Mathematical Medicine and Biology (2007) **24**, 413−426
doi:10.1093/imammb/dqn001
Advance Access publication on February 22, 2008

The ant colony algorithm for feature selection in high-dimension gene expression data for disease classification

K. R. ROBBINS†, W. ZHANG AND J. K. BERTRAND
Department of Animal and Dairy Science, The University of Georgia, Athens, GA 30602, USA

AND

R. REKAYA

Department of Animal and Dairy Science, Department of Statistics, and Institute of Bioinformatics, The University of Georgia, Athens, GA 30602, USA

[Received on 23 February 2007; revised on 16 January 2008; accepted on 17 January 2008]

The use of gene expression data to diagnose complex diseases represents an exciting area of medicine; however, such data sets are often noisy, requiring the selection of feature subsets to obtain maximum classification accuracy. Due to the high dimensions of many expression data sets, filter-based methods are commonly used, but often yield inconsistent results. Optimization algorithms can outperform filter methods, but often require preselection of features to achieve good results. To address the problems of many commonly used feature selection methods, the ant colony algorithm (ACA) is proposed for use on data sets with large numbers of features. The ACA is an optimization algorithm capable of incorporating prior information, allowing it to search the sample space more efficiently than other optimization methods. When applied to several high-dimensional data sets, the ACA was able to identify small subsets of highly predictive and biologically relevant genes without the need for extensive preselection of features. Using the selected genes to train a latent variable model yielded substantial increases in prediction accuracy when compared to several rank-based methods and results obtained in previous studies. The superiority of the ACA algorithm was validated through simulation.

Keywords: ant colony algorithm; disease diagnostics; feature selection; gene expression.

1. Introduction

The idea of using gene expression data for diagnosis and personalized treatment regimes represents a promising area of medicine and, as such, has been the focus of much research (Bagirov *et al.*, 2003; Golub *et al.*, 1999; Ramaswamy *et al.*, 2001). Many algorithms have been developed to classify disease types based on the expression of selected genes, and significant gains have been made in the accuracy of disease classification (Antonov *et al.*, 2004; Bagirov *et al.*, 2003). In addition to the development of classification algorithms, many studies have shown that improved performance can be achieved when using a selected subset of features, as opposed to using all available data (Peng *et al.*, 2003; Shen *et al.*, 2006; Subramani *et al.*, 2006). Increases in accuracy achieved through the selection of predictive features can complement and enhance the performance of classification algorithms, as well as improve the understanding of disease classes by identifying a small set of biologically relevant features (Golub *et al.*, 1999).

†Email: krobbin1@uga.edu

Ideally, one would like to select an optimal subset of features that would yield maximum predictive power for a given classification algorithm. In the case of high-dimensional data sets, this can be very computationally demanding; consequently, many statistical and rank-based methods are often used. While these methods are simple to implement and capable of quickly generating lists of selected features, their performance can vary greatly across different data types (Jeffery *et al.*, 2006); furthermore, many rank-based methods only give an indirect measurement of a single feature's predictive ability and do not take into account the contribution of a feature when grouped in a classifier (Shen *et al.*, 2006). As a result, these methods may select groups of highly correlated genes. This high collinearity could lower prediction accuracy due to larger uncertainties in parameter estimates and redundant information in the classifier. As such, there is a need for a feature selection method that can evaluate the contribution of a feature relative to all others in a classifier, is robust enough to yield consistent optimal performances across data types and can do so in a computationally feasible manner.

The idea of selecting a subset of features capable of best classifying a group of samples can be, and has been, viewed as an optimization problem. The genetic algorithm (GA), simulated annealing (SA) and other optimization and machine learning algorithms have been applied to the problem of feature selection (Lin *et al.*, 2006; Ooi & Tan, 2003; Peng *et al.*, 2003; Albrecht *et al.*, 2003). Though these methods are powerful, when dealing with thousands of features across multiple classes, the computational cost of these methods can be prohibitive. Previous results obtained with these methods, when dealing with large numbers of features, utilized filters to reduce the dimension of the data sets prior to implementation (Lin *et al.*, 2006; Peng *et al.*, 2003) or have produced relatively low prediction accuracies (Hong & Cho, 2006). The ant colony algorithm (ACA) is a machine learning technique that simulates the positive feedback system used by ant colonies to find the shortest route to a food source through the use of pheromone trails (Dorigio & Gambardella, 1997). Ants that choose a shorter path will transverse the distance at a faster rate, depositing more pheromone in the process. As the shorter path accumulates more pheromone, ants will begin to preferentially choose to follow that path, creating a positive feedback system. The communication of the ants through a common memory has a synergistic effect that, when coupled with more efficient searching of the sample space through the use of prior information, results in optimal solutions being reached in far fewer iterations than required for GA or SA (Dorigio & Gambardella, 1997). The algorithm also lends itself to parallelization, with ants being run on multiple processors, which can further reduce computation time, making its use more feasible with high-dimension data sets.

For this study, the ACA was implemented using the high-dimensional multi-class cancer gene expression (GCM) data set (Ramaswamy *et al.*, 2001), a colon cancer data set (Alon *et al.*, 1999) and a simulated data set with very limited pre-filtering and compared to several other rank-based feature selection methods, as well as previously published results to determine its efficacy as a feature selection method.

2. Methods

2.1 *Classification*

2.1.1 *Latent variable model.* A Bayesian regression model was used to predict the tumour type in the form of a probability $p_{ic}(y_{ic} = 1)$, with $y_{ic} = 1$ indicating that sample i is from tumour class c. The regression on the vector of binary responses $\mathbf{y_c}$ was done using a latent variable model (LVM), with l_{ic} being an unobserved, continuous latent variable relating to binary response y_{ic} such that

$$y_{ic} = \begin{cases} 1, & \text{if } l_{ic} \geqslant 0, \\ 0, & \text{if } l_{ic} < 0. \end{cases}$$

The liability l_{ic} was modelled using a linear regression model as

$$l_{ic} = \mathbf{X}_{ic}\boldsymbol{\beta}_c + e_{ic} \quad E(l_{ic}) = \mathbf{X}_{ic}\boldsymbol{\beta}_c \quad e_{ic} \sim N(0, 1),$$

where $\mathbf{X}_{ic}$ corresponds to row i of the design matrix $\mathbf{X}_c$ for tumour class c.

The link function of the expectation of the liability $\mathbf{X}_{ic}\boldsymbol{\beta}_c$ with the binary response y_{ic} was constructed via a probit model (West, 2003), yielding the following equations:

$$p_{ic}(y_{ic} = 1) = \boldsymbol{\Phi}(\mathbf{X_{ic}}\boldsymbol{\beta_c}) \quad \text{and} \quad p_{ic}(y_{ic} = 0) = 1 - \boldsymbol{\Phi}(\mathbf{X_{ic}}\boldsymbol{\beta_c}),$$

where $\boldsymbol{\Phi}$ is the standard normal distribution function.

For data sets containing multiple classes, subject i was classified as having tumour class c if $p_{ic}(y_{ic} = 1)$ was the maximum of the vector $\mathbf{p_i}$, containing all $p_{ic}(y_{ic} = 1)$, $c = 1, \ldots, nc$, where nc is the number of tumour classes in the data set. For binary data sets, subject i was classified as having tumour class c if $p_{ic}(y_{ic} = 1) > 0.5$.

For instances in which the number of selected features was greater than the number of subjects in the training data set, a singular value decomposition was applied to the data and a shrinkage estimator, known as a g-prior, was used to prevent overfitting (West, 2003). When the number of genes selected was less than the number of subjects in the training data set, a spectral decomposition was performed and principal components corresponding to eigenvalues close to zero were removed to avoid computational issues.

2.2 *Gene selection*

Filter- and wrapper-based methods were used to select features to form classifiers for each tumour class. Filter methods selected genes based on ranks determined by the sorted absolute values of fold changes (FC), t-statistics (T) and penalized t-statistics (PT) calculated for each gene for each tumour class. The wrapper method coupled the ACA with LVM (ACA/LVM) such that groups of genes were selected using the ACA and evaluated for performance using LVM.

2.2.1 *Fold change.*

The FC was computed as

$$\text{fc}_{mc} = |M_c - M_r|,$$

where M_c is the mean of the log base 2 of the gene expression of the tumour class of interest c and M_r is the mean of the log base 2 of the gene expression of the remaining tumour classes.

2.2.2 *t-Statistic.*

The T was calculated as

$$t_{mc} = \frac{|M_c - M_r|}{\text{Sp}\sqrt{1/n_c + 1/n_r}},$$

where M_c is the mean of the log base 2 of the gene expression of the tumour class of interest c, M_r is the mean of the log base 2 of the gene expression of the remaining tumour classes, Sp is the square root of the pooled variance, $n_c$is the number of subjects in the tumour class of interest c and $n_r$is the number of subjects in the remaining tumour types.

2.2.3 *Penalized t-statistic.* The PT was calculated as

$$\mathrm{pt}_{mc} = \frac{|M_c - M_r|}{a + \mathrm{Sp}\sqrt{1/n_c + 1/n_r}},$$

where M_c is the mean of the log base 2 of the gene expression of the tumour class of interest c, M_r is the mean of the log base 2 of the gene expression of the remaining tumour classes, a is the 90th percentile of the distribution of the pooled standard deviations of all m genes, Sp is the square root of the pooled variance, n_c is the number of subjects in the tumour class of interest and n_r is the number of subjects in the remaining tumour types.

2.2.4 *Ant colony optimization.* Artificial ants work as parallel units that communicate through a probability density function (PDF) that is updated by weights or 'pheromone levels', in this case determined by the performance of the selected features in classifying samples (Dorigio & Gambardella, 1997; Ressom *et al.*, 2006), where the probability of sampling feature m at time t is defined as

$$P_{mc}(t) = \frac{(\tau_{mc}(t))^{\alpha}\ \eta_{mc}^{\beta}}{\sum_{m=1}^{nf} (\tau_{mc}(t))^{\alpha}\ \eta_{mc}^{\beta}}, \tag{2.1}$$

where $\tau_{mc}(t)$ is the amount of pheromone for feature m (out of a total of nf features) of tumour class c at time t, η_{mc} is some form of prior information on the expected performance of feature m of tumour class c and α and β are parameters determining the weight given to pheromone deposited by ants and *a priori* information on the features, respectively.

For this study, the prior information (η_{mc}) was determined as

$$\eta_{mc} = \frac{\frac{f_{mc} - \min(\mathbf{f_c})}{\max(\mathbf{f_c}) - \min(\mathbf{f_c})} + \frac{t_{mc} - \min(\mathbf{t_c})}{\max(\mathbf{t_c}) - \min(\mathbf{t_c})} + \frac{\mathrm{pt}_{mc} - \min(\mathbf{pt_c})}{\max(\mathbf{pt_c}) - \min(\mathbf{pt_c})}}{3},$$

where $\mathbf{f_c}$ is a vector of all FC values for tumour class c, $\mathbf{t_c}$ is a vector of all T values for tumour class c and $\mathbf{pt_c}$ is a vector of all PT values for tumour class c. Values of α and β were set heuristically. After an extensive sensitivity analysis, it became clear that large values of β tended to give high weight to the prior information in detriment of the data. In contrary, very small values for β often resulted in slow convergence. Consequently, α and β were set to 1 and 0.3, respectively. These two values were chosen given their limited impact on the weight of the prior information and their improvement of the convergence of the procedure.

The ACA was initialized with all features having an equal baseline level of pheromone used to compute $P_m(0)$ for all features. Using the PDF as defined in (2.1), each of j artificial ants will select a subset S_k of n features from the sample space S containing all features. The pheromone level of each feature m in S_k is then updated according to the performance of S_k as

$$\tau_m(t+1) = (1-\rho)\tau_m(t) + \Delta\tau_m(t), \tag{2.2}$$

where ρ is a constant between 0 and 1 that represents the rate at which the pheromone trail evaporates and $\Delta\tau_m(t)$ is the change in pheromone level for feature m based on the performance of S_k and is set to zero if feature $m \notin S_k$. This process is repeated for all S_k.

The procedure can be summarized in the following steps:

(1) Each ant selects a predetermined number of genes.

(2) Training data are randomly split into two subsets for training (TDS) and validation (VDS) containing $3/4$ and $1/4$ of the data, respectively (none of the original validation data are used at any point in the ACA).
(3) Using the spectral decomposition of TDS, principal components are computed to alleviate effects of collinearity and selected for TDS and VDS by removing components with corresponding eigenvalues close to zero.
(4) Using TDS, an LVM is trained for each tumour class, and $p_{ic}(y_{ic} = 1)$ is predicted for every tumour class c for each sample i in VDS.
(5) The accuracy for each tumour class c is calculated as

$$\text{acc}_c = \frac{\sum_{i=1}^{nc} \boldsymbol{\Phi}(\mathbf{P_{ic}}\boldsymbol{\beta_c})/nc + \sum_{i=1}^{nr}[1 - \boldsymbol{\Phi}(\mathbf{P_{ic}}\boldsymbol{\beta_c})]/nr}{2}, \tag{2.3}$$

where $\mathbf{P_{ic}}$ contains principal component values for sample i for tumour class c, $\boldsymbol{\beta_c}$ is a vector of coefficients estimated using TDS, nc is the number of samples in VDS having tumour class c and nr is the remaining number of samples in VDS.
(6) The change in pheromone for each tumour class is calculated as

$$\Delta\tau_{mc}(t) = \text{acc}_c^{(1-\text{acc}_c)},$$

where acc_c is the accuracy for tumour type c as calculated using (2.3). This equation was chosen based on performance using real and simulated data.

Following the update of pheromone levels according to (2.2), the PDF is updated according to (2.1) and the process is repeated until some convergence criteria are met. As the PDF is updated, the selected features that perform better will be sampled at higher likelihoods by subsequent artificial ants which, in turn, deposit more 'pheromone', thus leading to a positive feedback system similar to the method of communication observed in real ant colonies. Upon convergence, the optimal subset of features is selected based on the level of pheromone trail deposited on each feature.

2.3 *GCM data set*

The data set contained 198 samples collected from 14 tumour types: BR (breast adenocarcinoma), PR (prostate adenocarcinoma), LU (lung adenocarcinoma), CO (colorectal adenocarcinoma), LY (lymphoma), BL (bladder transitional cell carcinoma), ML (melanoma), UT (uterine adenocarcinoma), LE (leukaemia), RE (renal cell carcinoma), PA (pancreatic adenocarcinoma), OV (ovarian adenocarcinoma), ME (pleural mesothelioma) and CNS (central nervous system). The data set was processed according to Ramaswamy *et al.* (2001) and contained the intensity values of 16063 probes generated using Affymetrix high-density oligonucleotide microarrays and calculated using Affymetrix GENECHIP software (http://www.broad.mit.edu/cgi-bin/cancer/publications/ pub_paper.cgi?mode=view&paper_id=61). Additionally, intensity values were thresholded to a minimum value of 20 and a maximum value of 16000. A log base 2 transformation was then applied to the data set. Genes with the highest expression values being less than two times the smallest were removed, leaving 14525 probes for analysis. For cross-validation, the data set was split into a training data set containing 144 subjects and a validation data set containing 54 samples.

2.4 *Colon cancer data set*

The data set contained 62 samples collected from normal colon tissue and colon cancer tumours (COA). The data set was processed according to Alon *et al.* (1999) and contained microarray gene expression for 2000 probes (http://microarray.princeton.edu/oncology/affydata/index.html). For cross-validation, the data set was randomly split into training data containing 56 samples and validation data containing 6 samples.

2.5 *Simulated data set*

A real data-based simulation was conducted using three of the 14 tumour types in the GCM data set. The tumour types were selected based on the relatively large number of tissue samples in each tumour type, respectively. After selection of tissue samples, expression levels for each of the samples were randomly shuffled across statuses and 50 genes were randomly selected to be up-regulated in one of the three tumour types. Gene expression values were then simulated as

$$y_{ij} = \exp_{hj} + r_{ij},$$

where y_{ij} is the simulated gene expression value of tissue sample i at gene j, $\exp_{hj}$ is the original gene expression for some random tissue sample h at gene j and r_{ij} is a normal random deviate for tissue sample i at gene j, simulated as

$$r_{ij} \sim N(0, v_j),$$

where v_j is equal to one-tenth the gene expression variance at gene j.

For genes selected to be up-regulated, a constant was added to the gene expression values of a randomly selected tumour type such that

$$E(T) = 3,$$

where $E(T)$ is the expectation of the T for genes selected to be up-regulated.

3. Results

The GCM data set has been a benchmark to compare the performance of classification and feature selection algorithms. Table 1 shows the best prediction accuracies obtained by methods used in this study and several previous studies genetic algorithm with silhouette statistics (Lin *et al.*, 2006), GA/maximum likelihood classification method (Ooi & Tan, 2003), maximal marginal linear programming (Antonov *et al.*, 2004), GA/support vector machines (Liu *et al.*, 2005) and SVM/recursive feature elimination (Ramaswamy *et al.*, 2001)) using independent test, performed on the same training and validation data sets originally formed by Ramaswamy *et al.* (2001) (GCM split) and leave one out cross-validation. The proposed ACA/LVM yielded substantial increases in accuracies over all other methods, with a 6.5% increase in accuracy over the next best results obtained using the GCM split (Antonov *et al.*, 2004). Furthermore, the ACA/LVM achieved increases of 13.9, 44.1 and 16.6% in accuracy over the FC/LVM, T/LVM and PT/LVM methods of feature selection, respectively.

The confusion matrices for the predictions obtained by the ACA/LVM, FC/LVM, PT/LVM and T/LVM using the GCM split are found in Tables 2–5. These tables show that the ACA/LVM performs as good as or better than the rank-based methods for every tumour type. Additionally, the ACA/LVM

TABLE 1 *Accuracy (%) of tumour class predictions using ACA and several previously published methods*

	GCM data set		
	GCM split†	Replicated splits	LOOCV‡
ACA/LVM(14525§)	90.7	84.8	—
FC/LVM(14525)	79.6	74.8	—
T/LVM(14525)	63.0	—	—
PT/LVM(14525)	77.8	74.4	—
AVG‖/LVM(14525)	79.6	74.8	—
GASS(1000)	81.5	—	81.3
GA/MLHD(1000)	76	—	79.8
MAMA	85.2	—	—
GA/SVM(1000)	—	—	81
SVM/RFE(16063)	60–77.8	—	—

GASS, genetic algorithm with silhouette statistics; MLHD, maximum likelihood classification method; MAMA, maximal marginal linear programming; SVM, support vector machines; RFE, recursive feature elimination.
†Split used by Ramaswamy *et al.* (2001).
‡LOOCV, leave one out cross-validation.
§Number of genes selected prior to the implementation of feature selection algorithm.
‖Weighted average of scaled fold change (FC), *t*-test (T) and penalized *t*-test (PT) values.

TABLE 2 *Confusion matrix for predictions obtained for the GCM data set using genes selected by the ACA*

	Predicted														
True	BR	PR	LU	CO	LY	BL	ML	UT	LE	RE	PA	OV	ME	CNS	
BR	2									1		1			4
PR	1	5													6
LU			4												4
CO				4											4
LY					6										6
BL		1				2									3
ML							2								2
UT								2							2
LE									6						6
RE										3					3
PA				1							2				3
OV												4			4
ME													3		3
CNS														4	4
	1	6	4	6	6	7	2	2	6	3	1	2	4	4	49/54

correctly predicted 50% of the BR samples, a tumour class that has traditionally yielded very poor results (Bagirov *et al.*, 2003; Ramaswamy *et al.*, 2001).

To further evaluate performance, each of the feature selection algorithms was tested using four additional random splits of the GCM data set, as well as several replications using the colon cancer data set of

TABLE 3 *Confusion matrix for best predictions obtained for the GCM data set using genes selected by the fold change (*50 *genes)*

	Predicted														
True	BR	PR	LU	CO	LY	BL	ML	UT	LE	RE	PA	OV	ME	CNS	
BR	0					3		1							4
PR	1	5													6
LU			3							1					4
CO				4											4
LY					6										6
BL		1				2									3
ML							2								2
UT								2							2
LE									6						6
RE										2	1				3
PA				1		1					1				3
OV						1						3	1		4
ME													3		3
CNS														4	4
	1	6	4	6	6	7	2	2	6	3	1	2	4	4	43/54

TABLE 4 *Confusion matrix for best predictions obtained for the GCM data set using genes selected by the penalized t-test (*10 *genes)*

	Predicted														
True	BR	PR	LU	CO	LY	BL	ML	UT	LE	RE	PA	OV	ME	CNS	
BR	0					3				1					4
PR	1	5													6
LU			4												4
CO				4											4
LY					6										6
BL		1				2									3
ML							2								2
UT								2							2
LE									6						6
RE										2	1				3
PA				2		1					0				3
OV						1						2	1		4
ME													3		3
CNS														4	4
	1	6	4	6	6	7	2	2	6	3	1	2	4	4	42/54

Alon *et al.* (1999) and a simulated data set. The best classification accuracies obtained for each algorithm are found in Table 6. Though the variance of prediction accuracy was high across replicates for all methods, the ACA/LVM algorithm yielded the best or equalled the best prediction accuracies for all replicates in each data set. When looking at the three filter methods, it can be seen that the best method varied depending on the replication and data set. These findings are in agreement with Jeffery *et al.* (2006).

TABLE 5 *Confusion matrix for predictions obtained for the GCM data set using genes selected by the t-test (*10 genes*)*

	Predicted														
True	BR	PR	LU	CO	LY	BL	ML	UT	LE	RE	PA	OV	ME	CNS	
BR	0		1			2			1						4
PR		5	1												6
LU			3							1					4
CO				4											4
LY					6										6
BL		1				2									3
ML							1				1				2
UT			1					1							2
LE									6						6
RE						1		2		0					3
PA				1		1					1				3
OV			1			2						0	1		4
ME	1												2		3
CNS	1													3	4
	1	6	4	6	6	7	2	2	6	3	1	2	4	4	34/54

TABLE 6 *Classification accuracies using several feature selection methods*

	Replication					
	1	2	3	4	5	Overall
	GCM (Ramaswamy *et al.*, 2001)					
ACA/LVM	90.7	83.3	79.6	81.5	88.9	84.8
FC/LVM	79.6	77.8	68.5	72.2	75.9	74.8
PT/LVM	77.8	77.8	66.7	68.5	81.5	74.4
AVG†/LVM	79.6	70.4	70.4	70.4	83.3	74.7
	Colon cancer (Alon *et al.*, 1999)					
ACA/LVM	83.3	100	83.3	100	100	93.3
FC/LVM	83.3	100	83.3	100	83.3	90
PT/LVM	83.3	100	83.3	100	83.3	90
T/LVM	83.3	100	83.3	100	83.3	90
	Simulated					
ACA/LVM	81.3	75	81.3	75	87.5	80
FC/LVM	68.8	56.3	75	62.5	68.8	63.8
PT/LVM	68.8	68.8	75	68.8	81.3	72.5
T/LVM	81.3	75	75	62.5	81.3	75

†Weighted average of scaled fold change (FC), *t*-statistic (T) and penalized *t*-test values (PT).

Due to a lack of any good criterion for determining an objective cut-off value for the rank-based methods used in this study, several values were used and evaluated. Table 7 shows the number of genes needed for each tumour type to achieve the best results, averaged across all replicates. It can be seen that, for 13 of the 18 tumour classes, the ACA/LVM selects fewer genes than the rank-based methods. It should be noted that for COA, the ACA/LVM required fewer genes to achieve the 90% accuracy

TABLE 7 *Number of genes selected for each tumour type using ACA and other feature selection methods*

	ACA	FC	PT	AVG†	T
GCM (Ramaswamy *et al.*, 2001)					
BR	3.4	18	14	18	—
PR	4.8	18	14	18	—
LU	2	18	14	18	—
CO	7.8	18	14	18	—
LY	6.6	18	14	18	—
BL	19.6	18	14	18	—
ML	4.6	18	14	18	—
UT	7.6	18	14	18	—
LE	3.2	18	14	18	—
RE	16	18	14	18	—
PA	14.6	18	14	18	—
OV	17.2	18	14	18	—
ME	5	18	14	18	—
CNS	5.6	18	14	18	—
Colon cancer (Alon *et al.*, 1999)					
COA	33.3	50	38	—	16
Simulated					
SC1	4.8	29	52	—	36
SC2	5	29	52	—	36
SC3	5	29	52	—	36

†Weighted average of scaled fold change (FC), *t*-statistic (T) and penalized *t*-test (PT) values.

reported for the rank-based methods. Unlike many rank-based methods, breaks in the pheromone levels provide a more objective way of selecting the number of genes per tumour type when using the ACA. Plots illustrating the selection of features for four tumour classes are found in Fig. 1.

To examine the degree of collinearity present in the top genes, as selected by ACA/LVM and the rank-based methods, the top 30 features selected for BR and CNS were clustered using *k*-means (Hartigan & Wong, 1979) and then correlated. The *k*-means algorithm was implemented using the R function *k*-means (R Development Core Team, 2006). The correlations between selected features can be seen in the form of heat matrices found in Fig. 2 where red (yellow) colour indicates low (high) correlations. The BR and CNS tumour classes were selected because they have very weak and very strong classifiers, respectively. When looking at collinearity in the features selected for CNS, there appear to be no substantial differences between methods; however, when looking at BR, features selected by ACA/LVM show far less collinearity than the rank-based methods. In fact, unlike the rank-based methods in which substantially more collinearity is observed with BR than CNS, the ACA/LVM shows very similar heat signatures for both groups of features.

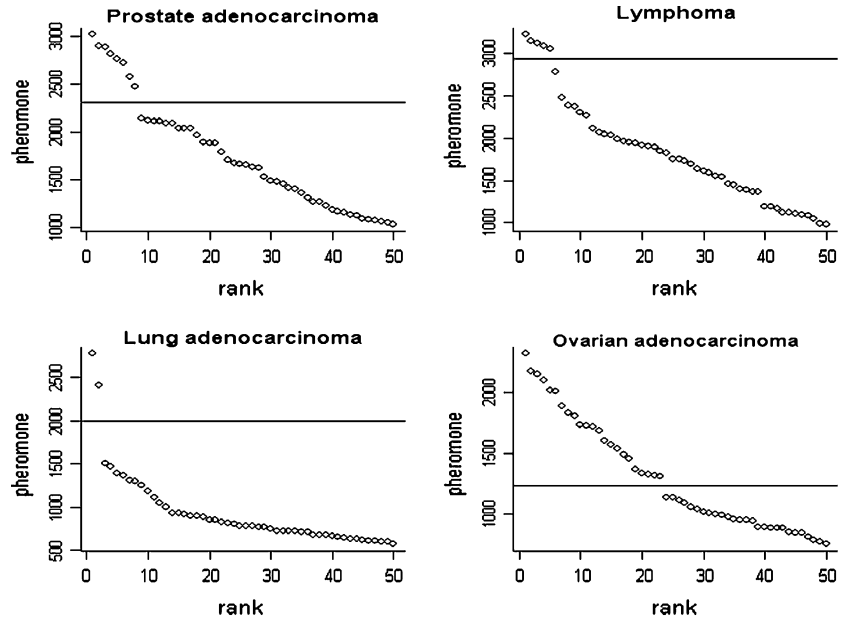

FIG. 1. Plots illustrating the selection of features (genes) using pheromone levels for four tumour types. Genes on top of the break-up point in the pheromone function (horizontal line) were selected.

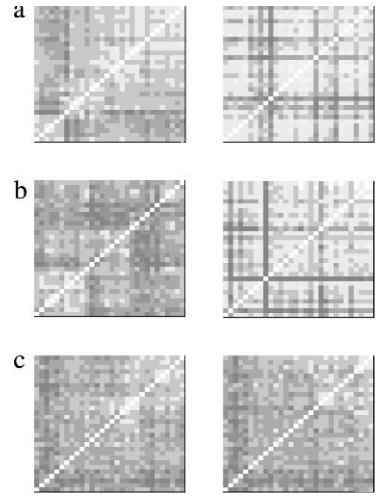

FIG. 2. Heat matrices between the top 30 genes selected for CNS (column 1) and BR (column 2) tumours based on (a) fold change, (b) penalized *t*-test and (c) ACA (dark, low correlation; light, high correlation).

4. Discussion

The performance of the ACA/LVM model was superior to not only the filter-based methods used in this study but also several reported results using the GCM data set. The ACA/LVM consistently yielded higher accuracies than the filter-based methods, for which ranks varied across replications and data sets.

The breaks in pheromone levels observed with the most predictive genes also provided more objective selection criteria for identifying top features, unlike the filter methods used in this study, in which truncation points were somewhat arbitrary. The objective selection criteria and robustness of the ACA, within the confines of the three data sets used in this study, make it a superior method for clinical applications, as it could enable a single procedure to be effectively applied to varied applications. The use of filter-based methods in such scenarios would require different combinations of truncation points and scoring methods for each data set, a highly impractical endeavour.

The superiority of the ACA/LVM when compared to models using GA indicates the ACA's utility, as compared to other optimization methods, when working with high-dimension data sets. The ACA's ability to incorporate prior information in the optimization process provides several advantages over other optimization algorithms when dealing with large numbers of features. The inclusion of prior information in the pheromone function focuses the selection process on genes that should yield better results without the need for an explicit truncation of the data, which was needed to achieve good results with the GA (Hong & Cho, 2006; Lin *et al.*, 2006; Liu *et al.*, 2005; Ooi & Tan, 2003; Peng *et al.*, 2003). Truncation of large numbers of genes could *a priori* eliminate genes from consideration that, though they may not have high predictive ability alone, could contribute the predictive power of an ensemble of genes. Additionally, depending on the method of truncation, the reduced gene list could be highly redundant (Lin *et al.*, 2006; Shen *et al.*, 2006), further reducing the informativeness of preselected genes. Conversely, when removing a small number of features in a large data set, the truncated data set may be too large for efficient convergence of the algorithm (Lin *et al.*, 2006). Additionally, the inclusion of prior information allows the ACA to be coupled with many other types of feature selection methods, making the ACA a versatile feature selection tool.

The reduction in the collinearity of genes as selected by ACA, particularly in tumour types yielding poor performance with filter methods, could be a source of the ACA's superior performance. Due to the reduction in the redundancy of selected features, fewer genes were needed for accurate classification in many of the tumour types. Combined with the fact that the ACA evaluates features in groups rather than individually, this should enable the ACA to identify clusters of genes with unique expression patterns, each contributing to the overall power of a classifier. These clusters of features, in addition to improving classification accuracy, could elucidate some of the biological mechanisms underlying the tumours of interest (Golub *et al.*, 1999). To this end, the ACA identified several small subsets of genes capable of obtaining high accuracies in cross-validation for many of the 14 tumour types contained in the GCM data set. Furthermore, using simulated data, 68.7% of the genes selected by the ACA were truly differentially expressed genes as opposed to only 41.1% of the genes identified by the highest performing rank-based method.

For LU tumours, the ACA identified two genes capable of classifying LU tumour samples with high accuracy in each of the five replicates. The selected genes, SP-B and SP-A, both encode pulmonary surfactant proteins which are necessary for lung function. Another tumour class, with which the ACA was able to select a small number of highly predictive genes, was CNS. As with the LU tumour type, the genes selected by the ACA were very consistent from replication to replication. The gene encoding for APCL protein had the highest pheromone levels in all five replicates and was the only gene required to achieve high prediction accuracy in Replicate 5. APCL protein is a homologue of APC, a known tumour suppressor that interacts with microtubules during mitosis (Akiyama & Kawasaki, 2006). The gene encoding MAP1B, a protein found to be important in synaptic function of cortical neurons, was also identified as being highly predictive of CNS tumour types. Several other genes selected by the ACA, found in 'supplemental materials', were identified in a previous study (Antonov *et al.*, 2004).

In contrast to the LU and CNS tumour types, BR samples were consistently predicted with low accuracies. These findings are in agreement with previous results (Bagirov *et al.*, 2003; Ramaswamy *et al.*, 2001). Unlike the gene list obtained for LU and CNS tumour types, the gene lists for BR tumours were highly variable, suggesting potentially high heterogeneity in these tumour samples. Despite dissimilarities between the genes selected across replications, the ACA did identify SEPT9 as being highly predictive in four of the five replicates. The protein encoded by this gene has been shown to be involved in mitosis of mammary epithelial cells (Nagata *et al.*, 2003) and has been associated with both ovarian and breast neoplasia (Scott *et al.*, 2006). The identification of this gene by the ACA demonstrates its ability to identify biologically relevant features in challenging data sets.

5. Conclusions

When applied to several high-dimensional data sets, the ACA achieved higher prediction accuracies than all other feature selection methods examined. In contrast to previous applications of optimization algorithms, the ACA yielded high accuracies without the need to preselect a small percentage of genes. Furthermore, the ACA was able to identify small subsets of genes related to both tissue of origin and neoplasia, demonstrating the algorithm's ability to identify highly predictive and biologically relevant genes in data sets with large numbers of features.

References

Akiyama, T. & Kawasaki, Y. (2006) Wnt signaling and the actin cytoskeleton. *Oncogene*, **25**, 7538–7544.

Albrecht, A., Vinterbo, S. A. & Machado, L. O. (2003) An epicurean learning approach to gene-expression data classification. *Artif. Intell. Med.*, **28**, 75–87.

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

Antonov, A. V., Tetko, I. V., Mader, M. T., Budczies, J. & Mewes, H. W. (2004) Optimization models for cancer classification: extracting gene interaction information from microarray expression data. *Bioinformatics*, **20**, 644–652.

Bagirov, A. M., Ferguson, B., Ivkovic, S., Saunders, G. & Yearwood, J. (2003) New algorithms for multi-class cancer diagnosis using tumor gene expression signatures. *Bioinformatics*, **19**, 1800–1807.

Dorigio, M. & Gambardella, L. M. (1997) Ant colonies for the travelling salesman problem. *Biosystems*, **43**, 73–81.

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

Hartigan, J. A. & Wong, M. A. (1979) A K-means clustering algorithm. *J. R. Stat. Soc.*, **28**, 100–108.

Hong, J. & Cho, S. (2006) Efficient huge-scale feature with speciated genetic algorithm. *Pattern Recognit. Lett.*, **27**, 143–150.

Jeffery, I. B., Higgins, D. G. & Culhane, A. (2006) Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, **7**, 359.

Lin, T., Liu, R., Chen, C., Choa, Y. & Chen, S. (2006) Pattern classification in DNA microarray data of multiple tumor types. *Pattern Recognit.*, **39**, 2426–2438.

Liu, J. J., Cutler, G., Li, W., Pan, Z., Peng, S., Hoey, T., Chen, L. & Ling, X. B. (2005) Multiclass cancer classification and biomarker discovery using GA-based algorithms. *Bioinformatics*, **21**, 2691–2697.

NAGATA, K., KAWAJIRI, A., MATSUI, S., TAKAGISHI, M., SHIROMIZU, T., SAITOH, N., IZAWA, I., KIYONO, T., ITOH, T. J., HOTANI, H. & INAGAKI, M. (2003) Filament formation of MSF-A, a mammalian septin, in human mammary epithelial cells depends on interactions with microtubules. *J. Biol. Chem.*, **278**, 18538–18543.

OOI, C. H. & TAN, P. (2003) Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, **19**, 37–44.

PENG, S., XU, Q., LING, X. B., PENG, X., DU, W. & CHEN, L. (2003) Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS Lett.*, **555**, 358–362.

RAMASWAMY, S., TAMAYO, P., RIFKIN, R., MUKHERJEE, S., YEANG, C., ANGELO, M., LADD, C., REICH, M., LATULIPPE, E., MESIROV, J. P., POGGIO, T., GERALD, W., LODA, M., LANDER, E. S. & GOLUB, T. R. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA*, **98**, 15149–15154.

R DEVELOPMENT CORE TEAM. (2006) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at http://www.R-project.org.

RESSOM, H. W., VARGHESE, R. S., ORVISKY, E., DRAKE, S. K., HORTIN, G. L., ABDEL-HAMID, M., LOFFREDO, C. A. & GOLDMAN, R. (2007) Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics*, **23**, 619–626.

SCOTT, M., MCCLUGGAGE, W. G., HILLAN, K. J., HALL, P. A. & RUSSELL, S. E. H. (2006) Altered patterns of transcription of the septin gene, SEPT9, in ovarian tumorigenesis. *Int. J. Cancer*, **118**, 1325–1329.

SHEN, R., GHOSH, D., CHINNAIYAN, A. & MENG, Z. (2006) Eigengene-based linear discriminant model for tumor classification using gene expression microarray data. *Bioinformatics*, **22**, 2635–2642.

SUBRAMANI, P., SAHU, R. & VERMA, S. (2006) Feature selection using Haar wavelet power spectrum. *BMC Bioinformatics*, **7**, 432.

WEST, M. (2003) Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Stat.*, **7**, 723–732.