on the latent functions is the mode point of the posterior distribution, i.e. $\boldsymbol{f}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{f}} \mathcal{P}(\boldsymbol{f}|\mathcal{D})$. This is a convex programming problem that guarantees a unique solution. The Laplace approximation refers to carrying out the Taylor expansion for $\mathcal{P}(\boldsymbol{f}|\mathcal{D})$ at the MAP point and retaining the terms up to the second order [MacKay, 1992]. The evidence can then be approximated as an explicit expression analytically:

$$\mathcal{P}(\mathcal{D}|\boldsymbol{\theta}) \approx \exp(-\mathcal{S}(\boldsymbol{f}_{\mathrm{MAP}}))|\mathbf{I} + \Sigma\Lambda_{\mathrm{MAP}}|^{-\frac{1}{2}} \qquad (4)$$

[2] For multi-label classification problems, the *softmax* function can be employed as the likelihood function for multinomial labels, as discussed by Williams and Barber [1998].

[3] Monte Carlo sampling methods can provide a good approximation to the posterior distribution of $\boldsymbol{\theta}$, but might be prohibitively expensive to use for high-dimensional problems.

where $\mathcal{S}(\boldsymbol{f}) = \frac{1}{2}\boldsymbol{f}^T\Sigma^{-1}\boldsymbol{f} - \sum_{i=1}^{n}\ln\mathcal{P}(y_i|f(x_i))$, $\mathbf{I}$ is an $n \times n$ identity matrix, and Λ_{MAP} is a diagonal matrix whose ii-th entry is $\frac{\partial^2 - \ln\mathcal{P}(y_i|f(x_i))}{\partial^2 f(x_i)}$ at the MAP estimate.

The gradients of the approximate evidence (4) with respect to the model parameters $\boldsymbol{\theta}$ can be derived analytically (refer to Chu and Ghahramani [2004] for detailed formulae). Gradient-based optimization methods can then be employed to search for the maximizer of the evidence $\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}}\mathcal{P}(\mathcal{D}|\boldsymbol{\theta})$. Since there might be several local maxima on the curve of $\mathcal{P}(\mathcal{D}|\boldsymbol{\theta})$, it is possible that the optimization problem may stick at local maxima in the determination of $\boldsymbol{\theta}$. We can avoid poor local maxima by maximizing (4) several times starting from several different initial states, and simply choose the one with the highest evidence as our preferred choice $\boldsymbol{\theta}^\star$.

2.2 Prediction

At the optimal model parameters $\boldsymbol{\theta}^\star$, let us take a test sample x_t for which the target y_t is unknown. The correlations between the test case x_t and the training samples $\{x_i\}$ are defined by the covariance function $\mathcal{K}(x_t, x_i)$ as in (1). The predictive distribution over ordinal labels y_t is

$$\mathcal{P}(y_t|x_t, \mathcal{D}, \boldsymbol{\theta}^\star) = \Phi\left(\frac{b_{y_t} - \mu_t}{\sqrt{\sigma^2 + \sigma_t^2}}\right) - \Phi\left(\frac{b_{y_t-1} - \mu_t}{\sqrt{\sigma^2 + \sigma_t^2}}\right). \tag{5}$$

where $\mu_t = \boldsymbol{k}^T\Sigma^{-1}\boldsymbol{f}_{\text{MAP}}$, $\sigma_t^2 = \mathcal{K}(x_t, x_t) - \boldsymbol{k}^T(\Sigma + \Lambda_{\text{MAP}}^{-1})^{-1}\boldsymbol{k}$, and $\boldsymbol{k} = [\mathcal{K}(x_t, x_1), \mathcal{K}(x_t, x_2), \ldots, \mathcal{K}(x_t, x_n)]^T$. The predictive label is decided as

$$\hat{y}_t = \arg\max_i \mathcal{P}(y_t = i|x_t, \mathcal{D}, \boldsymbol{\theta}^\star). \tag{6}$$

2.3 Forward Selection

The optimal values of κ_ℓ's can be determined by the maximizer of the evidence $\boldsymbol{\theta}^\star$, denoted as $\kappa_\ell^\star$'s, which indicates the relevance level of the genes to the labels. Based on these values $\kappa_\ell^\star$'s, we can sort the genes in descending order from relevant to irrelevant accordingly.

It is desirable to further select a minimal subset of the top-ranked genes as the biomarkers for modelling, denoted as $\mathcal{M}$, while keeping the accuracy of the resulting model and reducing the computational overhead. For this purpose, we need to define a quality criterion for the quality of a particular biomarker set. The leave-one-out (LOO) validation error is popularly used,[4] which is evaluated as

$$\text{LOO Error} = \sum_t \delta(\hat{y}_t \neq y_t) \tag{7}$$

[4] For the microarray datasets with dozens of samples, this might result in the same LOO error for multiple biomarker sets, which makes it difficult to discern the difference in performance. It is acceptable to employ other quality criteria, such as the predictive probability of misclassifications in LOO which is defined as $\sum_{t:\hat{y}_t \neq y_t} - \ln\mathcal{P}(y_t|x_t, \mathcal{D}, \boldsymbol{\theta}^\star)$ where $\mathcal{P}(y_t|x_t, \mathcal{D}, \boldsymbol{\theta}^\star)$ is computed as in (5) and $\hat{y}_t$ is defined as in (6).

Table 1. The outline of our algorithm for gene selection.

Initialize	generate k folds of the data set and $i = 1$
Loop	while $i \leq$k, leave the i-th fold out
	1. maximize evidence on the remaining k-1 folds optimization package returns the optimal $\boldsymbol{\theta}^\star$
	2. sort the genes by the optimal values of ARD parameters
	3. run forward selection to compute the LOO error (7)
	4. identify the minimal gene set $\mathcal{M}_i$
	5. $i = i + 1$
Ranking	rank the genes by the *number of hits* in the sets $\{\mathcal{M}_i\}_{i=1}^k$
Selection	run forward selection to compute quality criteria
	identify the minimal gene set $\mathcal{M}^\star$
Exit	return the set of selected genes $\mathcal{M}^\star$

where $\sum_t$ means the sum over all LOO validation cases, $\hat{y}_t$ is defined as in (6) and $\delta(s)$ is 1 if s is true, otherwise 0.

We can carry out LOO validation on a progressively larger biomarker set, adding one gene at a time as ordered by the gene ranking. Here a linear covariance function, defined as $\sum_{\ell=1}^{d} x_i^\ell x_j^\ell$ without ARD parameters, is employed in the Gaussian process modelling. The inclusion of a relevant gene should result in a decrease of the LOO error criterion (7). The gene set $\mathcal{M}$ that yields the minimal LOO error is identified as the set of biomarkers that contain the most informative genes for predicting target labels.

2.4 Algorithm

The optimal values of ARD parameters are estimated by maximizing the approximate evidence, which is also known as type-II maximum likelihood estimate. Qi et al. [2004] have shown that the evidence optimization can lead to overfitting by picking one from numerous linear classifiers that can correctly classify the limited training data. This potential difficulty becomes more serious on gene expression datasets with only dozens of samples. To address this problem, we propose a resampling procedure as the outer loop of our algorithm. The outline of our algorithm is given in Table 1. We found this algorithm to be robust both to overfitting and local minima problems.

Given a gene expression dataset, we randomly generated k folds after preprocessing. One fold was left out in turn and evidence optimization was carried out using the samples in the remaining k-1 folds. We maximized the evidence (4) several times starting from different initial states, and simply chose the one with the highest evidence as the optimal $\boldsymbol{\theta}^\star$. Based on the optimal values of the ARD parameters, the genes were sorted in descending order from relevant to irrelevant accordingly. In forward selection, we added one top-ranked gene each time into the gene subset $\mathcal{M}_i$ and then carried out LOO cross validation using the linear covariance function $\sum_{\ell=1}^{d} x_i^\ell x_j^\ell$ on the training samples in the remaining k-1 folds. The minimal subset that yielded the minimal LOO error was identified as $\mathcal{M}_i$. This procedure was repeated k times, and k subsets $\{\mathcal{M}_i\}$ were obtained. The number of times each gene was selected

in the k subsets $\{\mathcal{M}_i\}$ was used as the final criterion for gene ranking, which we refer to as *number of hits*. Genes with same *number of hits* are further ranked by the average ARD values. We carried out forward selection again based on the final gene rank to identify the minimal subset of relevant genes $\mathcal{M}^\star$.

3 RESULTS AND DISCUSSION

Three publicly accessible gene expression datasets, related to colon, leukaemia and prostate cancer, were analyzed using our algorithm. In all cases, the expression levels of each sample were first normalized to zero-mean and unit variance and then the expression levels of each gene were again normalized to zero-mean and unit variance over all the samples. We tackled two kinds of tasks with our algorithm, i.e., normal versus tumor (binary classification) and Gleason score prediction (ordinal regression).

3.1 Normal versus Tumor

Many popular gene ranking methods employ the t-statistic as a criterion to measure the variance of the expression levels in different classes for each gene [Alon et al., 1999, Furey et al., 2000]. Variants of the t-statistic, such as the measure of correlation proposed by Golub et al. [1999] and Fisher's discriminant criterion adapted by Pavlidis et al. [2001], have also been extensively applied. The t-statistic-like methods make the assumption that the data are described by a Gaussian distribution. However, according to Deng et al. [2004] and others, the normality condition often cannot be met in real gene expression datasets with very limited samples. Nonparametric tests, e.g. the Wilcoxon rank sum test, are superior to the t-test in this case.

As a preprocessing step, we used the Wilcoxon rank sum test on the normalized expression data to remove the most uninformative genes. The significance level was fixed at p=0.01, and the p-values were calculated using all the samples.[5] We then generated 10 folds of the whole data set for the resampling step in Table 1. The detailed results on these three datasets are reported in the following.

The colon cancer dataset, originally analyzed by Alon et al. [1999], contains expression levels of $d = 2000$ genes from 40 tumor and 22 normal colon tissues.[6] There are 373 genes significantly differentially expressed in the rank sum test at the significance level of p=0.01.

The leukaemia dataset, originally studied by Golub et al. [1999],[7] contains expression values of $d = 7129$ genes from 47 samples of acute myeloid leukaemia (AML) and 25 samples of acute lymphoblastic leukaemia (ALL). There are 1169 genes significantly differentially expressed at the significance level of p=0.01.

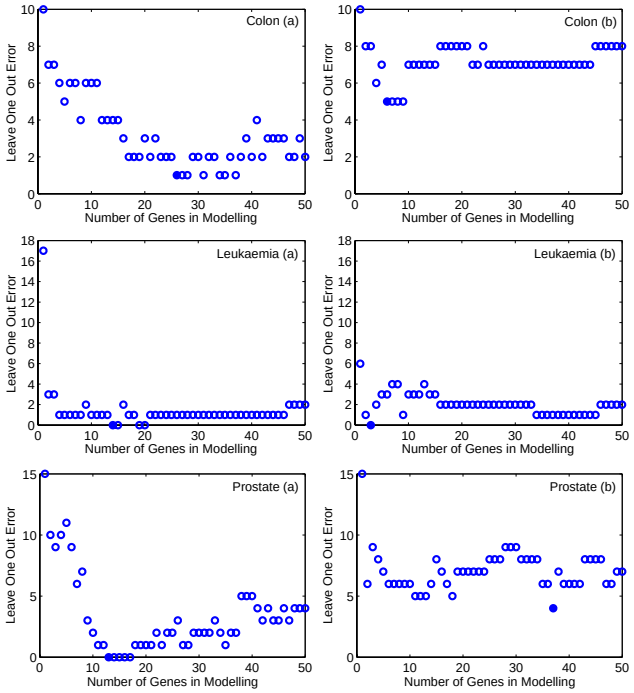

Fig. 1. The leave-one-out error using the top-ranked genes of the three datasets. The top-ranked 50 genes are progressively used in the modelling and the corresponding LOO error (7) are shown as circles. In the left-hand figures the genes are ranked by the *number of hits* of our algorithm, while in the right-hand figures the genes are ranked by their p-values of the Wilcoxon rank sum test. The filled circles indicate the set of selected genes with minimal LOO error.

Singh et al. [2002] carried out microarray expression analysis on 12600 genes to identify genes that are correlated with the distinction of prostate tumor from normal.[8] Fifty-two samples of prostate tumor and fifty samples of normal cells were investigated. There are 2717 genes significantly differentially expressed at the significance level p=0.01.

The left part of Figure 1 presents the results of the LOO error for the 50 top-ranked genes sorted by the *number of hits* of our algorithm, along with that for the genes ranked by the p-values of the rank sum test in the right part. A lower LOO error can be achieved using the gene rank of our algorithm, although this may involve using more genes than when using the p-value rankings on the colon and leukaemia data.

The selected genes are listed in Table 2 - 4 with more descriptions. In Table 2, we found that all the 8 genes selected by Shevade and Keerthi [2003] and 6 of 7 genes selected by Guyon et al. [2002] are also in our list. In Table 3, 8 of 9 genes selected by Shevade and Keerthi [2003] are also selected by our algorithm. The six genes in bold face were also identified by Golub et al. [1999] as being part of their 50 gene signature

[5] Since we are using the ranking by p-value as a preprocessing step, it was unnecessary for us to apply any correction for multiple testing or false discovery rate.

[6] Available at http://microarray.princeton.edu/oncology/affydata/index.html.

[7] The dataset is available at http://www.genome.wi.mit.edu/MPR.

[8] The dataset is available at http://www.genome.wi.mit.edu/MPR/prostate.

Table 2. The selected 26 genes in Colon cancer data. "Index" denotes the serial number of the selected gene in the original data. "Hits" is the *number of hits* criterion used in our algorithm. "Rank" denotes the rank in the p-values of Wilcoxon rank sum test. "SLR (8)" denotes the rank in the 8 genes selected by the sparse logistic regression algorithm of [Shevade and Keerthi, 2003]. "RFE (7)" denotes the rank in the 7 genes selected by recursive feature elimination using the support vector machines of [Guyon et al., 2002].

Index	GAN	Description	Hits	Rank	SLR (8)	RFE (7)
377	Z50753	h.sapiens mrna for gcap-ii/uroguanylin precursor	10	1	1	-
1772	H08393	collagen alpha 2(xi) chain (homo sapiens)	9	10	2	7
576	D14812	human mrna for orf, complete cds	9	232	7	-
792	R88740	atp synthase coupling factor 6, mitochondrial precursor	9	183	5	3
1924	H64807	placental folate transporter (homo sapiens)	9	98	4	1
493	R87126	myosin heavy chain, nonmuscle (gallus gallus)	8	2	3	-
732	R67343	immediate-early regulatory protein ie-n	8	120	-	-
1843	H06524	gelsolin precursor, plasma (human)	7	9	6	-
1473	R54097	translational initiation factor 2 beta subunit (human)	7	90	-	-
1231	H49870	mad protein (homo sapiens)	7	64	-	-
14	H20709	myosin light chain alkali, smooth-muscle isoform (human)	7	32	-	-
1346	T62947	60s ribosomal protein l24 (arabidopsis thaliana)	6	191	8	2
1360	H09719	tubulin alpha-6 chain (mus musculus)	6	291	-	-
1549	H11084	vascular endothelial growth factor (cavia porcellus)	6	117	-	-
1210	R55310	mitochondrial processing peptidase	6	212	-	-
663	Z17227	h.sapiens mRNA for transmenbrane receptor protein	6	277	-	-
1668	M82919	human gamma amino butyric acid (gabaa) receptor beta-3 subunit mRNA	6	189	-	-
1555	L38929	Homo sapiens protein tyrosine phosphatase delta mRNA, complete cds	6	288	-	-
1579	M31516	human decay-accelerating factor mRNA	5	370	-	-
1920	J04102	human erythroblastosis virus oncogene homolog 2 (ets-2) mRNA	5	218	-	-
1570	H81558	procyclic form specific polypeptide b1-alpha precursor	5	289	-	4
211	T47424	insulin receptor substrate-1 (homo sapiens)	5	267	-	-
1400	M59040	human cell adhesion molecule (cd44) mrna, complete cds	4	343	-	6
1221	R62549	putative serine/threonine-protein kinase b0464.5 in chromosome iii	4	244	-	-
1935	X62048	h.sapiens wee1 hu gene.	3	202	-	-
1916	T41204	p14780 92 kd type v collagenase precursor	3	357	-	-

Table 3. The selected 14 genes in Leukaemia data. "Index" denotes the serial No. of the selected gene. "Hits" is the *number of hits* used in our algorithm. "Rank" denotes the p-value rank in the Wilcoxon rank sum test. "SLR (9)" denoted the rank in the 9 genes selected by Shevade and Keerthi [2003]. The boldfaced genes were selected in the 50 gene signature of Golub et al. [1999].

Index	GAN	Description	Hits	Rank	SLR(9)
4951	Y07604	NDP kinase	10	93	2
4847	**X95735**	**Zyxin**	10	4	3
1779	M19507	MPO Myeloperoxidase	10	29	1
1834	**M23197**	**CD33 antigen**	9	1	6
6184	M26708	PTMA Prothymosin alpha	9	133	5
4196	**X17042**	**PRG1 Proteoglycan 1**	9	32	-
2288	**M84526**	**DF (adipsin)**	8	15	-
1829	M22960	PPGB (galactosialidosis)	8	28	-
6283	M65214	TCF3 Transcription factor 3	7	46	-
1882	**M27891**	**CST3 Cystatin C**	7	3	8
3252	U46499	glutathione s-transferase	6	6	-
3847	**U82759**	**HoxA9**	6	74	4
6169	M13690	C1NH	6	212	9
6041	L09209	APLP2	6	5	-

which distinguished AML from ALL. The gene selections are further visualized in Figure 2 by presenting the covariance matrices in colour. The covariance matrices turn out to be clearly blocked using the selected genes. The samples in same class are generally positively correlated, whereas the samples in different classes are negatively correlated.

Table 4. The selected 13 genes in Prostate cancer data. "Index" denotes the serial No. of the selected gene in the original data. "Hits" denotes the *number of hits* of our algorithm. "Rank" denotes the p-value rank in Wilcoxon rank sum test.

Index	Description	Hits	Rank
6185	X07732:hepatoma mRNA for serine protease **hepsin**	10	1
10234	AF055376:transcription factor C-MAF mRNA	10	163
11871	U21689:Human glutathione S-transferase-P1c gene	10	97
5890	AJ001625:Homo sapiens mRNA for Pex3 protein	10	38
5045	AL080150: cDNA DKFZp434D174	10	85
7623	X51345:Human jun-B mRNA for JUN-B protein	10	386
9172	AI207842:ao89h09.x1 Homo sapiens cDNA, 3 end	10	6
6390	AI093155:qa97g04.x1 Homo sapiens cDNA, 3 end	9	917
7539	X04297:human mrna for Na,K-atpase alpha-subunit	9	287
12495	M98539:Human prostaglandin D2 synthase gene	9	129
4438	AI275081:ql65b10.x1 Homo sapiens cDNA, 3 end	8	512
11942	D00017:humlic homo sapiens mrna for lipocortin II	8	45
7139	AF025887:Homo sapiens GSTA4 mRNA	8	1062

To estimate the predictive accuracy of our algorithm, we report in Table 5 the test error rates of a 10-fold cross validation experiment. One fold was left out for test in turn, and a Gaussian process model was trained on the remaining 9 folds using a gene subset selected by the rank sum test or the proposed algorithm separately. Note that the gene selection was carried out by using the samples in the 9 training folds only,

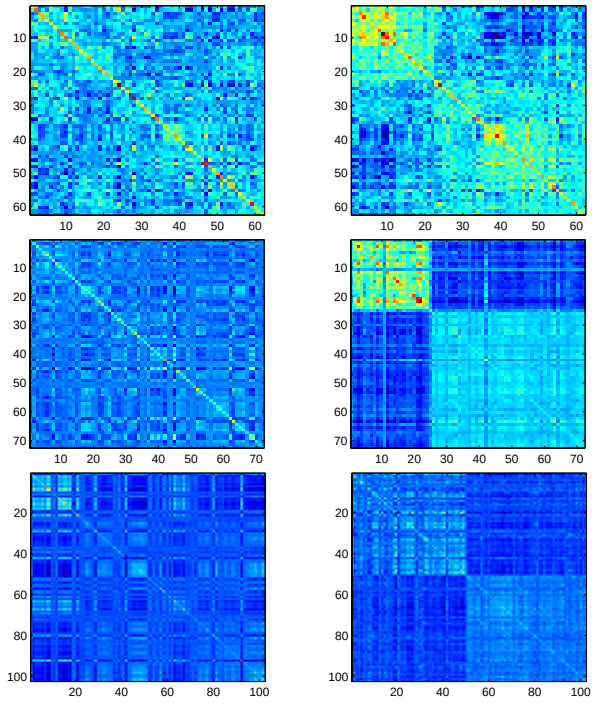

Fig. 2. The covariance matrices for the binary classification tasks. The covariance matrix is the $n \times n$ covariance matrix whose ij-th elements are defined by the linear covariance function $\sum_{\ell=1}^{d} x_i^\ell x_j^\ell$. In the left-hand figures the covariance matrices were evaluated over all the original genes, whereas in the right-hand figures the covariance matrices were evaluated over the genes selected by our algorithm. The samples have been grouped by their labels. The pairs in rows from top to bottom are for the Colon, Leukaemia and Prostate datasets accordingly. Red indicates that the sample pair is positively correlated, while blue indicates negatively correlated.

Table 5. Test error rates in the 10-fold cross validation experiments. The Wilcoxon rank sum test and the proposed algorithm were applied to select the gene subset for modelling separately using the training samples in 9 folds only, and then tested on the unused fold. "All Genes" denotes that all the genes were used in modelling, "Rank Sum Test" denotes that the subset of genes with p-values lower than 0.01 in the Wilcoxon rank sum test were used, and "Biomarkers" denotes that the gene subset selected by our algorithm was used. Test error rates averaged over the 10 folds are reported along with the standard deviation. The integers in the brackets are the total test error number over the 10 folds.

Dataset	All Genes	Rank Sum Test	Biomarkers
Colon	22.38±19.12%(14)	16.19±17.60%(10)	16.19±13.65%(10)
Leukaemia	17.44±8.02%(13)	7.08±9.63(5)	6.67±9.25%(5)
Prostate	14.73±12.67%(15)	12.82±10.66(13)	8.81±9.74%(9)

and then tested on the unused fold. We observed that the validation results using hundreds of genes selected by rank sum test are always better than that using all the original genes. The improvement is especially significant on the leukaemia dataset. Our algorithm can further reduce the number of selected genes to less than 50, and yields competitive performance on the colon and leukaemia datasets and much better results on the prostate.

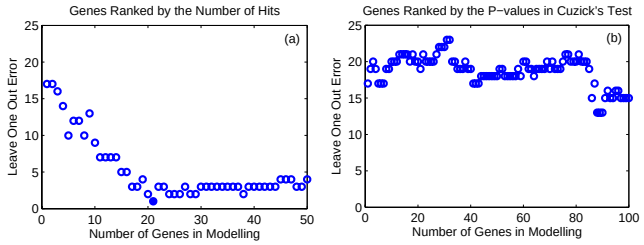

Fig. 3. The leave-one-out error for the task of predicting the Gleason score, using the top-ranked gene sets of the prostate data set. The top-ranked genes are progressively used in the modelling and the corresponding LOO error numbers are presented in the graphs (a) and (b) respectively.

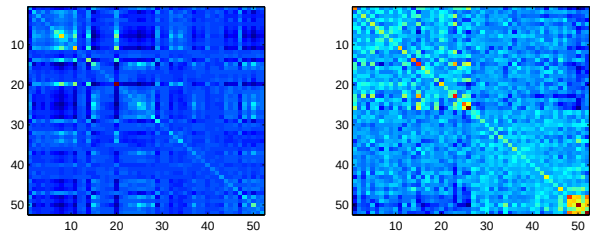

Fig. 4. The covariance matrices for the task of predicting Gleason score. As in Figure 2, the left graph presents the covariance matrix evaluated over all the original genes, while the right graph presents the matrix evaluated over the selected genes by our proposed algorithm. The samples have been grouped by their ordinal scales.

3.2 Gleason Score Prediction

The Gleason score is based exclusively on the architectural pattern of the glands of the prostate tumor. It evaluates how effectively the cells of any particular cancer are able to structure themselves into glands resembling those of the normal prostate. The ability of a tumor to mimic normal gland architecture is called its differentiation. The Gleason grading from very well differentiated (grade 1) to very poorly differentiated (grade 5) is usually done for the most part by viewing a low magnification microscopic image of the tumor. There are two types of Gleason scores, type I and type II, both of which have 5 scales. Hereafter, Gleason score refers to the sum of the grades of the two types.

Singh et al. [2002] investigated fifty-two samples of prostate tumor to identify a subset of the 12600 genes correlated with pathological features. For each sample, the Gleason score given by the pathologist ranges from 6 to 10. Singh et al. [2002] treated the Gleason scores as continuous variables in their analysis. We argue that the Gleason score are ordinal variables in nature rather than continuous variables, as the grades are ordered as ranks and the metric distances between the adjacent grades are not defined. Predicting the Gleason score from the gene expression data is thus a typical ordinal regression problem. In our experiments, as only 6 samples had a score greater than 7, we merged them as the top level, leading to three levels $\{=6, =7, \geq 8\}$ with 26, 20 and 6 samples respectively. We generated 6 folds in the resampling

procedure, and present the quality criteria for the top 50 genes ranked by the *number of hits* in Figure 3(a). The minimal LOO error number was observed when the top 21 genes were used. The selected 21 genes are listed in Table 6 with detailed descriptions. We further visualized the selected genes in Figure 4 by presenting the covariance matrices in colour. We observed three clearly blocked regions for the three ordinal scales in the covariance matrices using the selected genes. Moreover, the samples of the level 6 are strongly negatively correlated to the samples of level ≥ 8.

Cuzick's test is a Wilcoxon-like test for trend across ordered groups [Lehmann, 1998]. The informative genes can be selected based on the p-values of the Cuzick test. The LOO error numbers using the 100 top-ranked genes are presented in Figure 3(b). When more than 80 genes are used in modelling, the LOO error becomes smaller than that obtained using the top-ranked gene only. A much lower LOO error was obtained by our algorithm using the top 21 ranked genes. We also tried the Kruskal-Wallis rank sum test, which is designed for the case of multiple categories [Lehmann, 1998]. Since this test is insensitive to the ordering information among the ordinal scales, the LOO errors are always greater than that using the first top-ranked gene. This observation also implies that multi-classification methods should not be generally applied to tackle ordinal regression problems.

3.3 Discussion

The models we have developed to discriminate between normal and tumor tissues (prostate and colon cancer datasets) and between AML and ALL are very promising and reflect to some degree what is known of the biology of these systems. A representative case is hepsin in Table 4 (a gene selected in the signature discriminating between normal and tumor prostate samples). Hepsin is a cell surface serine protease that is known to be markedly upregulated in human prostate cancer. Overexpression of hepsin in a mouse model of non-metastasizing prostate cancer has no impact on cell proliferation, but causes disorganization of the basement membrane and promotes primary prostate cancer progression and metastasis to liver, lung and bone [Klezovitch et al., 2004].

Of particular interest are the models linking the degree of differentiation of prostate tumor (Gleason score) to the molecular state of tumor cells. In their original attempt Singh et al. [2002] have identified genes whose expression was correlated to this pathological variable. There are two major limitations in their approach. Firstly, genes are selected individually rather than in combination. The second limitation is that the Gleason score is not a continuous variable but a categorical one, as mentioned earlier.

Our approach significantly improves on the previous study by providing a statistical model representing the Gleason score as an ordered categorical variable. The molecular signature we have developed is robust and has good explanatory power. Although the signature we have identified does not include any of the genes originally selected by Singh et al. [2002] we have observed some degree of functional overlap. Both signatures, in fact, include genes involved in insulin response (IGF-I in our model, i.e. #3 in Table 6, and Insulin-like growth factor binding protein 3 in the model developed by Singh et al. [2002]) and contain members of the complement component pathway (Complement component 2 in the original analysis and Complement component 7 in our model, i.e. #21 in Table 6). Interestingly, the large majority of the genes in the models we have developed to explain the degree of differentiation of the tumor are known to be associated to tumor physiology or are related to molecular functions that are highly informative of the molecular events underling the pathology. Table 6 shows a functional classification of the selected genes. The most striking feature of our model is that seven genes are either tumor suppressor genes or oncogenes and therefore are known to be directly involved in the neoplastic process. Our signature contains five genes with tumor-suppressor activity. Of these, three have a demonstrated function in prostate cancer. The expression of the lysyl oxidase-like protein (LLP, #13 in Table 6) gene has been reported to be progressively lost in primary prostate cancer and associated metastatic lesions [Ren et al., 1998] and is inactivated by methylation and loss of heterozygosity in human gastric cancers [Kaneda et al., 2004]. These observations are strongly supportive for a role of LLP as a tumor suppressor gene in solid tumors. The expression of IGF1 is also decreased in human prostate cancer. A clear tumor-suppressive activity in prostate cancer has been demonstrated through an apoptotic mechanism [Mutaguchi et al., 2003]. Another gene selected in our model with a demonstrated tumor suppressive activity in prostate cancer is the inducible cAMP early repressor (CREM/ICER, #11). This gene is an important mediator of cAMP antiproliferative activity that specifically affects the tumorigenicity of prostate cancer cell without affecting their growth [Memin et al., 2002]. Phosphatidylethanolamine N-methyltransferase (PEMT, #9) is an enzyme in liver that catalyzes the stepwise methylation of phosphatidylethanolamine to phosphatidylcholine. PEMT protein decreased in pre-neoplastic nodules and virtually disappeared in hepatocellular carcinoma induced by aflatoxin B_1. Transfection experiments demonstrated that the loss of PEMT function may contribute to malignant transformation of hepatocytes [Tessitore et al., 2000]. This enzyme is expressed at similar levels in liver and prostate cells (estimated by looking at the frequency of ESTs in the Unigene database) and therefore it is reasonable to hypothesize a similar role may be shared in these different organs. Last of the tumor suppressor genes included in the model is the RbA p48 gene (#6). The protein encoded by this gene has been demonstrated to mediate the retinoblastoma protein tumor suppressor activity [Qian et al., 1993]. RbA would, in fact, be a component of the histone deacetylase complex that associates with the retinoblastoma protein [Nicolas et al., 2000]. Two genes encoding proteins with oncogenic activity have also been selected in

Table 6. The selected 21 genes in Prostate cancer data for predicting the Gleason score. "#" denotes the serial number in the list. "Hits" denotes the *number of hits* criterion used in our algorithm. "Cuzick" denotes the rank in the p-values of the Cuzick test for trend.

#	Index	Description	Functional role	Hits	Cuzick
1	583	AJ010232:Homo sapiens mRNA for **RET finger protein-like 3**	Oncogene related	6	1
2	7714	AA630312:ac08f05.s1 Homo sapiens cDNA	Not annotated	5	440
3	9264	X57025:Human **IGF-I mRNA** for insulin-like growth factor I	Tumor supressor-like	5	534
4	6118	AW043690:wy80b07.x1 Homo sapiens cDNA	Not annotated	4	10
5	11213	D84361 Human mRNA for p52 and p64 isoforms of N-Shc	Secretion and signalling	4	82
6	7049	X74262:H.sapiens **RbAp48 mRNA** encoding retinoblastoma binding protein	Tumor supressor-like	4	1329
7	8424	AF022375:Homo sapiens vascular endothelial growth factor mRNA	Vascularization	4	5586
8	10617	AW007029:ws49c09.x1 Homo sapiens cDNA	Not annotated	3	367
9	6897	AB029821:Homo sapiens mRNA for **phosphatidylethanolamine N-methyltransferase**	Tumor supressor-like	3	507
10	8484	U81561:Human protein tyrosine phosphatase receptor pi (PTPRP) mRNA	Signaling	3	41
11	4681	S68271:cyclic AMP-responsive element modulator (**CREM**)	Tumor supressor-like	3	208
12	4325	AF104942:Homo sapiens ABC transporter MOAT-C (**MOAT-C**) mRNA	Transport drug resistance	3	4445
13	5837	U24389:Human **lysyl oxidase-like** protein gene	Tumor supressor-like	3	5
14	7076	AF017307:Homo sapiens Ets-related transcription factor (**ERT**) mRNA	Oncogene related	3	85
15	9878	U90028:Homo sapiens bicaudal-D (BICD) mRNA	Migration and motility	3	562
16	10787	HSU83661 Homo sapiens multidrug resistance protein 5 (**MRP5**) mRNA	Transport drug resistance	3	1103
17	11233	HUMRPTK Homo sapiens receptor protein-tyrosine kinase (HEK11) mRNA	Migration and motility	3	36
18	6749	AB028978:Homo sapiens mRNA for KIAA1055 protein	Not annotated	3	400
19	10764	AF024710 Homo sapiens vascular endothelial growth factor (**VEGF**) mRNA	Vascularization	3	1934
20	5809	J02931:Human placental **tissue factor** (two forms) mRNA	Vascularization	3	376
21	8878	J03507:Human complement protein **component C7 mRNA**	Complement	2	86

the model. These are ERT (#14) and RET (#1). The proteins of the ETS family are transcription factors involved in signal transduction, cell cycle progression, and differentiation. It has been demonstrated that cell neoplastic transformation is associated with a dramatic increase in ETS transcriptional activity [de Nigris et al., 2001]. The RET proto-oncogene encodes a protein that belongs to the tyrosine kinase growth factor receptor family. The RET proto-oncogene is expressed in human prostate cancer xenografts and prostate cancer cell lines [Dawson et al., 1998].

Angiogenesis is another important process in the development of the tumor and it is represented in our model by two genes. These are VEGF (#19) and one of its main regulators, the gene encoding for Tissue factor (#20). VEGF is the only mitogen that specifically acts on endothelial cells and its function is key to the development of tumor angiogenesis in vivo [Affara and Robertson, 2004]. Tissue factor (TF), when produced by tumor cells, has been implicated in the regulation of new blood vessels formation through its ability to concurrently induce the expression of angiogenic molecules such as vascular endothelial cell growth factor (VEGF), while inhibiting the expression of anti-angiogenic molecules. The expression of TF has been directly linked to vascularization in prostate cancer [Abdulkadir et al., 2000]. Another molecular function represented in our model and with great relevance in tumor physiology is the ability to develop drug resistance. MRP5/MOAT-C (represented twice in the model we have developed, i.e. #16 and #12 in Table 6) is a drug resistant gene that has been implicated in the transport of cyclic nucleotides from cultured cells or isolated tissues [Wielinga et al., 2003].

Our model representing the degree of tumor differentiation is particularly interesting since most of the genes are directly linked to the molecular events underlying tumor progression (tumor suppressor genes, oncogenes and vascularization markers) or are related to cellular function relevant to cancer physiology (motility and secretion). The function of genes represented in our models suggests that the ability of tumor cells to aggregate into glandular-structures may be correlated to the regulation of proliferation and survival. Of interest is also the link between vascularization and the degree of tumor differentiation. This link is strongly supported by our model (both VEGF and one of its activators have been selected). Ultimately the ability to develop resistance to anti-cancer drugs could also be linked to the degree of differentiation of the tumor. Our results demonstrate how multi-gene markers that may be initially developed with a diagnostic or prognostic application in mind are also useful as an investigative tool to reveal associations between specific molecular and cellular events and features of tumor physiology.

4 CONCLUSIONS

We have presented a feature selection algorithm based on Gaussian processes for biomarker discovery associated with ordinal (including binary) clinical phenotypes. This algorithm is clearly superior to the simple ranking method using the rank sum test. Our results on the three microarray datesets are very promising and supported by existing biological knowledge. Moreover, our algorithm can be directly applied for biomarker discovery in large scale proteomics and metabolomics datasets and this is a focus of our future work.

ACKNOWLEDGEMENTS

We would like to thank the Institute for Pure and Applied Mathematics (IPAM) at UCLA, where part of this work was carried out. WC thanks S. Sathiya Keerthi for many discussions. WC, ZG and DLW acknowledge support from NIH Grant Number 1P01GM63208.

REFERENCES

Abdulkadir, S. A., G. F. Carvalhal, Z. Kaleem, W. Kisiel, P. A. Humphrey, W. J. Catalona, and J. Milbrandt. Tissue factor expression and angiogenesis in human prostate carcinoma. *Hum Pathol*, 31 (4):443–447, 2000.

Affara, N. I. and F. M. Robertson. Vascular endothelial growth factor as a survival factor in tumor-associated angiogenesis. *In Vivo*, 18 (5):525–542, 2004.

Alon, U., N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine. 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, 1999.

Chu, W. and Z. Ghahramani. Gaussian processes for ordinal regression. Technical report, Gatsby Unit, University College London, 2004. www.gatsby.ucl.ac.uk/∼chuwei/paper/gpor.pdf.

Dawson, D. M., E. G. Lawrence, G. T. MacLennan, S. B. Amini, H. J. Kung, D. Robinson, M. I. Resnick, E. D. Kursh, T. P. Pretlow, and T. G. Pretlow. Altered expression of RET proto-oncogene product in prostatic intraepithelial neoplasia and prostate cancer. *J Natl Cancer Inst*, 90(7):519–523, 1998.

de Nigris, F., T. Mega, N. Berger, M. V. Barone, M. Santoro, G. Viglietto, P. Verde, and A. Fusco. Induction of ETS-1 and ETS-2 transcription factors is required for thyroid cell transformation. *Cancer Res.*, 61(5):2267–2275, 2001.

Deng, L., J. Pei, J. Ma, and D. L. Lee. A rank sum test method for informative gene discovery. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*, pages 410–419, Washington, USA, 2004.

Furey, T., N. Christianini, N. Duffy, D. Bednarski, M. Schummer, and D. Hauessler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

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

Guyon, I., J. Weston, and V. Barnhill, S. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

Kaneda, A., K. Wakazono, T. Tsukamoto, N. Watanabe, Y. Yagi, M. Tatematsu, M. Kaminishi, T. Sugimura, and T. Ushijima. Lysyl oxidase is a tumor suppressor gene inactivated by methylation and loss of heterozygosity in human gastric cancers. *Cancer Res.*, 64 (18):6410–6415, 2004.

Klezovitch, O., J. Chevillet, J. Mirosevich, R. Roberts, R. Matusik, and V. Vasioukhin. Hepsin promotes prostate cancer progression and metastasis. *Cancer Cell*, 6(2):185–195, 2004.

Lehmann, E. L. *Nonparametrics: Statistical Methods Based on Ranks*. Prentice Hall, rev. 1st edition, 1998.

Li, Y., C. Campbell, and M. Tipping. Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics*, 18(10):1332–1339, 2002.

MacKay, D. J. C. A practical Bayesian framework for back propagation networks. *Neural Computation*, 4(3):448–472, 1992.

MacKay, D. J. C. Bayesian methods for backpropagation networks. *Models of Neural Networks III*, pages 211–254, 1994.

Memin, E., G. Yehia, R. Razavi, and C. A. Molina. ICER reverses tumorigenesis of rat prostate tumor cells without affecting cell growth. *Prostate*, 53(3):225–231, 2002.

Mutaguchi, K., H. Yasumoto, K. Mita, A. Matsubara, H. Shiina, M. Igawa, R. Dahiya, and T. Usui. Restoration of insulin-like growth factor binding protein-related protein 1 has a tumor- suppressive activity through induction of apoptosis in human prostate cancer. *Cancer Res.*, 63(22):7717–7723, 2003.

Neal, R. M. *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics. Springer, 1996.

Nicolas, E., V. Morales, L. Magnaghi-Jaulin, A. Harel-Bellan, H. Richard-Foy, and D. Trouche. RbAp48 belongs to the histone deacetylase complex that associates with the retinoblastoma protein. *J Biol Chem*, 275(13):9797–9804, 2000.

Pavlidis, P., J. Weston, J. Cai, and W. N. Grundy. Gene functional classification from heterogeneous data. In *Proceedings of the Fifth International Conference on Computational Molecular Biology*, pages 242–248, 2001.

Qi, Y., T. P. Minka, R. W. Picard, and Z. Ghahramani. Predictive automatic relevance determination by expectation propagation. In *Proceedings of the Twenty-first International Conference on Machine Learning*, pages 671–678, 2004.

Qian, Y.-W., Y.-C. J. Wang, R. E. Jr.. Hollingsworth, D. Jones, N. Ling, and E. Y.-H. P. Lee. A retinoblastoma-binding protein related to a negative regulator of Ras in yeast. *Nature*, 364(6438): 648–652, 1993.

Ren, C., G. Yang, T. L. Timme, T. M. Wheeler, and T. C. Thompson. Reduced lysyl oxidase messenger RNA levels in experimental and human prostate cancer. *Cancer Res.*, 58(6):1285–1290, 1998.

Shevade, S. K. and S. S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19 (17):2246–2253, 2003.

Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209, 2002.

Tessitore, L., E. Sesca, and D. E. Vance. Inactivation of phosphatidylethanolamine N- methyltransferase-2 in aflatoxin-induced liver cancer and partial reversion of the neoplastic phenotype by PEMT transfection of hepatoma cells. *Int J Cancer*, 86(3):362–367, 2000.

Wielinga, P. R., I. van der Heijden, G. Reid, J. H. Beijnen, J. Wijnholds, and P. Borst. Characterization of the MRP4- and MRP5-mediated transport of cyclic nucleotides from intact cells. *J Biol Chem*, 278(20):17664–17671, 2003.

Williams, C. K. I. and D. Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.