Session: 18 **Abstract: 15** **Email: krobbin1@uga.edu**

The use of the ant colony algorithm for analysis of high-dimension gene expression data sets.

K. R. Robbins, S. Joseph, W. Zhang, J. K. Bertrand, and R. Rekaya
Correspondence: krobbin1@uga.edu

ABSTRACT

The analysis of microarray data is quickly becoming common place in the field of animal science; however, due to the high dimensions and complex structure of expression information, traditional statistical models may be inadequate for the analysis of such data. To address issues associated with commonly used methods for the identification of predictive genes sets, the ant colony algorithm (ACA) is proposed for use on data sets with large numbers of features and complex structures. The ACA is an optimization algorithm capable of modeling complex data structures without the need for explicit parameterization. The incorporation of prior information and communication between simulated ants allow the ACA to search the sample space more efficiently than other optimization methods. When applied to a high-dimensional cancer microarray data set, the ACA was able to identify small subsets of highly predictive and biologically relevant genes without the need for simplifying assumptions. Using genes selected by the ACA to train a latent variable model yielded increases in prediction accuracy of 16.6% and 6.5% when compared to genes sets selected by test statistics and other optimization models. Furthermore, the ACA was able to converge to good solutions without the need for significant truncation of the data, as required by the other optimization algorithms. The ACA was also able to achieve higher prediction accuracies using fewer selected genes than the test statistics. This was attributed to ability of ACA to model the complex gene interactions, yielding gene lists far less redundant than those selected by test statistics used by statistical methodologies.

INTRODUCTION

With the dawn of the “omics” era in biological sciences it has become possible to collect thousands of data points on traits of interest, be it single nucleotide polymorphisms, gene expression, protein expression, or metabolomic information. With this information it is hoped that the mechanisms underlying traits of interest can be elucidated and understood with a resolution never before possible; however, with the high-dimensions of these datasets, identifying biological relevant features can be challenging. Due to the high cost of these technologies, the number of biological replicates is often quite small relative to the number of data points captured per subject. In the area of statistical genomics this can lead to simplifying assumptions that do not hold true. In the analysis of gene expression data, models are often nested within gene as there are not enough degrees of freedom to estimate all possible gene interactions (Wofinger et al. 2001). Though these methods can be effective in identifying genes with significant marginal contributions they do not take into account the contribution of a feature when grouped with other important features (Shen et al., 2006). As a result, these methods may select groups of highly correlated genes, which in turn, may reduce the predictive power of selected genes. As such, these methods may not be suitable for applications in human medicine and breeding genetics, were the ultimate goal is to identify genomic features that can predict a given phenotype such as disease status, drug response or offspring performance.

For applications in which highly predictive sub-sets of features are needed, machine learning and optimization algorithms may be better suited than nested models. These methodologies require no explicit modeling of data structures, but rely on simple algorithms that are often based on natural processes. The ant colony algorithm (ACA) is a machine learning technique that simulates the positive feed-back system used by ant colonies to find the shortest route to a food source through the use of pheromone trails (Dorigio and Gambardella, 1997). Each simulated ant evaluates a set of features and updates a pheromone function, which serves as a common memory for all simulated ants, based on the performance of that classifier. The communication between ants has a synergistic effect that results in optimal solutions being reached in a computationally efficient manner (Dorigio and 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 feasible for high dimension data sets.

For this study the ACA was implemented using the high-dimensional GCM data-set (Ramaswamy et al., 2001), containing 16,063 genes and 14 tumor classes, with very limited pre-filtering, and compared to several other feature selection methods, as well as previously published results to determine its efficacy in identifying highly predictive classifiers.

MATERIALS AND METHODS

Classification

A Bayesian regression model was used to predict tumor type in the form of a probability $p_{ic}(y_{ic}=1)$, with $y_{ic} = 1$ indicating that sample i is from tumor 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:

$$y_{ic} = \begin{cases} 1 \text{ if } l_{ic} \geq 0 \\ 0 \text{ if } l_{ic} < 0 \end{cases}$$

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

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

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

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

$$p_{ic}(y_{\mathrm{ic}}=1)=\mathbf{\Phi(X_{ic}}\boldsymbol{\beta}_\mathbf{c}\mathbf{)} \text{ and } p_{ic}(y_{\mathrm{ic}}=1)=1-\mathbf{\Phi(X_{ic}}\boldsymbol{\beta}_\mathbf{c}\mathbf{)}$$

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

Subject i was classified as having tumor class c if $p_{ic}(y_{\mathrm{ic}}=1)$ was the maximum of the vector $\mathbf{p_i}$, containing all $p_{ic}(y_{\mathrm{ic}}=1)$ c=1,…, *nc*, where *nc* is the number of tumor classes in the data set.

Gene Selection

Features were selected using nested models and the ACA. For nested models fold changes (FC), t-statistics (T), and penalized t-statistics (PT) were calculated for each gene separately. The ACA evaluated groups of genes and updated pheromone levels based on prediction accuracy obtained using LVM.

*Ant Colony Algorithm***:** 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 and Gambardella, 1997; Ressom et al., 2006), where the probability of sampling feature *m* at time *t* is defined as:

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

where $\tau_{mc}(t)$ is the amount of pheromone for feature *m* (out of a total of *nf* features) of tumor class c at time *t*; η_{mc} is some form of prior information on the expected performance of feature *m* of tumor class c; 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 the average FC, T, and PT scores of a gene for a given tumor type.

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 equation (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:

$$\tau_{mc}(t+1)=(1-\rho)*\tau_{mc}(t)+\Delta\tau_{mc}(t) \qquad (2)$$

where ρ is a constant between 0 and 1 that represents the rate at which the pheromone trail evaporates; $\Delta\tau_{mc}(t)$ is the change in pheromone level for feature *m* for tumor class *c* 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

Following the update of pheromone levels according to equation (2), the PDF is updated according to equation (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 rate 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 select based on the level of pheromone deposited on each feature.

GCM data set

The data set contained 198 samples collected from 14 tumor 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), LU (leukemia), RE (renal cell carcinoma), PA (pancreatic adenocarcinoma), OV (ovarian adenocarcinoma), ME (pleural mesothelioma), and CNS (central nervous system). The unedited data set contained the intensity values of 16063 probes generate using Affymetrix high density oligonucleotide microarrays, and calculated using Affymetrix GENECHIP software (Ramaswamy et al, 2001). Following the thresholding of intensity values to a minimum value of 20 and a maximum value of 16000, a log base 2 transformation was 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.

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 (GASS (Lin et al., 2006), GA/MLHD (Ooi and Tan, 2003), and MAMA (Antonov et al., 2004)) using independent test, performed on the same training and validation data sets originally formed by Ramaswamy et al., 2001 (GCM split). The proposed ACA 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 achieved increases of 13.9%, 40%, and 16.6% in accuracy over the FC, T, and PT methods of feature selection, respectively.

Table 1. Accuracy (%) of tumor class predictions using ant colony algorithm (ACA) and several previously published methods.

Method	GCM split[a]	Replicated splits
ACA(14525[b])	90.7	84.8
FC(14525)	79.6	74.8
T(14525)	64.8	____
PT(14525)	77.8	74.4
AVG[c](14525)	79.6	74.8
GASS(1000)	81.5	____
GA/MLHD(1000)	76	____
MAMA	85.2	____

[a] Split used by Ramaswamy et al 2001; [b] Number of genes selected prior to the implementation of feature selection algorithm; [c] Weighted average of scaled fold change, t-test, and penalized t-test values.

To examine the degree of collinearity present in the top genes, as selected by ACA and the nested methods, the top 30 features selected for BR were clustered using k-means (R Development Core Team, 2006) and then correlated. The correlations between selected features can be seen in the form of heat matrices found in Figure 1 where lighter shades indicate high correlations and darker shades indicate low correlations. It is clear, when looking at the heat matrices, that the features selected using nested models exhibit substantially more collinearity.

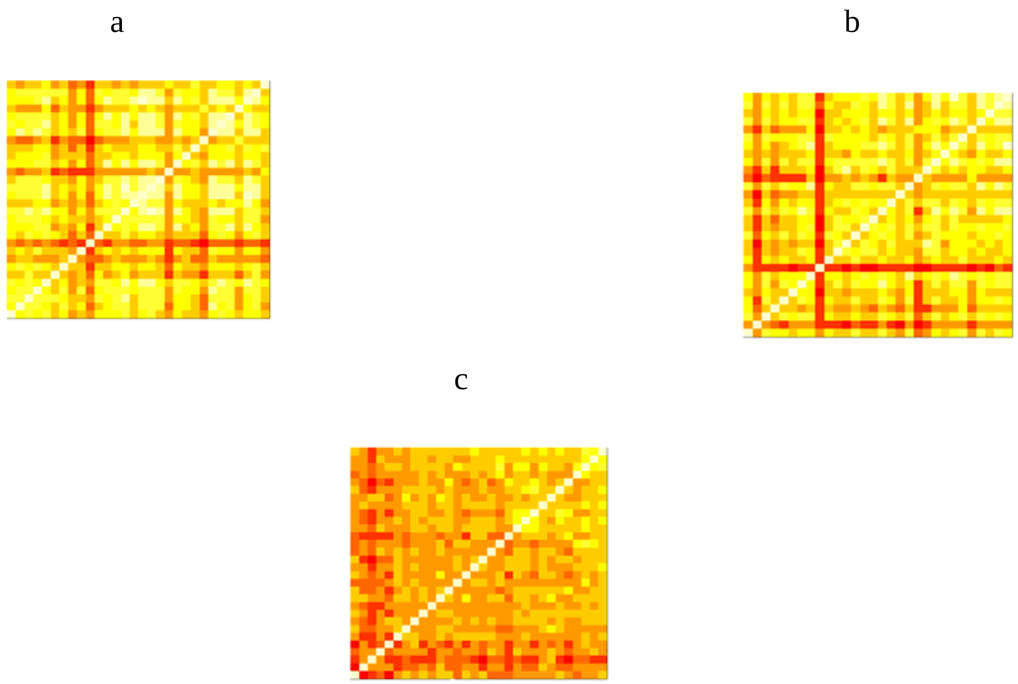

Figure 1. Heat matrices between the top 30 genes selected for breast adenocarcinoma tumors based on: a. fold change; b. penalized t-test; and c. ant colony algorithm (red/orange = low correlation, white/yellow = high correlation).

DISCUSION

The performance of the ACA model was superior, not only to the nested methods used in this study, but to several reported results using the GCM data set. The ACA consistently yielded superior accuracies using fewer genes than the nested methods. 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 (Lin et al., 2006; Ooi and Tan 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 to 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 pre-selected genes.

The reduction in the collinearity of genes as selected by ACA, particularly in tumor 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. 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. To this end the ACA identified several small subsets of genes capable of obtaining high accuracies in cross validation for many of the 14 tumor types contained in the GCM data set.

REFERENCES

Antonov, A.V., Tetko, I.V., Mader, M.T., Budczies, J. and H. W. Mewes 2004, 'Optimization models for cancer classification: extracting gene interaction Information from microarray expression data', *Bioinformatics,* 20, 644-652.

Dorigio, M. and L. M. Gambardella 1997, 'Ant colonies for the travailing salesman problem', *BioSystems,* 43, 73-81.

Lin, T., Liu, R., Chen, C., Choa, Y. and S. Chen 2006, 'Pattern classificationin DNA microarray data of multiple tumor types', *Pattern Recognition,* 39, 2426-2438. interactions with microtubules', *J. of Biol. Chem.,* 278, 18538-18543.

Ooi,C.H. and P. Tan 2003, 'Genetic algorithms applied to multi-class prediction for the analysis of gene expression data', *Bioinformatics,* 19, 37-44.

R Development Core Team 2006, 'R: A language and environment for statistical computing' R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

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. and T. R. Golub 2001, 'Multiclass cancer diagnosis using tumor gene expression signatures', *Proc. Natl. Acad. Sci.,* 98, 15149-15154.

Ressom, H.W., Varghese, R.S., Orvisky, E., Drake, S.K., Hortin, G.L., Abdel-Hamid, M. Loffredo, C.A. and R. Goldman 2006, 'Ant colony optimization for biomarker identification from MALDI-TOF mass spectra', *Proc. of the 28th EMBS Annual Inter. Conf.*, 4560-4563.

Shen, R., Ghosh, D., Chinnaiyan, A. and Z. Meng, 2006, 'Eigengene-based linear discriminant model for tumor classification using gene expression microarray data' *Bioinformatics*, 22, 2635-2642.

West, M. 2003, 'Bayesian factor regression models in the "Large p, Small n" paradigm', *Bayesian Statistics*, 7, 723-732.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P, Afshari, C. and R. S. Paules, 2001, 'Assessing gene significance from cDNA microarray expression data via mixed models', *J Comput Biol*, 8(6), 625-637.