STATISTICAL METHODS FOR THE ANALYSIS OF COMPLEX GENOMIC DATA

by

KELLY R. ROBBINS

(under the direction of Romdhane Rekaya and J. Keith Bertrand)

ABSTRACT

The use of genomic technology has the potential to provide invaluable insight into the mechanisms of several important traits. Unfortunately this information comes at a cost, in terms of the high-dimensions and sometimes poor quality of the data. One potential application of genomics is the diagnosis of diseases, such as Alzheimer's disease, with ambiguous and confounding clinical markers. Of course to predict disease statuses, an algorithm must first be trained using a data set in which disease statuses are known without error. In the case of incipient Alzheimer's disease this is rarely the case. To this end a misclassification algorithm was applied to a data set containing healthy individuals and incipient Alzheimer's patients to examine the effects of potential misclassification on diagnostic accuracy. Results obtained without invoking the misclassification algorithm showed limited predictive power of the model. When the misclassification algorithm was invoked significant increase in the model's predictive ability were obtained. These results demonstrate the utility of the misclassification algorithm in data sets containing potential misdiagnosis.

In addition to potential misdiagnosis, the high-dimensions of genomic data sets can also pose substantial issues for statistical analysis. Due to the large number of features in many genomic datasets, explicit modeling of gene interactions is often infeasible. To eliminate the need for simplifying assumptions a machine learning algorithm, referred to as the ant colony

algorithm (ACA), was adapted for analysis of high-dimension genomic data. In a study examining the selection of predictive gene expression patterns, the performance of the ACA was compared to several standard methodologies. When applied to high-dimensional data sets, the ACA was able to identify small subsets of highly predictive genes, yielding superior prediction accuracy when compared to several standard feature selection methods. In an application involving single nucleotide polymorphism marker data, a modified ACA was implemented to identify markers associated with a binary trait under the influence of interacting loci. When compared to marginal effects models, the ACA demonstrated superior performance under several simulation scenarios with p-values for associated SNP being more significant using the ACA, resulting in substantial increases in power.

INDEX WORDS: Ant colony optimization, genomics, latent variable model, logistic regression, misclassification algorithm

STATISTICAL METHODS FOR THE ANALYSIS OF COMPLEX GENOMIC DATA

by

KELLY R. ROBBINS

B.S., The University of Tennessee, 2002

M. S., The University of Georgia, 2004

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2007

STATISTICAL METHODS FOR THE ANALYSIS OF COMPLEX GENOMIC DATA

by

KELLY R. ROBBINS

Co-Major Professors: Romdhane Rekaya
Joseph Keith Bertrand

Committee: Ignacy Misztal
Samuel Aggrey

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2007

DEDICATION

I would like to dedicate this work to the loving memory of my grandfather, Dr. Robert Baker. His life has, and always will, serve as inspiration for my family and myself.

ACKNOWLEDGEMENTS

Throughout the course of my studies I have been blessed to be surrounded by many amazing people, whose support and encouragement were instrumental in the successful completion of my doctoral studies. I would like to begin by thanking my major professors Dr. Rekaya and Dr. Bertrand. It has been an extremely rewarding experience working with you both, and I will always be grateful for all of the opportunities you have given me. Without your knowledge, guidance, and support, I wouldn't be where I am today. I would also like to thank Drs. Misztal and Aggrey. I have truly enjoyed working with you over the past five years. Your guidance as committee members has been invaluable. In addition to my committee, I would like to thank Robyn, Matt Spangler, and Travis, these past five years would have been incredibly difficult without your friendship and advice.

To my friends and wonderful girlfriend, you have kept me sane through these past few years, and for that I am grateful. Sarah, your love and support have meant so much to me, whenever I was feeling overwhelmed you were there to keep me focused on what is really important. Frank, your never ending series of disasters have served as constant distractions from the stress of my work. The fact that you almost killed me has allowed me to appreciate life like never before. Matt, destroying you at Halo, foosball, basketball, and just about every competitive activity imaginable, has always managed to cheer me up during difficult times.

Finally, and most importantly, I would like to thank my wonderful family for all their love and support. Grandma, thank you for keeping me in your prayers. Mom, your wisdom and guidance have been so important throughout my life, and I can't thank you enough. Reigan, your friendship and unique perspective have been invaluable; my conversations with you are always

priceless. Dad, having been through this before, your advice has been greatly appreciated and extremely helpful throughout my graduate studies.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

In the last decade there has been a rapid expansion of new technologies enabling the efficient collection of large quantities of genomic data in both economically important and model organisms. These high-throughput technologies, capable of capturing thousands of data points on single nucleotide polymorphisms (SNP), gene expression, protein expression, and metabolomic data, have ushered in the 'omic' era of quantitative biology. Unfortunately the high-throughput nature of these technologies has proven to be a double-edged sword, as the sometimes poor data quality, high-dimensions, and complex structure of these datasets have made effective modeling and extraction of useful information difficult.

Issues surrounding data quality, specifically the misclassification of dependent variables, represent a pressing need in human medicine and animal production. Misclassification is a common problem in areas of human medicine dealing with neurological disorders, which in many cases have broad and overlapping symptoms. In animal science, traits such as meat quality and confirmation scores may have misclassification due to the categorization of continuous traits. Similarly non-return to heat, a measurement of fertility, suffers from failures to accurately detect heat. The presence of these misclassified subjects can substantially erode a model's power to predict and diagnose, as well as its ability to identify, or select, features associated with the trait of interest.

The high-dimensions and complex structure of genomic data can also have substantial and negative impacts on feature selection. Control of experiment-wide error rates in these high-

dimension datasets can severely limit the power to detect causative mutations and expression profiles related to traits of interest. Due to the complex interactions among genes and their products, large numbers of parameters must be estimated from data sets with relatively small replication. As a result, models are often applied that analyze each gene independently or account for only marginal effects of markers, ignoring important epistatic relationships.

Clearly, these issues represent important needs that must be addressed if the full benefits of the 'omic' era are to be reaped. Methodologies must be developed to account for potential misclassification in poorly defined and difficult to measure traits. Techniques for feature selection need to account for the complex interactions inherent in genomic data, and given the high-dimensions of these data, these methodologies must be computationally efficient. To this end, the goals of the current studies are to 1) refine and apply a misclassification algorithm to gene expression data to examine the potential misclassification of incipient Alzheimer's patients and examine the effects of said misclassification on the prediction accuracy of a trained machine learning algorithm and 2) develop and apply feature selection methodologies capable of efficiently identifying biologically relevant and highly predictive gene expression profiles and single nucleotide polymorphisms in the presence of complex genomic structures.

CHAPTER 2

REVIEW OF LITERATURE

Misclassification and Machine Learning

Misclassification of dependent variables is a major issue in many areas of science that can arise when indirect markers are used to classify subjects, or when continuous traits are treated as categorical. In human medicine this can have significant impacts on diagnostic accuracy (Gould et al. 2007; Ramsley, 2006), particularly for neurological disorders such as attention deficit hyperactivity disorder (ADHD) and Alzheimer's disease (AD). In a review, Yeh et al. (2004) concluded that uncertainty in the status of ADHD patients limited researchers' ability to identify underlying genetic factors that could be used for genetic screening. This is evidenced by the fact that the true prevalence of ADHD is unknown, despite the use of multiple clinical tests (Graetz et al., 2001; Gadow et al., 2000; Sarkis, 2000). Similarly (AD) can be very difficult to diagnosis, particularly in patients suffering from the earliest stages of AD due to the similarities in the cognitive decline of incipient AD patients and declines associated with the normal aging process (Ash et al., 2000; Haroutunian et al., 1999). This confounding of the effects of aging and incipient AD could lead to significant misdiagnosis rates and has spurred research to develop more accurate methods of diagnosis to take advantage of drugs that can delay the development of the more devastating stages of the disease when administered to incipient AD patients (Mueller et al. 2005).

In animal science, misclassification can negatively affect both genetic improvement and the ability to ascertain the biological mechanisms for traits of interest; however, this has been the focus of little research. Scores used to evaluate animal conformation and meat quality could have

some uncertainty due to the continuous nature of the underlying factors, particularly when dealing with animals or meat samples with values near class boundaries. Studies examining the effects of uncertainty in the binary variable of success at first insemination found that failure to account for uncertainty can lead to biased parameter estimates (Sapp et al., 2005; Spangler et al., 2006). Rekaya et al. (2001) found, using simulated data where 5.6% of the binary data were misclassified, that failure to account for misclassification resulted in biased parameter estimates, with the true values of falling outside the 95% high density posterior interval.

Fortunately, when dealing with traits influenced by genetic factors, expression profiles and genetic markers can provide direct measurements of the underlying mechanisms controlling phenotypic responses. As a result, a great deal of research has been focused on the use of this genomic information for disease diagnostics and class prediction. In the area of disease diagnostics, several machine learning algorithms have been developed that train classifiers, a set of criteria for class prediction, on data sets of known outcomes. These trained classifiers are then used to predict the statuses of samples with unknown classifications.

Using a statistical latent variable model (LVM), disease statuses and categorical variables can be modeled using expression data by assuming a continuous underlying random variable controlling the trait of interest such that:

$$y_i = \begin{cases} 1 & \text{if } l_i \geq 0 \\ 0 & \text{if } l_i < 0 \end{cases}$$

The liability l_i can be modeled using a linear regression model as:

$$l_i = \mathbf{X}_i\boldsymbol{\beta} + e_i \qquad E(l_i) = \mathbf{X}_i\boldsymbol{\beta} \qquad e_i \sim N(0,1) \qquad (1)$$

where $\mathbf{X}_i$ corresponds to row i of the matrix **X,** containing explanatory responses, such as gene expression.

The link function of the expectation of the liability $\mathbf{X}_i\boldsymbol{\beta}$ with the binary response y_i is then constructed via a probit model yielding the following equations**:**

$$p_i(y_i = 1) = \boldsymbol{\Phi}(\mathbf{X_i}\boldsymbol{\beta}) \text{ and } p_i(y_i = 1) = 1 - \boldsymbol{\Phi}(\mathbf{X_i}\boldsymbol{\beta}) \quad (2)$$

where $\boldsymbol{\Phi}$ is the standard normal distribution function, yielding the following relationships:

$$y_i = \begin{cases} 1 & \text{if } \boldsymbol{\Phi}(\mathbf{X_i}\boldsymbol{\beta}) \geq 0.5 \\ 0 & \text{if } \boldsymbol{\Phi}(\mathbf{X_i}\boldsymbol{\beta}) < 0.5 \end{cases}$$

Through the use of these link functions, regressions for the observed classifications can be fit on the continuous underlying scale. Using the estimated regression coefficients, or trained classifier, unknown statuses can be predicted. In several studies LVM, as well as other approaches such as support vector machines (SVM), decision trees (DT), and neural networks (NN), have achieved high prediction accuracies, thus demonstrating the utility of genomic information for class prediction (West et al. 2001, Ressom et al. 2007; Albetar et al., 2006; Jung et al., 2007).

Gulob et al. (1999) used a classifier trained using gene expression values captured in microarray experiments to distinguish between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) tumors. Using a leave-one-out cross-validation scheme 36 out of 38 samples were classified correctly. Ramaswamy et al. (2001) used gene expression data to classify tissues into 14 different tumor types using an SVM and obtained a classification accuracy of 78%. In the field of animal science these methodologies could be used to train classifiers and identify biomarkers for reproductive, fitness, and meat quality traits, which can be difficult and costly to measure. Predictions obtained from these classifiers or the biomarkers themselves could then be incorporated in to evaluations or used a selection tools. Unfortunately,

the performance of these algorithms is contingent on the assumption that the statuses of the training data are known without error, which may not always be the case.

When misclassification is present, prediction accuracies and confidences can be reduced. Zhang et al. (2006) found, that as the number of misclassified samples in the training set increased, the certainty of class predictions decreased. Furthermore, the selection of features, (genes, proteins, etc.) for use in training classifiers and selection of livestock, require the identification of causal relationships with the trait of interest. In situations where the true phenotypic value of the trait is not known without error, the power to detect these relationships could be severely reduced (Edwards et al. 2005). As such there is a need to develop methodologies to identify potentially misclassified subjects, and examine the effects of misclassification on class prediction accuracy and experimental power to detect causal relationships.

An approach treating the statuses of dependent variables as a mixture of correct and incorrect classifications lends itself well to Bayesian methodologies. Rekaya et al. (2001) sampled a vector of Bernoulli random variables ***m*** taking on the values of 0 and 1 for correctly and incorrectly classified samples, respectively. The sampling of each value m_i, i=1, ..,n was based on a binomial misclassification probability related to the fit of the model to the observed record and can be represented by the following equations:

$$p(m_i = 1 \mid \boldsymbol{\pi}, \mathbf{m}_{-i}, \mathbf{X}, \boldsymbol{\beta}, \mathbf{r}) = \frac{\pi[\boldsymbol{\Phi}(\mathbf{X}_i\boldsymbol{\beta})]^{(1-z_i)}[1-\boldsymbol{\Phi}(\mathbf{X}_i\boldsymbol{\beta})]^{z_i}}{K}$$

and

$$p(m_i = 0 \mid \boldsymbol{\pi}, \mathbf{m}_{-i}, \mathbf{X}, \boldsymbol{\beta}, \mathbf{r}) = \frac{(1-\pi)[\boldsymbol{\Phi}(\mathbf{X}_i\boldsymbol{\beta})]^{z_i}[1-\boldsymbol{\Phi}(\mathbf{X}_i\boldsymbol{\beta})]^{(1-z_i)}}{K}$$

where π is the probability of observing misclassification in the data sets and is sampled from a Beta distribution; $\mathbf{\Phi(X_i\beta)}$ can be obtained through equation (1) and (2); and

$$K = \pi[p_i(\mathbf{Q_i D\gamma})]^{(1-z_i)}[1-p_i(\mathbf{Q_i D\gamma})]^{z_i} + (1-\pi)[p_i(\mathbf{Q_i D\gamma})]^{(1-z_i)}[1-p_i(\mathbf{Q_i D\gamma})]^{z_i}.$$

The binary statuses for records were switched for all iterations in which the corresponding value m_i is sampled as 1. This approach was successfully implemented with a LVM threshold model to account for uncertainty in the status of binary random variables using both simulated and real dairy cattle data.

In an application to breast cancer tumors, using a real data based simulation, Zhang et al. (2006) successfully applied a Bayesian misclassification algorithm with a LVM for prediction of breast cancer status, identifying all misclassified samples using gene expression. In addition to observed decreases in the certainty of predictions, it was found that the presence and number of misclassified samples in the training data set greatly affected the selection of discriminative genes. These results suggest that misclassification can not only affect prediction accuracy and certainty, but the identification of important biomarkers as well and should be tested on traits that are traditionally difficult to classify such as incipient AD.

Feature selection

Issues affecting the identification of predictive and important genomic features, or feature selection, are not limited to potential misclassification. The high dimensions and complex structure of high through put data can have large impacts on the identification of informative features (Marchini et al. 2005; Brinza et al. 2006, Shen et al. 2006). Jeffery et al. (2006) found that the performance of various feature selection methods varied greatly depending on the number of features selected and the structure of the data sets being analyzed, with each methodology selecting substantially different list of features, yielding varied performance in

prediction accuracy. Since the choice of feature selection methods can have such considerable impacts on both the prediction accuracy of classifiers as well as the biological understanding of important traits, methodologies must be developed that can consistently identify biologically relevant and predictive features.

Methods of feature selection fall into three main categories: filter methods, wrapper methods, and embedded methods. Filter methods use values such as t-test, fold change, signal to noise ration, or penalized t-test, to name a few. These values are calculated for each feature independently and all features having a value greater than some pre-defined threshold are selected. Wrapper methods use the marginal contribution of a feature to the prediction accuracy of a classifier or the performance of a selected group of features as the selection criteria, and can be implemented in a manner similar to several common model selection procedures. Embedded methods select features as part of the training process of some machine learning methods, such as SVM. Since embedded methods are specific to a select few machine learning algorithms, discussion will focus on filter and wrapper methods.

Due to the large number of features and unknown epistatic relationships encountered in exploratory studies, a prohibitively large number of parameters must be estimated to fully model the structure of the data. Given the high number of parameters and the relatively low numbers of biological replicates in genomic studies, nested models, which ignore potential gene interactions, are often used (Wolfinger et a l., 2001; McClurg et al., 2006). Given the ease of implementation and computational speed of the nested models, filter based methods are commonly used for feature selection; however, when selecting features for class prediction, the evaluation of each gene separately can lead to the selection of highly redundant features that tend to yield poor and inconsistent performance in prediction accuracy (Shen et al. 2006). Additionally, these methods

may fail to select genes that yield only moderate prediction power when evaluated alone but provide substantial information when combined.

The use of wrapper methods, which examine the performance of groups of genes, can address many of the issues associated with filter methods; however the high-dimensions of genomic data sets make it computationally impossible to examine all possible combinations of features. This problem of dimensionality can be viewed as an optimization problem, for which several algorithms commonly referred to as optimization algorithms, have been developed. These algorithms are designed to search large sample spaces for globally optimal solutions and have been applied to wide range of problems ranging from protein folding prediction to interactions of robots on assembly lines (Shen et al., 2004; Shymygelska and Hoos et al., 2005; Kreiger et al., 2000; Ding et al, 2005).

Many of these algorithms are derived from natural processes such as the entropy of atoms (simulated annealing, SA), the nervous system (NN), and natural selection (genetic algorithm, GA) (Albrecht et al. 2003; Jung et al., 2007; Huang and Chang, 2006). While these algorithms yield optimal results for many applications (Albrecht et al. 2003), some studies suggest that they may be inefficient and poorly suited for high-dimensional genomic data sets (Hong and Cho, 2005) expand. Lin et al. (2006) found that, while the GA achieved good solutions on the relatively small SRBCT data set (Khan et al., 2001) , when applied to the high-dimension multi-class GCM cancer data set (Ramaswamy et al. 2001) the GA only converged to good solutions after the truncation of the expression of over 14,000 genes.

While some optimization algorithms may be insufficient for use on high-dimension genomic data sets, Dorigo and Gambardella (1997) developed an algorithm based on the behavior of ant colonies capable of reaching optimal solutions more efficiently. The ant colony

algorithm (ACA) mimics the ability of real ant colonies to find the shortest route to a food source using communication in the form of chemical pheromone trails. Ants that choose shorter paths will transverse the distance more quickly, thus depositing pheromone at faster rate. As the pheromone level builds on the shortest path, ants will begin to preferentially choose the better path, leading to a positive feed back system which results in all ants eventually taking the shortest route. The ACA accomplishes this through the use of a statistical pheromone function in the form of a probability density function (PDF):

$$P_m(t) = \frac{(\tau_m(t))^{\alpha} \eta_m^{\beta}}{\sum_{m=1}^{nf} (\tau_m(t))^{\alpha} \eta_m^{\beta}} \tag{3}$$

where $\tau_m(t)$ is the amount of pheromone for path *m* at time *t*; η_m is some form of prior information on the expected performance of path *m*; α and β are parameters determining the weight given to pheromone deposited by ants and a priori information on the features used to construct the paths, respectively.

The PDF is updated each iteration based on the performance of a subset of features using the following equation:

$$\tau_m(t+1) = (1-\rho) * \tau_m(t) + \Delta\tau_m(t) \tag{4}$$

where ρ is a constant between 0 and 1 that represents the rate at which the pheromone trail evaporates; $\Delta\tau_m(t)$ is the change in pheromone level for feature *m* based on the performance of a the selected path. As the weights, or pheromone, on paths yielding optimal performance build, the probability that ants in subsequent iterations choose the optimal path increases. This simulated ant colony yields a positive feed back system similar to the one observed in real ant colonies.

The procedure can be summarized in the following steps:

1) Each ant selects a path.
2) The change pheromone for each path is calculated ($\Delta\tau_m(t)$).
3) Following the update of pheromone levels according to equation (4), the PDF is updated according to equation (3) and the process is repeated until some convergence criteria are met.

Unlike other algorithms, such as the GA where good solutions may take several iterations to disseminate across the pedigree through mating and selection, the update of the statistical pheromone equation allows goods solutions, obtained by one ant, to be immediately communicated to all ants. This property, combined with the incorporation of prior information which focuses the search for features on regions of the sample space more likely to yield good solutions, could make the ACA more suitable for use on high-dimension expression data sets.

Dorigio and Gambardella (1997) applied the ACA to the traveling salesman problem, an application commonly used to test the performance of optimization algorithms. Given a certain number of cities to visit and the distances between them, several algorithms were implemented to determine the order in which the cities should be visited to minimize the total distance traveled. It was observed that the communication between ants showed a synergistic effect as the CPU time taken to reach optimal solutions when ants communicated decreased as the number of ants increased from 1 to 10. When ants shared no communication, CPU time remained constant as the number of ants increased. Furthermore, it was shown that the ACA converged to optimal solutions in far less iterations than GA, NN, or SA algorithms, demonstrating the superior efficiency of the ACA.

The application of Dorigio and Gambardella (1997) was very well suited to the mechanism of the pheromone trail; however when applying ACA to feature selection problems

the use of the pheromone trail is more abstract. In this case of feature selection, the path is a subset of features select by the ants. The performance of the features can be computed as the prediction accuracy obtained from a classifier trained using the selected feature subset and then used to update the PDF. Ressom et al. (2007) applied the ACA to matrix assisted laser desorption/ionization time-of-flight (MALDI-TOF) spectral data to identify peaks associated with hepatocellular cancer (HCC). The ACA was implemented with SVM and selected eight out of 228 peaks as being highly predictive. The eight peaks yielded 94% sensitivity and 100% specificity in distinguishing HCC cases from cirrhosis cases in a cross validation study.

While the results of these studies demonstrate the effectiveness and efficiency of the ACA for feature selection, the 228 features in Ressom et al. (2007) study are far fewer than the number of features encountered in many gene expression studies. In order to fully evaluate the efficacy of the ACA as a feature selection method for genomic studies, its performance must be evaluated on a high-dimension data set with little to no pre-filtering, as required by other optimization algorithms.

SNP association

Of course the potential applications of the ACA go far beyond the selection of features for class prediction. The identification of causative mutations through the use of sequence mapping has been the focus of intensive research over the past decade. With the use of single nucleotide polymorphism (SNP) it is now possible to saturate the genome with markers. SNP are bi-allelic mutations involving a single base and account for approximately 90% of human genetic variation. SNP occur every 100 to 300 base pairs, yielding a degree of resolution never before possible.

Due to the bi-allelic nature of SNP, single markers are often insufficient to adequately explain variation in complex traits. As a result, haplotypes, groups of linked SNP alleles, are often used in association studies. Given the high-dimensions of SNP data sets, haplotypes are often formed using adjacent SNP in a methodology referred to as the sliding window approach (McClurg et al. 2006). This methodology utilizes a window of *k* SNP in width and slides across the genome *h* SNP at a time. For each window, haplotypes effects are estimated and tested for significant associations. In the case of binary traits, such as disease traits, haplotype effects are estimated as log odds ratios (*lor*) using logistic regression. The relationship between the *lor* and the binary response can be expressed as:

$$y_i = \begin{cases} 1 & \text{if } lor_i \geq 0 \\ 0 & \text{if } lor_i < 0 \end{cases}$$

The log odds ratio lor_i is modeled as:

$$lor_i = \ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{X}_i\boldsymbol{\beta}$$

where P_i = probability ($y_i = 1$).

The link function of the log odds ratio $\mathbf{X}_i\boldsymbol{\beta}$ with the binary response y_i gives the following equations**:**

$$p_i(y_i = 0) = \frac{1}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})} \text{ and } p_i(y_i = 1) = \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}$$

yielding the following relationships:

$$y_i = \begin{cases} 1 \text{ if } \dfrac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})} \geq 0.5 \\ 0 \text{ if } \dfrac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})} < 0.5 \end{cases}$$

In human applications, SNP marker maps have been used to identify disease related genes for a wide variety of conditions. Hugot et al. (2001) used both micro satellite markers and SNP to examine a region on ch16 believed to be associated with Chrohn's disease. Analysis identified 3 SNP mutations associated with Chronh's disease located in the coding region of NOD2, a gene involved in the activity of microbial pathogen receptors. Woon et al. (2007) examined the role of BMAL1 in type-2 diabetes and hypertension in an association study involving 1,304 individuals. Though tests of individual SNP showed no significant association, haplotype association tests did show evidence of association with two haplotypes for type-2 diabetes and hypertension, respectively. Though the obtained p-values were small ($<2.4 \text{ X } 10^{-3}$), methods for controlling experiment-wise error rate were unclear, thus illustrating issues in maintaining experimental power faced when conducting SNP association studies..

In animal science, marker maps have used to identify regions of the genome associated with economically important traits. Stone et al. 2005 examined bovine chromosome 5 for association with carcass traits in beef cattle. After performing a Bonferroni correction for multiple comparisons, 2 haplotypes at the phosphodiesterase 1B locus were found to be significantly associated with fat thickness and rib fat. In addition to identifying causative mutations, simulation studies have explored the use of genome marker scans in genetic evaluations for the estimation of genomic breeding values which may allow the selection of younger animals thus reducing generation intervals and increasing response to selection. Meuwissen et al. 2001 simulated a population with markers placed every 1 CM with each flanked region containing a QTL. When estimating interval effects using a Bayesian methodology, correlations between simulated and estimated genomic breeding values of .85 were obtained. These results suggest genomic information can be used to accurately predict

breeding values; however, the simulations were overly simplistic and, as such, results should be viewed as optimistic until they can be validated in a real animal population.

While there has been much focus on identifying mutations with large marginal effects for the purposes of disease treatment and animal selection, recent studies have found that gene interactions may play important roles in many complex traits. Coutinho et al. 2007 examined potential interactions between genes involved in autism. All possible 2-way and 3-way combinations of genotypes at seven different loci were examined. Using cross-validation and permutation testing, a significant interaction between two of the loci was detected. Barendse et al. 2007 found significant epistasis between the calpain 1 and alpastatin, genes that affect meat tenderness in several beef cattle breeds. Given these findings, accounting for gene interactions could be important in association studies examining complex traits.

Though these studies where able to identify significant gene interactions, they only accounted for genes previously identified as having significant marginal effects on the traits of interest. This may not represent an optimal strategy, as important epistatic genes may not always show significant marginal effects. Pickrell et al. (2007) showed in a simulation study that many tradition marginal effects models had little power to detect causative mutations in the presence of various types of two-locus gene interactions. Similarly, Marchini et al. 2005 examined the power of models considering only marginal effects compared to those considering two way interactions, using an exhaustive search, as well as, a two stage approach with the first stage selecting markers based on marginal effects. Results showed that under simulated epistasis, the interaction models had greater power to detect both loci; however due to high computational cost, the proposed methodology would require truncation of the data set based on marginal effects when dealing with large numbers SNP.

Given that genome-wide association studies consider thousands of SNP, potentially under the control of a large number of interacting loci, methodologies need to be developed that can account for potential gene interactions in high-dimension data sets. These methodologies should be computationally efficient, eliminating the need to pre-select markers using ineffective marginal effects models, and permitting the use of permutation testing to control family-wise error rates. As with applications involving feature selection in gene expression data sets, optimization algorithms like the ACA could provide the most efficient method for identifying significant associations in the presence of gene interactions in high-dimension data sets.

Clearly, the field of statistical genomics represents an exciting and challenging area of research with potential to increase understanding of human disease and improve selection of important traits in the livestock industry. While the promise of this area of science is beginning to be realized, there are several issues in the analysis of genomic data, resulting from poor data quality, high-dimensions, and the complex nature of many import traits, that must be addressed before the full potential of genomics research can be realized.

Literature Cited

Ashe, K. H. 2000. Learning and Memory in Transgenic Mice Modeling Alzheimer's Disease. Learning & Memory. 8(6):301-308.

Albitar, M., Potts, S. J., Giles, F. J., O'Brian, S., Keating, M., Thomas, D., Clarke, C., Jilani, I., Agullar, C., Estey, E., and H. Kantarjian. 2006. Proteomic-based prediction of clinical behavior in adult acute lymphoblastic leukemia. Cancer. 106(7):1587-1594.

Albrecht, A., Vinterbo, S.A. and L. O. Machado. 2003. An epicurean learning approach to gene-expression data classification. Artif. Intell in Medicine. 28:75-87.

Barendse, W., Harrison, B. E., Hawken, R. J., Ferguson, D. M., Thompson, J. M., Thomas, M. B., and R. J. Bunch. 2007. Epistasis between Calpain 1 and its inhibitor Calpastatin within breeds of cattle. Genetics 176:2601-2610.

Brinza, D., He, J., and A. Zelikovsky. 2006. Combinatorial search methods for Multi-SNP disease association. Proceedings of the 28th EMBS Annual Internation Conference. 5802-5805.

Coutinho, A. M., Sousa, I., Martins, M. et al. 2007. Evidence for epistasis between SLC6A4 and ITGB3 in autism etiology and in the determination of platelet serotonin levels. Hum. Genet. 121:243-256.

Ding, Y. P., Wu, Q. S., and Q. D. Su. 2005. Multivariate Calibration Analysis for metal porphyrin mixtures by an ant colony algorithm. Analytical Sciences. 21:327-330.

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

Edwards, B. J., Haynes, C., Levenstein, M. A., Finch, S. J., and D. Gordon. 2005. Power and sample size calculations in the prescence of phenotype errors for case/control genetic association studies. BMC Genetics. 6:18.

Gadow, K. D., Nolan E. E., Litcher, L., Carlson, GA, Panina, N. Golovakha, E. Sprafkin, J. and E. J. Bromet. 2000. Comparison of attention-deficit/hyperactivity disorder symptom subtypes in Ukrainian schoolchildren. J. Am. Acad. of Child Adolesc. Psychiatry. 39:1520-1527.

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. and E. S. Lander 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 286:531-537.

Gould, C. F., Ly, J. Q., Lattin G. E., Beall, D. P., and J. B. Sutcliffe. 2007. Bone tumor mimics: avoiding misdiagnosis. Curr. Probl. Radiol. May/June :124-141.

Graetz, B. W., Sawyer, M. G., Hazell, P. L., Arney F., P. A. Baghurst. 2001. Validity of DSM-IV ADHD subtypes in a ntionally representative sample of Australian children and adolescents. J. Am. Acad. Child Adolesc. Psychiatry. 40:1410-1417.

Haroutunian, V., Purohit, D. P., and Perl, D. P. 1999. Neurofibrillary tangles in nondemented elderly subjects and mild Alzheimer disease. Archives of Neurology. 56(6):713-8.

Hong, J. and S. Cho 2006. Efficient huge-scale feature selection with speciated genetic Algorithm. Pattern Recognition Lett. 27:143-150.

Huang, H. L. and F. L. Chang. 2007. ESVM: Evolutionary support vector machine for automatic feature selection and classification of microarray data. BioSystems. 90(2):516-528.

Hugot, J. P., Chamaillard, M., Zouali, H. et al. 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature. 411:599-603.

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

Jung, E., Kim, J., Kim, M., Jung, D. H. et al. 2007, Artificial neuron network models for prediction of intestinal permeability of oligopeptides. BMC Bioinformatics. 8:245.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F. Schwab, M., Antonescus, C. R., Peterson, C., and C. S. Meltzer. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neuron networks. Nat Med. 1:673-679.

Kreiger, M. J. B., Billeter, J. B., and Keller, L. 2000. Ant-like task allocation and recruitment in cooperative robots. Nature. 406:992-995.

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.

Marchini, J., Donnelly, P., and L. R. Cardon. 2005. Genome-wide stregies for detecting multiple loci that influence complex diseases. Nat. Genetics. 37:413-417.

McClurg, P., Pletcher, M. T., Wiltshire, T. and A. I. Su. 2006. Comparative analysis of haplotype association mapping algorithms. BMC Bioinformatics. 7:61.

Mueller, S. G., Weiner, M. W., Thai, L. J., Petersen, R. C., Jack, C. R., Jagust, W., Trojanowski, J. Q., Toga, A. W., and L. Beckett. 2005. Ways towards an early diagnosis in Alzheimer's disease: The Alzheimer's disease neuroimaging initiative (ADNI). Alzheimers Dement. 1(1):55-66.

Meuwissen, T. H. E., Hayes, B. J., and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps.Genetics.157:1819-1829.

Pickrell, J., Clerget-Darpoux, F., and C. Bourgain. 2007. Power of genome-wide association studies in the presence of interacting loci. Genet. Epidemiol. [Epub ahead of print].

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.

Ramsley, S. E. 2006. Unipolar or bipolar ? Improving diagnostic confidence with adult pateint. American Academy of Nurse Practitioners. 19:172-178.

Rekaya, R., Weigel, K. A. and Gianola, D. 2001. Threshold model for misclassified binary responses with applications to animal breeding. Biometrics. 57:1123-11299.

Ressom, H.W., Varghese, R.S., Orvisky, E., Drake, S.K., Hortin, G.L., Abdel-Hamid, M. Loffredo, C.A. and R. Goldman. 2007. Peak selection from MALDI-TOF mass spectra using ant colony optimization. Bioinformatics. 23(5):619-626.

Sapp, R. L., Spangler, M. L., Rekaya, R., and J. K. Bertrand. 2005. A simulation study for the analysis of uncertain binary resposes: Application to first insemination success in beef cattle. Gener. Sel. Evol. 37:615-634.

Sarkis, E. H. 2000. 'Model' behavior. Science. 287:2160-2163.

Schmitt, F. A.,Davis, D. G., Wekstein, D. R., Smith C. D., Ashford, J. W., and W. R. Markesbery. 2000. Preclinical AD revisited: neuropathology of cognitively normal older adults. Neurology. 55:370-376.

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.

Shymygelska, A. and H. H. Hoos. 2005. An ant colony optimization algorithm for the 2D and 3D hydrocarbon polar protein folding program. BMC Bioinformatics. 6:30.

Spangler, M. L., Sapp R. L., Rekaya, R., and J. K. Bertrand. 2006. Success at first

insemination in Australian Angus cattle: Analysis of uncertain binary responses. J. Anim. Sci. 84:20-24.

Stone, R. T., Casas, E., Smith, T. P. L., Keele, J. W., Harhay, G., Bennett, G. L., Koohmaraie, M. ,Wheeler, T. L., Shackelford, S. D., and W. M. Snelling. 2005. Identification of genetic markers for fat deposition and meat tenderness on bovine chromosome 5: Development of a low-density single nucleotide polymorphism map. J. Anim. Sci. 83:2280-2288.

West, M., Blanchette, C., Dressman, H., Huang, E. R., Ishida, S., Spang, R., Zuzan, H., Olson Jr, J. A., Marks, J. R. and J. R. Nevins. 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc. Natl. Acad. Sci. 98:11462-11467.

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.

Woon , P. Y., Kaisaki, P. J., Braganca, J., Bihoreau, M. T., Levy, J. C., Farrall, M., and D. Gauguir. 2007. Aryl hydrocarbon receptor nuclear translocator-like (BMAL1) is associated with susceptibility to hypertension and type 2 diabetes. Proc. Natl. Acad. Sci. 104(36):14412-14417.

Yeh, M., Morley, K. I., and W. D. Hall. 2004. The policy and ethical implications of genetic research on attention deficit hyperactivity disorder. Australian and New Zealand Journal of Psychiatry. 38:10-19

Zhang W., Rekaya, R. and K. Bertrand. 2006. A method for predicting disease subtypes in presence of misclassification among training samples using gene expression: application to human breast cancer. Bioinformatics. 22:317-325.

CHAPTER 3

CLASSIFICATION OF INCIPIENT ALZHEIMER PATIENTS USING GENE EXPRESSION DATA: DEALING WITH POTENTIAL MISDIAGNOSIS[1]

[1] Robbins, K. R., Joseph, S., Zhang, W., R. Rekaya and J. K. Bertrand. 2006. *Online Journal of Bioinformatics.* 7 (1) : 22-31. Reprinted here with permission of publisher

Abstract

A latent-threshold model and misclassification algorithm were implemented to predict the Alzheimer's disease (AD) status of 16 subjects using gene expression data. Each of the 16 subjects was initially classified as healthy or incipient AD using clinical tests. To examine possible age effects on the diagnosis of incipient AD, two datasets were created containing the age unadjusted (D1) and age adjusted (D2) expression of the 100 most informative genes. Control and incipient subjects were clustered into old and young age classes which were then used for age adjustments. Results obtained without invoking the misclassification algorithm showed limited predictive power of the model using either D1 or D2. When the misclassification algorithm was invoked, four subjects were identified as being potentially misdiagnosed. Results obtained after adjustment of the AD status (switching of the binary status) of these four samples showed a significant increase in the model's predictive ability. Further examination of the misdiagnosed samples, using plots and χ^2 tests, showed that the gene expression of these samples agreed more with the new than the initial classification. Similar results were obtained using either D1 or D2. Interestingly, all of the misdiagnosed subjects were originally classified as either an old control or a young incipient. These results suggest that gene expression can be used to improve AD diagnosis by identifying potentially misdiagnosed subjects in the training set. Moreover, it was found that age may have little influence on genes highly correlated with AD status, but it could affect diagnosis based on clinical tests.

Key words: Latent-threshold model, Misclassification algorithm, Alzheimer's disease

Introduction

Alzheimer's disease (AD) is a degenerative disease of nerve cells in the cerebral cortex that leads to atrophy of the brain and senile dementia. Patients with this devastating disorder lose their ability to encode new memories, first of trivial and then of important details of life. AD is the most common form of age-related dementia and one of the most serious health problems in the USA. AD is a major cause of loss of intellectual function in middle-aged and elderly people. The number of individuals with AD is expected to approach 14 million by the year 2050. In 1998, the annual cost for the care of patients with AD in the United States was approximately $40,000 per patient.

While symptoms for AD are well defined and very pronounced in later stages of the disease, diagnosis of patients with incipient AD has proven to be somewhat difficult The Mini-Mental State Exam (MMSE) is one clinical test used to detect the presence of AD, however the effects of normal aging can be confounded with the effects of AD in incipient cases when using cognitive tests such as MMSE (Ashe, 2000; Glasko et al., 1990). Neurofibrillary tangles (NFT) are another common AD indicator used to aid in the confirmation of AD diagnosis; however some recent studies have shown that NFT may be associated with normal aging, and may not be sufficient for accurate diagnosis of AD (Haroutunian et al., 1999; Price and Sisodia, 1998; Schmitt et al., 2000; Snowdon, 1997).

Although indicators of AD, such as MMSE and NFT, are not highly effective in detecting the presence of incipient AD, neuropathological data indicate that the brain changes associated with AD begin well before clinical symptoms are established (Price and Sisodia, 1998). This coupled with evidence that many genes play a role in the development of AD**,** suggest that the

use of genetic information, such as gene expression, could yield more accurate diagnosis of incipient AD (Goate et al., 1991; Scheuner et al., 1996; Tanzi, 1999).

The use of microarray expression profiling for the classification and subtype discovery of diseases has been proposed and frequently investigated; however, such algorithms work under the assumption that all classifications in the training set are correct (Dudoit et al., 2002; Golub et al., 1999; Khan et al., 2001; Yeoh et al., 2002). In cases where the diagnostic tools for classification of training samples are not highly reliable, as with MMSE, this assumption may not hold true. If training samples are in fact misclassified, these algorithms may select genes that have little or no predictive ability. Zhang et al. (2006) found that when misclassification occurred in the training set, the predictive ability of the algorithm used was greatly reduced. In cases where the certainty of training sample classification is in question, a method to identify and correct potential misclassifications may be needed to effectively predict disease status.

For this study a threshold model and misclassification algorithm were implemented, utilizing both age adjusted and age unadjusted gene expression, to identify subjects that were misclassified using clinical diagnostic tools, to predict disease status, and to examine the effects of age on the diagnosis and onset of incipient AD.

Material and Methods

The dataset used for this study contained gene expression measures on mRNA extracted from hippocampus brain tissue, MMSE scores, NFT scores, and age at death for 31 subjects. Methods for mRNA extraction, calculation of the average difference of probe set intensities, and disease classification, based on MMSE and NFT, are described by Blalock et al. (2004). A summary of the data can be found in Table 3.1.

Two datasets were created for use in predicting disease status. The first dataset (D1) contained the natural log of the average difference of probe set intensities (LGE). To examine any effects of aging on the expression of AD related genes, a dataset containing age adjusted LGE was created (D2). However, due to the large difference in the mean ages of healthy and incipient AD subjects (Table 3.1) the modeling of age directly as a covariate was problematic. It should be noted that this confounding did not exist between control and more advanced cases of AD as presented in Table 3.1. To address this problem, subjects in the control and incipient AD classes were placed into young and old age groups. Using these classes, the LGE for each gene was adjusted using the following fixed effect model:

$$LGE_{ijk} = age_i + status_j + age * status_{ij} + e_{ijk} \quad (1)$$

where LGE_{ijklm} is the log gene expression for subject k; age_i is the age class i; $status_j$ is the disease status j; $age*status_{ij}$ is the interaction of age i and disease status j; and e_{ijk} the random residual. Using the estimates from equation (1), the age adjusted LGE were calculated as:

$$LGE^*_{ijk} = LGE_{ijk} - age_i \quad (2)$$

A Bayesian regression model, as developed by West (2001), was used to predict disease status in the form of a probability $p_i(y_i=1)$, with $y_i = 1$ indicating an incipient AD disease status for subject i and $y_i = 0$ indicating a healthy status. The regression on the vector of binary response **y** was done using a latent variable model, with l_i being an unobserved, continuous latent variable relating to binary response y_i such that:

$$y_i = \begin{cases} 1 & \text{if } l_i \geq 0 \\ 0 & \text{if } l_i < 0 \end{cases}$$

The liability l_i was modeled using a linear regression model as:

$$l_i = \mathbf{X}_i\boldsymbol{\beta} + e_i \qquad E(l_i) = \mathbf{X}_i\boldsymbol{\beta} \qquad e_i \sim N(0,1) \tag{3}$$

where $\mathbf{X}_i$ corresponds to row i of the matrix **X,** containing explanatory responses.

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

$$p_i(y_i = 1) = \boldsymbol{\Phi}(\mathbf{X}_i\boldsymbol{\beta}) \text{ and } p_i(y_i = 1) = 1 - \boldsymbol{\Phi}(\mathbf{X}_i\boldsymbol{\beta}) \tag{4}$$

where $\boldsymbol{\Phi}$ is the standard normal distribution function, yielding the following relationships:

$$y_i = \begin{cases} 1 & \text{if } \boldsymbol{\Phi}(\mathbf{X}_i\boldsymbol{\beta}) \geq 0.5 \\ 0 & \text{if } \boldsymbol{\Phi}(\mathbf{X}_i\boldsymbol{\beta}) < 0.5 \end{cases}$$

For this application the explanatory responses contained in the matrix **X** were the LGE of the test subjects. Genes were selected based on differential expression and the correlations between LGE and the binary responses of subjects in the training set and differential expression. The 100 most influential genes were selected and then used to estimate regression coefficients or gene effects in the vector **β**. The process for gene selection was repeated for each replication of the validation procedure.

Due to the fact that 100 genes were used, the dimensions of **X** and **β** were much larger then the number of binary response in **y,** thus creating the need to perform a dimension reduction on **X** and **β**. This was done using singular value decomposition (SVD) such that:

$$\mathbf{X} = \mathbf{QDP}' \tag{6}$$

where **X** is the *n* (number of samples) by *m* (number of genes) matrix of LGE, **Q** is an *n* by *n* orthogonal matrix, **D** is an ordered *n* by *n* diagonal matrix, and **P** is an *m* by *m* orthogonal matrix, where the first *n* columns of **P** are the right-hand singular vectors of **X**.

Using the results of the SVD on **X,** equation (3) can be re-written in matrix-vector notation as:

$$\boldsymbol{l} = \mathbf{QD}\boldsymbol{\gamma} + \boldsymbol{e} \tag{7}$$

where $\boldsymbol{\gamma}$ an *nx1* vector of "super" gene effects. Further,

$$y_i = \begin{cases} 1 & \text{if } \boldsymbol{\Phi}(\mathbf{Q}_i\mathbf{D}\boldsymbol{\gamma}) \geq 0.5 \\ 0 & \text{if } \boldsymbol{\Phi}(\mathbf{Q}_i\mathbf{D}\boldsymbol{\gamma}) < 0.5 \end{cases} \tag{8}$$

where $\boldsymbol{l}$ is a vector of *n* liabilities corresponding to *n* binary responses in the vector **y;** $\mathbf{Q}_i$ is row *i* of the matrix **Q,** and $\boldsymbol{\gamma}$ is equal to $\mathbf{P}'\boldsymbol{\beta}$**.**

The predictive ability of this model was tested using a "leave one out" validation procedure, in which the binary response for one subject was treated as unknown, and regression coefficients were calculated using the remaining *n-1* subjects in the training dataset. The estimated coefficients were then used to predict the disease status for the record being treated as missing. This process was repeated until each of the 16 subjects had been treated as having missing records and subsequently had their disease status predicted.

To examine the possibility of misdiagnosis, the probability if miscoding (PM) was calculated for each sample in the validation data set using a Bayesian approach derived by Rekaya et al. (2001). Let **r** be an unobserved vector of the "true" binary outcomes (incipient=1, control=0) of samples in the training set, which can be viewed as a realization of the vector **y,** and, as such, are bound by equation (7) and classification rule (8). Assume the observed vector of binary outcomes, **z,** is a noisy realization of **r**, where one or more of the binary outcomes are potentially misdiagnosed or misclassified. Misclassification occurs if some elements of **z** are switched from the "true" binary response in **r**, such that z_i does not equal r_i. Let **m** be a vector of indicator variables, where $m_i = 1$ if observation *i* is misclassified and $m_i = 0$ otherwise. Each

element of **m** and **r** can be modeled using a Bernoulli process, such that, assuming independence, their joint distribution, given $\boldsymbol{\gamma}$, π, **F,** and **D**, is:

$$p(\mathbf{m},\mathbf{r} \mid \boldsymbol{\gamma},\pi,\mathbf{F},\mathbf{D}) = \prod_{i=1}^{n} \pi^{m_i}(1-\pi)^{(1-m_i)}[p_i(\mathbf{Q_i D}\boldsymbol{\gamma})]^{r_i}[1-p_i(\mathbf{Q_i D}\boldsymbol{\gamma})]^{(1-r_i)} \quad (9)$$

where π is the probability of a misdiagnosis or misclassification occurring in the training set; and $p_i(\mathbf{F_i D}\boldsymbol{\gamma}) = \Phi(\mathbf{F_i D}\boldsymbol{\gamma})$ is the probability that subject r_i has incipient AD.

Using the relationship between r_i and z_i, given m_i, derived as:

$$r_i = (1-m_i)z_i + m_i(1-z_i) \quad (10)$$

coupled with equation (9), the conditional posterior distribution of m_i is given by:

$$p(m_i = 1 \mid \boldsymbol{\gamma},\boldsymbol{\pi},\mathbf{m}_{-\mathbf{i}},\mathbf{Q},\mathbf{D},\mathbf{r}) = \frac{\pi[p_i(\mathbf{Q_i D}\boldsymbol{\gamma})]^{(1-z_i)}[1-p_i(\mathbf{Q_i D}\boldsymbol{\gamma})]^{z_i}}{K} \quad (11)$$

and

$$p(m_i = 0 \mid \boldsymbol{\gamma},\boldsymbol{\pi},\mathbf{m}_{-\mathbf{i}},\mathbf{Q},\mathbf{D},\mathbf{r}) = \frac{(1-\pi)[p_i(\mathbf{Q_i D}\boldsymbol{\gamma})]^{z_i}[1-p_i(\mathbf{Q_i D}\boldsymbol{\gamma})]^{(1-z_i)}}{K} \quad (12)$$

where $K = \pi[p_i(\mathbf{Q_i D}\boldsymbol{\gamma})]^{(1-z_i)}[1-p_i(\mathbf{Q_i D}\boldsymbol{\gamma})]^{z_i} + (1-\pi)[p_i(\mathbf{Q_i D}\boldsymbol{\gamma})]^{(1-z_i)}[1-p_i(\mathbf{Q_i D}\boldsymbol{\gamma})]^{z_i}$ and PM = $p(m_i = 1 \mid \boldsymbol{\gamma},\boldsymbol{\pi},\mathbf{m}_{-\mathbf{i}},\mathbf{Q},\mathbf{D},\mathbf{r})$.

Assuming a Beta(1,3) as a prior for π, its conditional distribution is easily obtained as:

$$p(\pi \mid \boldsymbol{\gamma},\mathbf{m},\mathbf{Q},\mathbf{D},\mathbf{z}) \sim Beta(1+\sum_{i=1}^{n} m_i, n+3-\sum_{i=1}^{n} m_i) \quad (13)$$

All estimates of model parameters were obtained using a fully Bayesian approach (Rekaya et al., 2001; West et al., 2001; Zhange et al., 2006). Using the binomial approximation the normal distribution z-scores were calculated from the PM for each sample in the training set. The z-score with the highest absolute value (z*) was selected and then compared to the threshold

z_α. It was assumed that the misclassification of subjects in the training set would be a rare event and as such α was set to 0.05. If $z^* \geq z_\alpha$, using a one-sided test, the subject was identified as being potentially misclassified. This subject was then reclassified, and the validation step was repeated using the new classification. This process was repeated until no additional samples were identified as being potentially misclassified.

To examine the accuracy of the reclassification process, a subset of D1 was created that contained only samples that were not identified as being miscoded and were correctly classified in cross validation. This dataset (D3) contained four incipient AD and five healthy subjects. Using D3, the 25 genes most highly correlated with the disease status were selected. The mean LGE, for each gene, was computed for incipient (MI) and control (MC) samples, and an overall mean expression (MO) computed by averaging MI and MC. The LGE of reclassified samples, MI, and MC were deviated from MO and plotted. The number of times the reclassified samples fell on the incipient and control sides of MO, respectively, were summed, discarding LGE with absolute values less than 0.05 when deviated from MO, and used to calculate the $X^2_{(df=1)}$ as follows:

$$\chi^2_{(df=1)} = \frac{(nc - mc)^2}{mc} + \frac{(ni - mc)^2}{mc} \tag{14}$$

Where *nc* is the number of times the LGE fell on the control side of MO; *ni* is the number of times the LGE fell on the incipient side of MO; and *mc* is the average of *nc* and *ni*.

Results and Discussion

The ranges of MMSE and NFT scores, given in Table 3.2, show large overlapping regions between the scores for control and incipient AD groups. Clearly there are no well defined borders separating control and incipient AD, which could make diagnosis difficult when using these clinical tests. This seems to be reflected by the initial validation results found in Table 3.3.

When using either D1 or D2, the model had little power to correctly predict disease status given the original disease classification based on MMSE and NFT scores. In fact, when using D1, 6 out of 16 samples had their disease status wrongly predicted and an additional 2 samples, 6 and 8, could not be classified as their $p(status = 1)$ were equal to 0.5 resulting in an accurate classification probability of 50%. Similar results were observed when D2 (age adjusted expression) was used. These results are in concordance with the conclusions of Glasko et al. (1990) who found that MMSE was not sensitive to the presence of early AD. Furthermore previous microarray studies have found differential expression between healthy and Alzheimer's disease (AD) subjects for several genes, and Ashford and Mortimer (2002) concluded that non-familial AD is mainly a genetic disease (Loring et al. 2001; Ricciareli et al., 2004). This combined with the overlapping of AD indicators used for the classification of subjects in this dataset, suggests that the poor predictability of the model may be due, in large part, to potential misdiagnosis or misclassification in the training data set.

When potential misdiagnosis in the training set was postulated in the statistical model, four samples, subjects 4, 10, 12, and 15, were identified as being misclassified. Table 3.4 shows that, after iteratively reclassifying (switching of their binary status) of these four subjects, there were large increases in the prediction accuracy of the model using both datasets. It can be seen that in addition to the four reclassified subjects, the predictions for subjects 1 and 16 went from being incorrect to correct after reclassification. In fact, only three samples (subjects 8, 7, and 11) had their disease status wrongly classified using either D1 or D2 and an additional sample (subject 6) was not conclusively classified using D1. Based on these results, the consideration of potential misdiagnosis in the statistical model has increased the prediction power of the model from 50 to 75% and from 56 to 81% using D1 and D2, respectively. The predictions for subjects

6 and 8 changed little after reclassification, both remaining near 0.50 when using D1, suggesting, perhaps, that the gene expression measures of these subjects follow an irregular pattern unlike those of healthy and incipient AD subjects. The density plots of the posterior distributions of subjects 8 and 6 (results not shown), unlike all other subjects, were relatively flat, suggesting the LGE contributed little information to the sampling of their liabilities, possibly due to hybridization problems or additional heterogeneity in gene expression not accounted for when using binary classifications.

In order to further investigate, a validation was conducted excluding these subjects. As can be seen in Table 3.5, the exclusion of these subjects had only small effects on the model's predictive ability, suggesting that these subjects had only a minor influence on the gene selection process. With the exception of subject 8, the only subject that was correctly classified prior to recoding and incorrectly classified after recoding was subject 7. It should be noted that when α was relaxed to 0.08 subject 7 was identified as being miscoded, however, when α was set to 0.1 the algorithm deteriorated to a state in which previously recoded samples were continuously switched between the two disease statuses. This yielded results with little interpretive value, and suggested the need for more conservative levels of α to prevent recoding of subjects that were correctly classified using MMSE and NFT scores. Based on these results it is clear that the model performance improved after recoding subjects 15, 4, 12, and 10.

In addition to the recoded samples, χ^2 values were calculated for two subjects strongly believed to be correctly classified as healthy and incipient (AD). These reference samples were removed from D3 when selecting genes for calculation of their respective χ^2 values. The χ^2 values for samples 15, 4, and 12 were highly significant, with the majority of the gene expression values in agreement with the recoded status as seen in Table 3.6. When compared to the χ^2

values of the reference subjects 3 and 2, the LGE of these recoded subjects appear to fit the reclassified status significantly better than their initial status.

Although the reference samples did not have the most significant χ^2 values, as expected, plots of LGE shown in Figure 3.1 and Figure 3.2 show that, on average, the LGE of the reclassified samples were closer to MO then than the LGE of the reference subjects. Despite the fact that the χ^2 value for subject 10 was insignificant, its expression agreed with the recoded status more than with the original classification. While the LGE of subject 10 appear to be somewhat ambiguous when utilizing D3 to select genes, it seems that this subject's LGE agree more with the reclassified status when using all data, as shown by the validation result in Table 3.4. In addition to the LGE, the MMSE and NFT scores were examined for each of the reclassified samples. It was found that all of the recoded samples were within the range of the recoded status for MMSE or NFT, with subject 15 being within the range of both the MMSE and NFT for its recoded status. Such results lend support to the reclassification of these four samples.

Given the small sample size and the confounding of age with disease status, as seen in Table 3.1, it is difficult to construe the actual effect of age on MMSE, NFT, and LGE. One possible explanation for the similar performances of D1 and D2 is that age has little effect on the expression levels of the genes most highly correlated with AD status. This seems to be supported by the fact that, on average, the age class effects were less than 1.5% of the magnitude of the mean expression for selected genes, suggesting that differences in predictions between D1 and D2 could be the result of random noise added by age adjusting the LGE. The fact that all four of the recoded subjects were classified as either an old control or a young incipient could indicate that there may be some type of relationship between age, NFT, and MMSE, possible resulting in a higher rate of misdiagnosis of patients in these categories. Previous studies have shown that

age does affect AD indicators such as cognitive decline and NFT, but that these indicators do not necessarily signal the onset of AD. Chen and Fernandez (2000) concluded age was a risk factor for the onset of AD due to decreases in Ca^{2+} signaling in older individuals. These decreases in signaling led to accumulation of amyloid plaques and NFT, both major indicators of AD. Snowdon (1997), however, found that active individuals were less likely to develop AD despite large amounts of amyliod plaques and NFT. In a study using transgenic mice, Ashe (2000) concluded that a major difficulty in determining the relationship between cognitive decline and molecular markers was an inability to distinguish between age-dependent and age-independent effects.

Another plausible explanation for the similarities between the performances of D1 and D2 is the inability to effectively model age as a covariate resulted in uninformative age adjustments. The age classes used to create D2 may have only partially accounted for the age effect. If this were the case, the small differences observed when using D2 versus D1 could be the result of the partial accounting for the age effect, with the full effect of age being unobserved. Similarly, the pattern of misdiagnosis using clinical data could simply be artifacts of the data resulting from the small sample size. This may be supported by the fact that, if it is expected that MMSE decreases with age and NFT increases, misclassification would most likely occur when a subject was initially classified as a young control or an old incipient. Clearly, given the confounding of age with disease status and the small sample size, further investigation is needed to confirm the effect of age on the diagnosis of incipient AD when using gene expression or clinical tests.

Due to the overlapping of AD indicators between healthy and incipient AD subjects, and the possible effects of age, MMSE and NFT may not be adequate measures for diagnosis of

incipient AD. The use of LGE, coupled with the reclassification algorithm, appeared to greatly increase the accuracy of incipient AD diagnosis. While the use of LGE showed improved performance over traditional AD makers, utilization of such data for disease diagnosis is impractical, as it would require the collection of brain tissue for expression analysis. However, the treatment of AD in its early stages will require a better understanding of the genes involved in the development of incipient AD. Microarray studies provide an excellent tool for accomplishing such a task, but if multiple subjects are misclassified, as it appears was the case with these data, such studies would have little power to detect differential expression. The use of this algorithm as a preprocessing step for analysis of microarray data, in which miscoding is suspected to be present, could increase the power of detecting differentially expressed genes. In fact, preliminary results form mixed model analysis of gene expression detected no AD related genes as differentially expressed when using the original classifications; however, when the statuses of the 4 subjects identified as being misclassified were switched, differential expression of several AD related genes was detected. One such gene associated with amyloid plaques, VILIP-1, was found to down regulated, which is in agreement with findings by Schnurra et al.(2001)

From the results of this study it is clear that MMSE and NFT are inadequate for diagnosis of incipient AD. Conversely it was shown that the misclassification algorithm was capable of identifying miscoded data, and that the utilization of LGE can more accurately predict disease status, possibly allowing for more powerful microarray analysis. Additionally the similar performances of D1 and D2 may suggest that the effects of age are limited to AD markers such as MMSE and NFT, and have little effect on the expression of the genes most highly correlated

with the onset of AD. However, further research with a larger and more balanced dataset is needed to confirm the effects of age on the onset of incipient AD.

Acknowledgments

We would like to thank Dr. Eric M. Blalock for providing us with the data used for this study.

Literature Cited

Ashe, K. H. 2000. Learning and Memory in Transgenic Mice Modeling Alzheimer's Disease. Learning & Memory. 8(6):301-308.

Ashford, J. W. and J. A. Mortimer. 2002. Non-familial Alzheimer's disease is mainly due to genetic factors. Journal of Alzheimer's disease. 4:169-77.

Blalock, E. M., Geddes, J. W., Chen, K. C., Porter, N. M., Markesbery W. R. and P. W. Landfield. 2004. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc. Natl. Acad. Sci. 101: 2173-2178.

Chen, M. and H. L. Fernandez. 2000. How Important are Risk Factors in Alzheimer's Disease. Journal of Alzheimer's disease. 2:97-108.

Dudoit, S., Fridlyyand, J. and T. Speed. 2002. Comparison of discrimination methods for classification of tumors using gene expression data Journal of the American Statistical Association. 97:77-87.

Galasko, D., Klauber, M. R., Hofstetter, C. R., Salmon, D.P., Lasker, B. and L. J. Thal. 1990. The Mini-Mental State Examination in the early diagnosis of Alzheimer's disease. Archives of Neurology. 47:49-52.

Goate, A., Chartier-Harlin, M. C., Mullan, M., Brown, J., Crawford, F. and L. Fidani

1991. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature. 349:704-706.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gassenbeek, M., Mesirov, P., Coller, H., Loh, M. L., Downing, J. R., and M. A. Caligiuri. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 286:531-537.

Haroutunian, V., Purohit, D. P., and D. P. Perl. 1999. Neurofibrillary tangles in nondemented elderly subjects and mild Alzheimer disease. Archives of Neurology. 56(6): 713-718.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescuy, C. R., Peterson, C. and P. S. Meltzer. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 7: 673-679.

Loring, J. F., Wen, X., Lee, J. M., Seilhamer, J. and R. Somogyi R. 2001. A gene expression profile of Alzheimer's disease. DNA and Cell Biology. 20:683-695.

Price, D. L. and S. S. Sisodia. 1998. Mutant genes in familial Alzheimer's disease and transgenic Models. Annual Review of Neuroscience. 21: 479-505.

Rekaya, R., Weigel, K. A. and D. Gianola. 2001. Threshold model for misclassified binary responses with applications to animal breeding. Biometrics. 57:1123-1129.

Ricciarelli, R., d'Abramo, C., Massone, S., Marinari, U., Pronzato, M. and M. Tabaton. 2004. Microarray analysis in Alzheimer's disease and normal aging. International Union of Biochemistry and Molecular Biology Life. 56:349-354.

Scheuner, D., Eckman, C., Jensen, M., Song, X., Citron, M. and N. Suzuki. 1996.

Secreted amyloid beta-protein similar to that in the senile plaques of Alzheimer's disease is increased in vivo by the presenilin 1 and 2 and APP mutations linked to familial Alzheimer's disease. Nat. Med. 2:864-870.

Schmitt, F. A.,Davis, D. G., Wekstein, D. R., Smith C. D., Ashford, J. W., and W. R. Markesbery. 2000. Preclinical" AD revisited: neuropathology of cognitively normal older adults. Neurology. 55: 370-376.

Schurra, I., Berstein, H. G., Reiderer, P. and K. H. Braunewell. 2001. The neuronal Calcium sensor protein VILIP-1 is associated with amyloid plaques and extracellular tangles in Alzheimer's Disease and promotes cell death and tau phosphorylation in vitro: A link between calcium sensors and Alzheimer's disease. Neurobiology of Disease. 8: 900-909.

Snowdon, D A. 1997. Aging and Alzheimer's disease: lessons from the Nun Study. Gerontologist. 37:150-156

Tanzi, R. E. 1999. A genetic dichotomy model for the inheritance of Alzheimer's disease and common age-related disorders. The Journal of Clinical Investigation. 104:1175-1179.

West, M., Blanchette, C., Dressman, H., Huang, E. R., Ishida, S., Spang, R., Zuzan, H., Olson Jr, J. A., Marks, J. R. and J. R. Nevins. 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc. Natl. Acad. Sci. 98:11462-11467.

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

Yeoh, E. J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R.,

Behm, F. G., Raimondi, S. C., Relling, M. V. and A. Patel. 2002. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression Profiling. Cancer Cell. 1:133-143.

Zhang W., Rekaya, R. and J. K. Bertrand. 2006. A method for predicting disease subtypes in presence of misclassification among training samples using gene expression: application to human breast cancer. Bioinformatics. 22:317-325.

Table 3.1. Mean and standard deviation of age, MMSE and NFT scores for healthy and Alzheimer's patient[a]

Status	Mean Age	Age SE	Mean MMSE	MMSE SE	Mean NFT	NFT SE
Control	85.3	2.7	27.7	0.5	2.7	3.1
Incipient AD	91.9	2.3	24.3	3.0	17.5	21.7
Moderate AD	83.4	1.1	16.5	0.6	25.6	3.5
Severe AD	84	4.0	6	1.4	32.7	7.2

[a] AD = Alzheimer's Disease. MMSE = Mini-Mental Status Examination. NFT = Nerofibrilliary tangles. For this study, only control and incipient AD subjects were used for analysis.

Table 3.2. Ranges of MMSE and NFT scores for healthy and incipient Alzheimer's disease patients

AD indicator[a]	MMSE		NFT	
	Lower bound	Upper bound	Lower Bound	Upper Bound
Control	26	30	0	8
Incipient AD	20	29	5.5	65.8

[a]AD = Alzheimer's Disease. MMSE = Mini-Mental Status Examination. NFT = Nerofibrilliary tangles.

Table 3.3. Validation results obtained assuming that the input data was correct (without consideration of the misclassification) using age adjusted and age unadjusted data sets

Dataset		Age unadjusted data			Age adjusted data		
Subject	Original status[a]	P(Status = 1)[b]	90% HDPI[c] L	U	P(status = 1)	90% HDPI L	U
1	1	0.12	0.00	0.58	0.08	0.00	0.40
2	1	0.90	0.48	1.00	0.82	0.25	1.00
3	0	0.14	0.00	0.66	0.15	0.00	0.68
4	1	0.05	0.00	0.52	0.11	0.00	0.55
5	0	0.12	0.00	0.89	0.32	0.00	0.91
6	0	0.50	0.05	1.00	0.22	0.00	0.83
7	0	0.41	0.00	0.96	0.40	0.00	0.95
8	0	0.50	0.00	0.82	0.19	0.00	0.78
9	0	0.44	0.00	0.97	0.20	0.00	0.78
10	1	0.55	0.03	1.00	0.53	0.03	1.00
11	1	0.31	0.02	1.00	0.23	0.00	0.82
12	0	0.83	0.11	1.00	0.81	0.24	1.00
13	1	0.73	0.12	1.00	0.74	0.14	1.00
14	0	0.30	0.03	1.00	0.55	0.03	1.00
15	0	0.93	0.69	1.00	0.93	0.65	1.00
16	1	0.32	0.00	0.92	0.15	0.00	0.67

[a] Alzheimer's disease status based on clinical tests (0 = healthy; 1= incipient AD). [b] probability of an individual being incipient AD. [c] High probability density interval 90% for the classification probability (L = Lower bound, U= Upper bound).

Table 3.4. Validation results after reclassification of subjects identified as potentially miscoded using age adjusted and age unadjusted data sets

Dataset			Age unadjusted data			Age adjusted data		
Subject	Original status[a]	Recoded Status[b]	P(Status = 1) [c]	90% HDPI[d] L	U	P(status = 1)	90% HDPI L	U
1	1	1	0.72	0.11	1.00	0.82	0.28	1.00
2	1	1	0.97	0.97	1.00	0.97	0.99	1.00
3	0	0	0.11	0.00	0.57	0.11	0.00	0.56
4	1	0	0.04	0.00	0.17	0.05	0.00	0.18
5	0	0	0.11	0.00	0.54	0.12	0.00	0.57
6	0	0	0.50	0.03	1.00	0.46	0.00	0.96
7	0	0	0.87	0.39	1.00	0.86	0.36	1.00
8	0	0	0.56	0.04	1.00	0.56	0.04	1.00
9	0	0	0.07	0.00	0.35	0.06	0.00	0.24
10	1	0	0.27	0.00	0.87	0.33	0.00	0.92
11	1	1	0.37	0.00	0.81	0.28	0.00	0.86
12	0	1	0.92	0.61	1.00	0.94	0.67	1.00
13	1	1	0.86	0.37	1.00	0.85	0.33	1.00
14	0	0	0.30	0.00	0.90	0.30	0.00	0.89
15	0	1	0.91	0.57	1.00	0.91	0.55	1.00
16	1	1	0.83	0.27	1.00	0.83	0.26	1.00

[a] Alzheimer's disease status based on clinical tests (0 = healthy; 1= incipient AD). [b]Predicted Alzheimer's disease status (0 = healthy; 1= incipient AD). [c] probability of an individual being incipient AD. [d] High probability density interval 90% for the classification probability (L = Lower bound, U= Upper bound).

Table 3.5. Validation results after reclassification and removal of subjects 6 and 8 using age adjusted and age unadjusted data sets

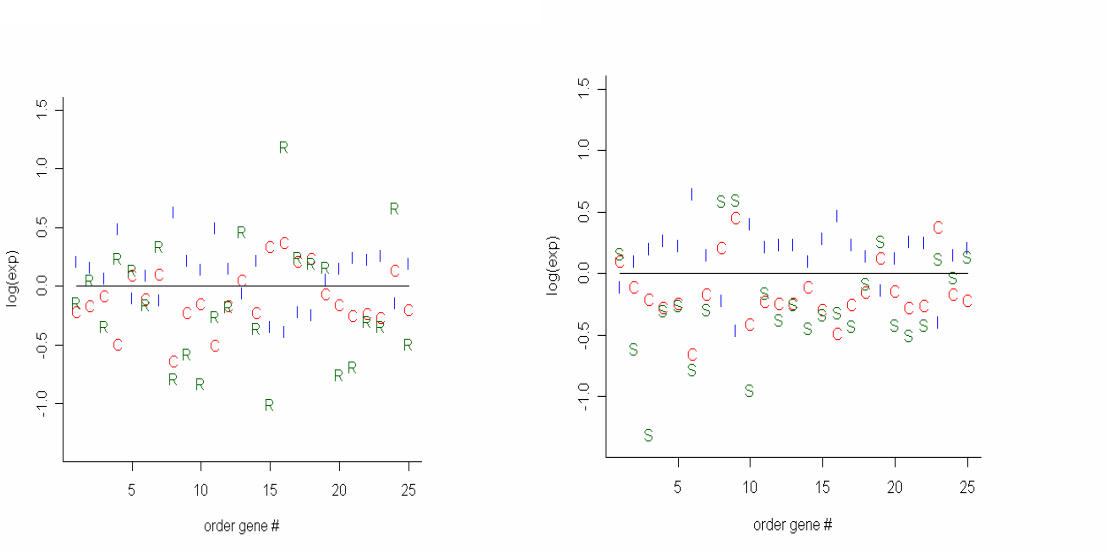
Dataset			Age unadjusted data			Age adjusted data		
Subject	Original status[a]	Recoded Status[b]	P(Status = 1)[c]	90% HDPI[d] L	U	P(status = 1)	90% HDPI L	U
1	1	1	0.82	0.25	1.00	0.80	0.22	1.00
2	1	1	0.94	0.76	1.00	0.93	0.71	1.00
3	0	0	0.12	0.00	0.59	0.12	0.00	0.59
4	1	0	0.05	0.00	0.18	0.05	0.00	0.19
5	0	0	0.11	0.00	0.55	0.12	0.00	0.57
7	0	0	0.84	0.26	1.00	0.85	0.32	1.00
9	0	0	0.10	0.00	0.47	0.06	0.00	0.26
10	1	0	0.38	0.00	0.94	0.37	0.00	0.92
11	1	1	0.37	0.00	0.94	0.27	0.00	0.87
12	0	1	0.92	0.61	1.00	0.91	0.57	1.00
13	1	1	0.78	0.20	1.00	0.77	0.19	1.00
14	0	0	0.41	0.00	0.94	0.37	0.00	0.99
15	0	1	0.83	0.27	1.00	0.83	0.27	1.00
16	1	1	0.79	0.20	1.00	0.80	0.21	1.00

[a] Alzheimer's disease status based on clinical tests (0 = healthy; 1= incipient AD). [b] predicted Alzheimer's disease status (0 = healthy; 1= incipient AD). [c] probability of an individual being incipient AD. [d] High probability density interval 90% for the classification probability (L = Lower bound, U= Upper bound).

Table 3.6. χ^2 values of reclassified subjects[a]

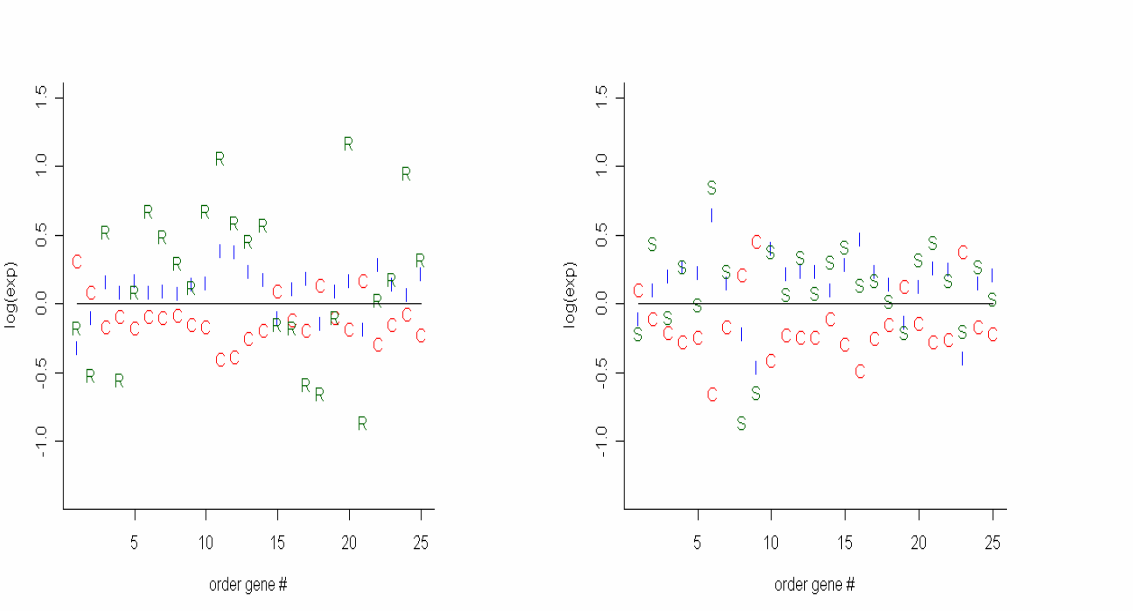
Subject	χ^2 (df = 1)	p-value	#I	#C	Original classification	Recoded classification	MMSE[b]	NFT[c]
4	20.17	<0.001	1	23	I	C	26	6.4
15	18.18	<0.001	21	1	C	I	27	1.3
12	6.76	0.01	19	6	C	I	28	0
10	.17	0.68	11	13	I	C	24	5.5
2	10.67	0.002	20	4	I	__	29	12
3	13.50	<0.001	3	21	C	__	30	8

[a] #I = the number of times the subject's gene expression fell on the incipient side of the overall mean gene expression. #C = the number of times the subject's gene expression fell on the control side of the overall mean gene expression. Subjects 2 and 3 were selected as references. [b] MMSE = the Mini-Mental State Exam. [c] NFT = neurofibrillary tangles.



a. b.

Figure 3.1. Mean expression of the most influential 25 genes for incipient (I) and control samples, plotted with the expression of a reference control sample (R) and the reclassified (from incipient to control) subject 4 (S). All expression levels were deviated from an overall mean: a) Genes were selected and mean expressions were calculated excluding the reference sample, and b) including the reference sample.



a. b.

Figure 3.2. Mean expression of the most influential 25 genes for incipient (I) and control samples, plotted with the expression of a reference incipient sample (R) and the reclassified (from control to incipient) subject 15 (S). All expression levels were deviated from an overall mean: a) Genes were selected and mean expressions were calculated excluding the reference sample, and b) including the reference sample.

CHAPTER 4

THE ANT COLONY ALGORITHM FOR FEATURE SELECTION IN HIGH DIMENSION GENE EXPRESSION DATA FOR DISEASE CLASSIFICATION[1]

[1] Robbins, K. R., W. Zhang, J. K. Bertrand, and R. Rekaya. Submitted to the *Journal of Mathematical Medicine and Biology*

Abstract

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 pre-selection 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 the 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 pre-selection 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. Superiority of the ACA algorithm was validated through simulation.

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).

Ideally, one would like to select an optimal sub-set 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 select 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 sub-set 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 and 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 datasets prior to implementation (Lin et al., 2006; Peng et al., 2006), or have produced relatively low prediction accuracies (Hong and Cho, 2006). 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). 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 though the use of prior information, results in optimal solutions being reached in far fewer iterations than required for GA or SA (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 more feasible with high dimension data sets.

For this study the ACA was implemented using the high-dimensional GCM data-set (Ramaswamy et al., 2001), a colon cancer data set (Alon et al., 1999), and a simulated dataset 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.

Introduction

Classification

Latent variable model: **A** Bayesian regression model was used to predict the 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_{ic} was constructed via a probit model (West, 2003) yielding the following equations:

$$p_{ic}(y_{ic} = 1) = \boldsymbol{\Phi}(\mathbf{X_{ic}}\boldsymbol{\beta}_\mathbf{c}) \text{ and } p_{ic}(y_{ic} = 0) = 1 - \boldsymbol{\Phi}(\mathbf{X_{ic}}\boldsymbol{\beta}_\mathbf{c})$$

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

For data sets containing multiple classes subject i was classified as having tumor 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,…, *nc*, where *nc* is the number of tumor classes in the data set. For binary data sets subject i was classified as having tumor class c if $p_{ic}(y_{ic} = 1) > .5$.

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 over-fitting (West, 2003). When the

number of genes selected was less then the number of subjects in the training data set a spectral decomposition was performed and principle components corresponding to eigenvalues close to zero were removed to avoid computational issues.

Gene Selection

Filter and wrapper based methods were used to select features to form classifiers for each tumor 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 tumor 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.

Fold change: The fold change was computed as:

$$fc_{mc} = \left| M_c - M_r \right|$$

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

t-statistic: The t-statistic was calculated as:

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

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

Penalized t-statistic: The penalized t-statistic was calculated as:

$$pt_{mc} = \frac{\left|M_c - M_r\right|}{a + Sp\sqrt{1/n_c + 1/n_r}}$$

where M_t is the mean of the log base 2 of the gene expression of the tumor class of interest c; M_r is the mean of the log base 2 of the gene expression of the remaining tumor 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 tumor class of interest; and n_r is the number of subjects in the remaining tumor types.

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 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:

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

where $\mathbf{f}_c$ is a vector of all fold change values for tumor class c; $\mathbf{t}_c$ is a vector of all t-statistic values for tumor class c; and $\mathbf{pt}_c$ is a vector of all penalized t-statistic values for tumor 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 the 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 equation (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:

$$\tau_m(t+1) = (1-\rho) * \tau_m(t) + \Delta\tau_m(t) \qquad (2)$$

where ρ is a constant between 0 and 1 that represents the rate at which the pheromone trail evaporates; $\Delta\tau_m(t)$ 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:

4) Each ant selects a predetermined number of genes.

5) Training data is randomly split into two subsets for training (TDS) and validation (VDS) containing ¾ and ¼ of the data respectively (none of the original validation data is used at any point in the ACA).

6) Using the spectral decomposition of TDS, principle components are computed to alleviate effects of collinearity and selected for TDS and VDS by removing components with corresponding eigenvalues close to zero.

7) Using TDS, a latent variable model is trained for each tumor class, and $p_{ic}(y_{ic}=1)$ is predicted for every tumor class c for each sample i in VDS.

8) The accuracy for each tumor class c is calculated as:

$$acc_c = \frac{\sum_{i=1}^{nc} \mathbf{\Phi}(\mathbf{P_{ic}}\boldsymbol{\beta}_\mathbf{c}) / nc + \sum_{i=1}^{nr} [1 - \mathbf{\Phi}(\mathbf{P_{ic}}\boldsymbol{\beta}_\mathbf{c})] / nr}{2} \qquad (3)$$

where $\mathbf{P_{ic}}$ contains principle component values for sample i for tumor class c; $\boldsymbol{\beta}_\mathbf{c}$ is a vector of coefficients estimated using TDS; *nc* is the number of samples in VDS having tumor class *c*; and *nr* is the remaining number of samples in VDS.

9) The change in pheromone for each tumor class is calculated as:

$$\Delta\tau_{mc}(t) = acc_c^{(1-acc_c)}$$

where acc_c is the accuracy for tumor type c as calculated using equation (3). This equation was chosen based on performance using real and simulated data.

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 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 in the level of pheromone trail 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 data set was processed according to Ramaswamy et al. (2001) and contained the intensity values of 16063 probes generate 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.

Colon Cancer data set

The data set contained 62 samples collected from normal colon tissue and colon cancer tumors (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.

Simulated data set

A real data based simulation was conducted using 3 of the 14 tumors types in the GCM data set**.** The tumor types were selected based on the relatively large number of tissue samples in each tumor 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 1 of the 3 tumor 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_{ij} 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:

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

where v_j is equal to is 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 tumor type such that:

$$E(T)=3$$

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

Results

The GCM data set has been a benchmark to compare the performance of classification and feature selection algorithms. Table 4.1 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), MAMA (Antonov et al., 2004), GA/SVM (Liu et al., 2005), and SVM/RFE

(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 (LOOCV). 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 TT/LVM using the GCM split can be found in Tables 4.2-4.5. These tables show that the ACA/LVM performs as good as or better than the rank based methods for every tumor type. Additionally the ACA/LVM correctly predicted 50% of the BR samples, a tumor 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 Alon et al. (1999), and a simulated data set. The best classification accuracies obtained for each algorithm can be found in Table 4.6. Though the variance of prediction accuracy was high across replicates for all methods; the ACA/LVM algorithm yielded the best, or equaled 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 Jefferey et al. 2006.

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 4.7 shows the number of genes needed for each tumor type to achieve the best results, averaged across all

replicates. It can be seen that, for 13 of the 18 tumor 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 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 tumor type when using the ACA. Plots illustrating the selection of features for 4 tumor classes can be found in Figure 4.1.

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 (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 4.2 where red (yellow) color indicates low (high) correlations. The BR and CNS tumor 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.

Discussion

The performance of the ACA/LVM model was superior, not only to the filter based methods used in this study, but 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 3 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 endeavor.

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 and Cho, 2006; Lin et al., 2006; Liu et al., 2005; Ooi and Tan et al., 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 pre-selected 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 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 in many of the tumor 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 mechanism underlying the tumors 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 tumor 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 tumors, the ACA identified two genes, capable of classifying LU tumor 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 tumor class, with which the ACA was able to select a small number of highly predictive genes, was CNS. As with the LU tumor 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 five. APCL protein is a homologue of APC, a known tumor suppressor that interacts with microtubules during mitosis (Akiyama and 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 tumor 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 tumor 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 BR and CNS tumor types, the gene lists for BR tumors were highly variable, suggesting potentially high heterogeneity in these tumor 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.

Conclusions

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

Literature Cited

Akiyama, T. and Y. Kawasaki. 2006. Wnt signaling and the actin cytoskeleton.

Oncogene. 25:7538-7544.

Albrecht, A., Vinterbo,S.A. and L. O. Machado. 2003. An epicurean learning approach to gene-expression data classification. Artif. Intell in Medicine. 28:75-87.

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

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.

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

Dorigio, M. and L. M. Gambardella. 1997. Ant colonies for the travailing 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. and E. S. Lander. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 286:531-537.

Hong, J. and S. Cho. 2006. Efficient huge-scale feature with speciated genetic algorithm. Pattern Recognition Lett. 27:143-150.

Jefferey, I.B., Higgins, D.G. and A. Culhane. 2006. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics. 7.

Li, J. and L. Wong. 2006. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. Bioinformatics. 15:725-734.

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.

Liu, J.J., Cutler, G., Li, W., Pan, Z., Peng, S., Hoey, T., Chen, L. and X. B. Ling. 2005. Multiclass cancer classification and biomarker discover 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. and M. Inagaki. 2003. Filament formation of MSF-A, a mammalan Septin, in human mammary epithelial cells depends on 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.

Peng, S., Xu, Q., Ling, X.B., Peng, X., Du, W. and L. Chen 2003. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. FEBS Letters. 555:358-362.

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.

Scott, M., McCluggage, W.G., Hillan, K.J., Hall, P.A. and S. E. H. Russell. 2006. Altered patterns of transcription of th septin gene, SEPT9, in ovarian tumorgenesis. Int. J. Cancer. 118:1325-1329.

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.

Subramani, P., Sahu, R. and S. Verma. 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 Statistics. 7:723-732.

Yeang, C.H., Ramaswamy, S., Tomaya, P., Mukherjee, S., Rifkin, R.M., Angelo, M., Reich, M., Lander, E., Mesirov, J. and T. Golub. 2001. Molecular classification of multiple tumor types. Bioinformatics. 1:1-7.

Table 4.1. Accuracy (%) of tumor class predictions using ant colony algorithm (ACA) and several previously published methods.

	GCM data set		
	GCM split[a]	Replicated splits	LOOCV[b]
ACA/LVM(14525[c])	90.7	84.8	____
FC/LVM(14525)	79.6	74.8	____
T/LVM(14525)	63.0	____	____
PT/LVM(14525)	77.8	74.4	____
AVG[d]/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	____	____

[a]Split used by Ramaswamy et al 2001; [b]Leave one out cross validation; [c]Number of genes selected prior to the implementation of feature selection algorithm; [d]Weighted average of scaled fold change (FC), t-test (T), and penalized t-test (PT) values.

Table 4.2. Confusion matrix for predictions obtained for the GCM data set using genes selected by the ant colony algorithm.

True\Predicted	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

Table 4.3. Confusion matrix for best predictions obtained for the GCM data set using genes selected by the fold change (50 genes)

True\Predicted	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 4.4. Confusion matrix for best predictions obtained for the GCM data set using genes selected by the penalized t-test (10 genes)

True\Predicted	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

Table 4.5. Confusion matrix for predictions obtained for the GCM data set using genes selected by the t-test (10).

True\Predicted	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 4.6. Classification accuracies using several feature selection methods

Replication	1	2	3	4	5	Over all
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[a]/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
TT/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

[a]Weighted average of scaled fold change (FC), t-test (PT), and penalized t-test values (T).

Table 4.7. Number of genes selected for each tumor 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

[a]Weighted average of scaled fold change (FC), t-test, and penalized t-test (PT) values.

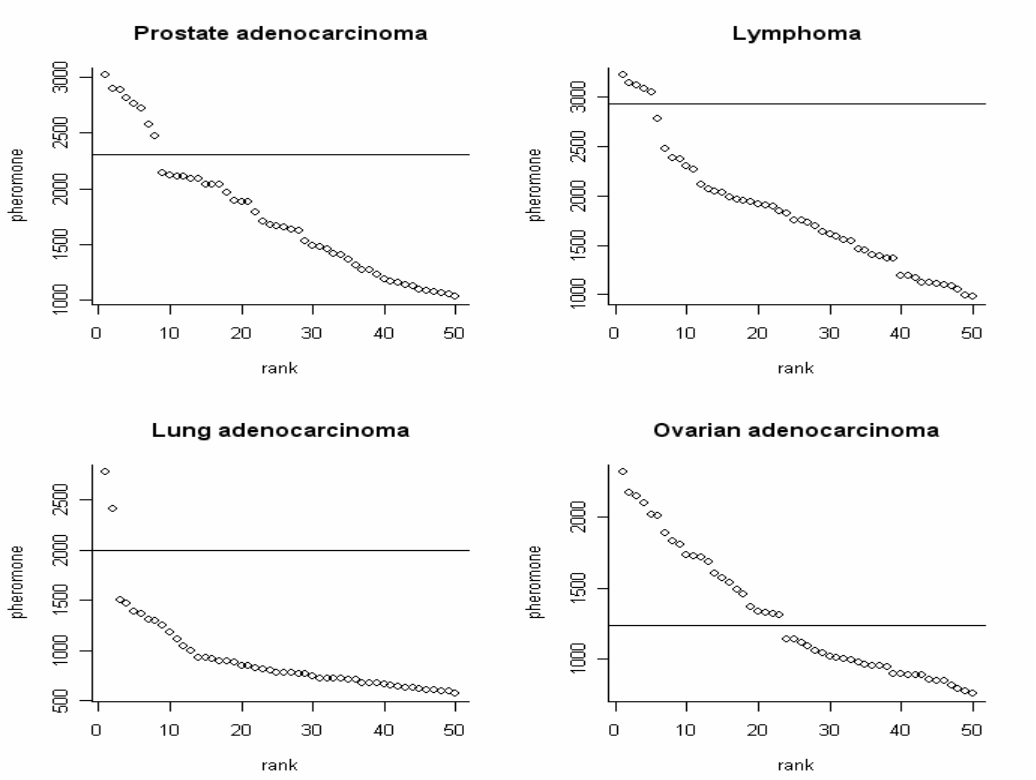


Figure 4.1. Plots illustrating the selection of features (genes) using pheromone levels for four tumor types. Genes on top of the breakup point in the pheromone function (horizontal line) were selected.

a

b

c

Figure 4.2. Heat matrices between the top 30 genes selected for CNS (column 1) and breast adenocarcinoma (column 2) tumors based on: a. fold change; b. penalized t-test; and c. ant colony algorithm (red = low correlation, yellow = high correlation).

CHAPTER 5

THE USE OF THE ANT COLONY ALGORITHM FOR THE DETECTION OF MARKER ASSOCIATIONS IN THE PRESENCE OF GENE INTERACTIONS[1]

[1] Roobbins, K. R., J. K. Bertrand, and R. Rekaya. To be submitted to the *Journal of Genetic Epidemeology*

Abstract

In recent years there has been much focus on the use of single nucleotide polymorphism (SNP) fine genome mapping to identify causative mutations for traits of interest; however many studies focus only on the marginal effects of markers, ignoring potential gene interactions. Simulation studies have show that this approach may not be powerful enough to detect important loci when gene interactions are present. While several studies have examined potential gene interaction, they tend to focus on a small number of SNP markers. Given the prohibitive computation cost of modeling interactions in studies involving a large number SNP, methods need to be develop that can account for potential gene interactions in a computationally efficient manner. This study adopts a machine learning approach by adapting the ant colony optimization algorithm (ACA), coupled with logistic regression on haplotypes, for association studies involving large numbers of SNP markers. The proposed method is compared to a sliding window approach (SW) for the formation haplotypes and a publicly available software package for genotype association (RG). Each algorithm was evaluated using a binary trait simulated using an epistatic model and HapMap ENCODE genotype data. Results show that the ACA outperforms SW and RG under several simulation scenarios, with p-values obtained using ACA being more significant for SNP in high linkage disequilibrium with causative mutations, resulting in substantial increases in power.

Introduction

In recent years there has been much focus on the use of single nucleotide polymorphism (SNP) fine genome mapping to identify causative mutations for traits of interest, and while putative mutations have been identified for several traits, these studies tend to focus on SNP with large marginal effects (Hugot et al., 2001; Woon et al., 2007). However, several studies have

found that gene interactions may play important roles in many complex traits (Coutinho et al., 2007; Barendse et al., 2007). Unfortunately, due to the high density of SNP maker maps, it is computationally infeasible to examine all possible interactions in genome-wide SNP association studies. As a result many studies examining gene interactions focus on a small number of SNP, previously identified as having strong marginal associations.

While this approach has shown some success, simulation studies conducted by Marchini et al. (2005) and Pickrell et al. (2007) showed that, in the presence of several types of gene interactions, there was reduced power to detect causative loci with models estimating only marginal effects. Although Marchini et al. (2005) implemented methods capable of modeling potential gene interactions; due to high computational cost they would require truncation of a large portion of marker data for genome-wide studies considering multiple interacting loci. As such there is a need for methodologies capable of identifying important genomic regions in the presence of potential gene interactions in studies examining large numbers of SNP.

Given the examination of all possible SNP interactions is computationally infeasible with large SNP marker maps, an alternative approach must be considered. One such approach would be to view the identification of groups of interacting SNP as an optimization problem, for which several algorithms have been developed. These algorithms are designed to search large sample spaces for globally optimal solutions and have been applied to wide range of problems from protein folding prediction to interactions of robots on assembly lines (Shymygelska and Hoos et al., 2005; Kreiger et al., 2000; Ding et al, 2005). Through the selection and evaluation of SNP from different regions of the genome using efficient searching methods, optimization algorithms should be able to explore potential gene interactions. Kooperberg et al. (2006) utilized an optimization algorithm, referred to as simulated annealing (SA), to examine interaction effects;

however, only 32 SNP were considered in the model selection process. For studies involving hundreds or even thousands of SNP, efficient algorithms are needed to search the sample space for optimal solutions.

One such algorithm, the ant colony algorithm (ACA), has been shown to be efficient in high-dimension data sets (Robbins et al. 2007). The ACA, developed by Dorigio and Gambardella (1997), is based on the mechanism by which ant colonies find the shortest route to a food source. Ants communicate through a chemical pheromone trail, deposited as they transverse a given path. Ants that choose a shorter path will transverse the distance at a faster rate, thus depositing more pheromone in the process. As the pheromone builds, ants will begin to preferentially choose the shorter path leading to a positive feed back system. Dorigio and Gambardella (1997) showed that the communication between ants had a synergistic effect allowing the ACA to reach optimal solutions in fewer iterations than other optimization algorithms. In the case of SNP association studies, the 'path' is represented by a haplotype formed from a selected subset of SNP, and performance is evaluated based on the fit of a logistic regression for binary traits.

For this study a modified ACA, enabling the use of permutation testing for global significance, was combined with a logistic regression (ACA/LR) and implemented on a simulated binary trait under the influence of interacting genes. The SNP data used for simulations were generated by the HapMap ENCODE project. The performance of the ACA/LR was evaluated and compared to models accounting for only marginal effects.

Materials and Methods

Logistic regression: Haplotype effects were estimated as log odds ratios (*lor*) using logistic regression (LR). The relationship between the *lor* and the binary response can be expressed as:

$$y_i = \begin{cases} 1 & \text{if } lor_i \geq 0 \\ 0 & \text{if } lor_i < 0 \end{cases}$$

The log odds ratio lor_i is modeled as:

$$lor_i = \ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{X}_i\boldsymbol{\beta} + e_i \qquad (1)$$

where P_i = probability ($y_i = 1$) and $\mathbf{X}$ is a matrix containing indicator variables for the haplotypes formed from the selected SNP. Haplotypes with less than two corresponding observations were discarded, and analysis was conducting on all remaining haplotypes.

The link function of the log odds ratio $\mathbf{X}_i\boldsymbol{\beta}$ with the binary response y_i gives the following equations**:**

$$p_i(y_i = 0) = \frac{1}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})} \text{ and } p_i(y_i = 1) = \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})} \qquad (2)$$

yielding the following relationships:

$$y_i = \begin{cases} 1 & \text{if } \dfrac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})} \geq 0.5 \\ 0 & \text{if } \dfrac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})} < 0.5 \end{cases}$$

Marginal effects model: For the model estimating marginal effects, both genotype and haplotype models were implemented. The genotype association method was implemented as an R function developed by Gonzalez et al., 2007. The haplotype approach was implemented using

a sliding window. This approach utilizes a window of *k* SNP in width which slides across the genome *h* SNP at a time. For each window, haplotypes were formed and effects were estimated using the logistic regression model. Using the link function in (2) the accuracies for the selected SNP were calculated as:

$$accuracy = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})} + \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{1}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})} \quad (3)$$

where n_1 represents the number of individual have a binary status of 1; and n_2 represents the number of individuals having a binary status of 0. To estimate individual SNP accuracies, the average of the accuracies for all haplotypes containing a given SNP were computed.

Ant colony algorithm: The ACA employs artificial ants that communicate through a probability density function (PDF) that is updated each iteration with weights or "pheromone levels", which are analogous to the chemical pheromones used by real ants. In the case of SNP association studies, the weights can be determined by the strength of the association between selected haplotypes and the trait of interest. Using the notation of Dorigio and Gambardella (1997) and Ressom et al. (2006), the probability of sampling SNP *m* at time *t* is defined as:

$$P_m(t) = \frac{(\tau_m(t))^{\alpha}\eta_m^{\beta}}{\sum_{m=1}^{nf}(\tau_m(t))^{\alpha}\eta_m^{\beta}} \quad (4)$$

where $\tau_m(t)$ is the amount of pheromone for SNP *m* at time *t*; η_m is some form of prior information on the expected performance of SNP *m*; α and β are parameters determining the weight given to pheromone deposited by ants and a priori information on the features, respectively. For the current study the prior information was the average accuracy of SNP *m*, as determined using the sliding window approach previously described.

The ACA was initialized with all SNP having an equal baseline level of pheromone used to compute $P_m(0)$ for all features. Using the PDF as defined in equation (4), each of *j* artificial ants will select a subset S_k of *n* SNP from the sample space S containing all SNP. Given the relationship between adjacent SNP, ants were allowed to randomly change SNP selections following a multinomial distribution. Changes in SNP selection are limited to the three adjacent SNP on either side of the originally selected SNP marker. The pheromone level of each feature *m* in S_k is then updated according to the performance of S_k as:

$$\tau_m(t+1) = (1-\rho) * \tau_m(t) + \Delta\tau_m(t) \qquad (5)$$

where ρ is a constant between 0 and 1 representing the rate at which the pheromone trail evaporates; $\Delta\tau_m(t)$ is the average change in pheromone level for feature *m* based on the performance of all S_k containing SNP *m*, and is set to zero if feature *m* was not selected by any of the artificial ants. The use of the average pheromone, rather than the sum of the pheromone for each S_k containing SNP *m*, enables the pheromone deposited on each SNP to reach an equilibrium relative to the performance of SNP *m*, provided ρ is less than 1. This equilibrium can be exploited to estimate significance through permutation testing.

The procedure can be summarized in the following steps:

1) Each ant selects a predetermined number of SNP.
2) Using selected SNP, haplotypes are formed and used for logistic regression.
3) The pheromone for each selected group of SNP *k* is calculated as:

$$pheromone_k = acc^{(1-acc)} \qquad (6)$$

where acc is the accuracy for calculated using equation (3).

4) The change in pheromone at time t ($\Delta\tau_m(t)$) is then calculated as the average of pheromone levels feature SNP m using equation (6).
5) Following the update of pheromone levels according to equation (5), the PDF is updated according to equation (4) and the process is repeated until pheromone levels have converged.

Data simulations: Genotype data on 90 unrelated individuals from the Japanese and Han Chinese data set were downloaded from the HapMap ECODE project website. Two 500 Kbp regions on chromosome 2, comprising 2047 polymorphic SNP, were used for simulations. All SNP haplotypes were assumed to be known with out error. The binary disease trait was simulated under a two locus epistatic model as seen in Table 5.1. The loci of the causative mutations were selected at random; with the frequencies of the causative mutations being .58 and .6. Although these frequencies might be considered high, it was necessary to restrict selection to SNP with mutant allele frequencies greater than .5. This was done to insure a reasonable simulated disease incidence of 15%.

Permutation testing was used to access global significance for all models used in the study. Statuses were random shuffled amongst subjects, with haplotype effects and genotype association p-values re-estimated for each new configuration of the response variables. The largest estimated haplotype effect or the smallest genotype association p-value from each permutation was saved to form an empirical distribution used for calculation of p-values. One hundred permutations were performed, yielding p-values accurate to 1%. Power was calculated as the proportion of times a given method identified at least one SNP marker in high LD ($r^2 \geq .90$) with a causative mutation.

Results

The p-values of the causative mutations, as well as, several SNP in high linkage disequilibrium (LD) with causative mutations can be found in Table 5.1. For each replicate the best results, using either two or three SNP haplotypes, are shown. In scenario 1, having the strongest epistatic effect, the ACA/LR detected highly significant associations in both causative regions of the chromosome in two of the three replications. In contrast, the other methods identified causative mutations in both regions in only one replicate, with SW/LR never identifying SNP in both regions. In scenario 2, where the epistatic effect was decrease by 50%, the ability to detect SNP in high LD with the causative mutations was significantly reduced for all methods, with the SW/LR unable to detect any of the associated SNP as being significant, and the RG method only detecting associated SNP in one replication. In contrast, the ACA/LR, despite showing reduced power, was able to detect a significant association in the region around SNP rs2049736 in two replications, and was able to detect a significant effect around both causative mutations in replication 1. Furthermore, p-values obtained from ACA/LR tended to be more significant in each of the two scenarios.

Estimates of power for the three methods can be found in Table 5.2. For SW/LR it can be seen that there is moderate to no power to detect SNP in high LD with causative mutations for scenario 1 and scenario 2, respectively. Though RG showed increased power in scenario 2 its performance was no better than that of SW/LR in scenario 1. Although the ACA/LR suffered from a reduction in power under scenario 2, the methodology had the highest power for both scenarios, yielding a 113% increase in overall power over the next best method, RG. To determine the rate of false discoveries obtained from ACA/LR, the cumulative distribution, based on LD with causative mutations, of SNP identified as being significantly associated with

simulated trait were plotted in Figure 5.1. The cumulative distribution shows the majority of identified SNP having LD between .35 and 1.00, with approximately 7.6% of SNP having LD below .35.

A plot illustrating the LD of all SNP with the two causative mutations can be found in Figure 2. The plot shows a large peak of high LD with SNP rs2049736, while the peak of high LD with SNP rs28953468 is substantially more narrow, and is preceded by a plateau of SNP in moderate LD. Plots of the associative effects each marker using the three methodologies are shown in Figure 3-4. The association plots obtained using ACA/LR show that the pheromone levels for each marker correspond well to the LD plot. Interestingly, in scenario 1 the ACA/LR was able to detect a significant association with many of the markers in moderate LD with rs28953468. Conversely, the plots of association obtained using RG and SW/LR tended to contain more noise.

Discussion

The substantial increase in power, observed when using ACA/LR, demonstrates the effectiveness of the ACA in accounting for epistasis. This is particularly evident in Figure 5.3, as the ACA/LR was able to detect SNP in moderate LD (.35-.45) with the causative mutation rs28953468. By jointly evaluating genotypes in different regions of the genome, the ACA was able evaluate the interacting loci and account for interaction effect through the pheromone function. As the algorithm burned in, the interacting SNP began to be selected together more frequently, further increasing the contribution of the epistatic effects to the pheromone deposited by the artificial ants. This positive feedback allowed the ACA to efficiently identify SNP in high LD with the causative mutations that had no significant marginal effects.

When considering SNP having LD with causative mutations less than .35 as false positives, a threshold chosen based on the LD plot in Figure 5.2, the ACA/LR had a false positive rate of 7.4%, as compared to 2.6% for RG (results not shown). The relatively high error rates were somewhat surprising given the strict control placed on family-wise error, with false positives present in 16.7% of the replications for both ACA/LR and RG. One possible explanation could be the small number of permutations conducted, as 100 permutations yield p-values accurate to only one tenth. Generally 500 permutations would be considered adequate; however given the high number of replicates conducted in this study, 500 permutations would be too computationally costly. Regardless of the cause, a false positive rate of 7.4% would generally be considered acceptable in high-dimension association studies for which the goal was to identify a small subset of markers for further evaluation (Benjamini and Yekutieli, 2005), especially when considering the increase in power obtained when using the ACA/LR.

Since association studies involving large numbers of SNP are generally exploratory in nature, the number of potentially interacting SNP would be unknown. This would necessitate the ACA/LR be robust relative to the number of SNP used to form haplotypes and the number of SNP interacting in the simulated model. In this regard, the number of SNP used by the ACA/LR, relative to the number of SNP interacting in the simulated model, appeared to have no effect in scenario 1, as the power of the methodology do not change when selecting different numbers of SNP. In scenario 2, the power increased when selecting more SNP than were causative of the trait, demonstrating that the number of SNP selected by the algorithm need not correspond to the number of causative loci for the model to perform optimally. Although these results suggest some level of robustness, several runs of the algorithm, selecting various numbers of SNP for haplotype formation, might best insure that optimal results are reached.

In the current study the estimated log odds ratios obtained using ACA/LR and SW/LR were the sum of the individual's two haplotypes, this corresponded closely to codominant model for genotype association. As such a codominant model was used for the RG method, though a dominance model would fit the data better given the recessive nature of the causative mutations. This was done to insure an unbiased comparison amongst the tested methodologies, since the purpose of the study was examine the consequences of ignoring potential gene interactions, and not to compare genotype and haplotype methodologies. That being said and given the different modes of inheritance for causative mutations, the ACA may demonstrate better performance if genotype analysis were adopted. Furthermore, given that a genome wide scan would have markers on multiple chromosomes; a haplotype approach would make little sense.

Conclusions

In the presence of simulated epistasis, models accounting only for marginal effects have little power to detect SNP in strong linkage disequilibrium with causative mutations. In contrast, the proposed optimization methodology obtained substantial increases in power, demonstrating the effectiveness of machine learning approaches for the analysis of marker association studies in which gene interactions may be present. Though the false positive rate was higher than expected, it was not substantially higher than the rates obtained using the genotype association methodology and was within generally accepted levels. Clearly for high-dimension association studies, methodologies capable of efficiently modeling gene interactions, such as the model proposed in this study, should yield superior performance in terms of the power to detect SNP in linkage disequilibrium with causative mutations.

Literature Cited

Hugot, J. P., Chamaillard, M., Zouali, H. et al. 2001. Association of NOD2 leucine-rich

repeat variants with susceptibility to Crohn's disease. Nature. 411:599-603.

Barendse, W., Harrison, B. E., Hawken, R. J., Ferguson, D. M., Thompson, J. M., Thomas, M. B., and R. J. Bunch. 2007. Epistasis between Calpain 1 and its inhibitor Calpastatin within breeds of cattle. Genetics 176:2601-2610.

Benjamini, Y. and D. Yekeli. 2005. Quantitative Trait Loci Analysis Using the False Discovery Rate. Genetics. 171:783-790.

Coutinho, A. M., Sousa, I., Martins, M. et al. 2007. Evidence for epistasis between SLC6A4 and ITGB3 in autism etiology and in the determination of platelet serotonin levels. Hum. Genet. 121:243-256.

Ding, Y. P., Wu, Q. S., and Q. D. Su. 2005. Multivariate Calibration Analysis for metal porphyrin mixtures by an ant colony algorithm. Analytical Sciences. 21:327-330.

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

Gonzalez, J. R., Armengol, L., Sole, X., Guino, E., Mercader, J. M., Estivill, X., and V. Moreno. 2007. SNPassoc: an R package to perform whole genome association studies. Bioinformatics. 23(5):644-645

Kooperburge, C., Bis, J. C., Marciante, K. D., Heckbert, S. R., Lumley, T., and B. M. Psaty. 2006. Logic regression for analysis of the association between genetic variation in the rennin-angiotensin system and myocardial infarction or stroke. Am. J. Epidemiol. 165:334-343.

Kreiger, M. J. B., Billeter, J. B., and Keller, L. 2000. Ant-like task allocation and recruitment in cooperative robots. Nature. 406:992-995.

Marchini, J., Donnelly, P., and L. R. Cardon. 2005. Genome-wide stregies for detecting

multiple loci that influence complex diseases. Nat. Genetics. 37:413-417.

Pickrell, J., Clerget-Darpoux, F., and C. Bourgain. 2007. Power of genome-wide association studies in the presence of interacting loci. Genet. Epidemiol. [Epub ahead of print].

Ressom, H.W., Varghese, R.S., Orvisky, E., Drake, S.K., Hortin, G.L., Abdel-Hamid, M. Loffredo, C.A. and R. Goldman. 2007. Peak selection from MALDI-TOF mass spectra using ant colony optimization. Bioinformatics. 23(5):619-626.

Robbins, K. R., Zhang, W., R. Rekaya, and J. K. Bertrand. 2007. Ant colony optimization for feature selection in high dimensionality data sets. J. Math Med Biol. (Submitted).

Shymygelska, A. and H. H. Hoos. 2005. An ant colony optimization algorithm for the 2D and 3D hydrocarbon polar protein folding program. BMC Bioinformatics. 6:30.

Woon , P. Y., Kaisaki, P. J., Braganca, J., Bihoreau, M. T., Levy, J. C., Farrall, M., and D. Gauguir. 2007. Aryl hydrocarbon receptor nuclear translocator-like (BMAL1) is associated with susceptibility to hypertension and type 2 diabetes. Proc. Natl. Acad. Sci. 104(36):14412-14417.

Table 5.1. Relative risk for simulated trait[a].

	Scenario 1				Scenario 2			
	ab	aB	Ab	AB	ab	aB	Ab	AB
Ab	1	1	1	1	1	1	1	1
aB	1	1	1	1	1	1	1	1
Ab	1	1	1	1	1	1	1	1
AB	1	1	1	15	1	1	1	10

[a] Risks are relative to the aa/bb genotype.

Table 5.2. Permutation p-values for SNP under simulated epistasis[a].

Scenario 1										
SNP	LD	RG	SW/LR	ACA/LR	RG	SW/LR	ACA/LR	RG	SW/LR	ACA/LR
			Rep 1			Rep 2			Rep 3	
rs2049736(409)	1.00	0.10	0.35	0.04*	0.09	0.06	0.01**	0.27	0.24	0.04*
rs7569458(349)	0.955	0.01**	0.13	0.03*	0.11	0.09	0.01*	0.42	0.32	0.07
rs28953468(2041)	1.00	0.03*	0.05	0.01**	0.06	0.14	0.01**	0.19	0.98	0.38
rs10490014(2040)	0.977	0.01**	0.05	0.01**	0.07	0.13	0.01**	0.12	0.98	0.45
Scenario 2										
SNP	LD	RG	SW/LR	ACA/LR	RG	SW/LR	ACA/LR	RG	SW/LR	ACA/LR
			Rep1			Rep 2			Rep 3	
rs2049736(409)	1.00	0.16	0.35	0.11	0.20	1.00	1.00	0.33	0.17	0.04*
rs7569458(349)	0.955	0.03*	0.13	0.03*	0.34	1.00	1.00	0.36	0.18	0.09
rs28953468(2041)	1.00	0.95	0.50	0.09	1.00	0.65	0.22	1.00	1.00	0.39
rs10490014(2040)	0.977	0.50	0.46	0.05*	0.81	0.99	0.56	1.00	1.00	0.34

[a] * indicates significance at α=.05; and ** indicates significance at α=.01

Table 5.3. Power calculations for ACA/LR and SW/LR[a].

	2 SNP Haplotypes	3 SNP Haplotypes	Average
		SW/LR	
Scenario 1	0.333	0.333	0.333
Scenario 2	0.000	0.000	0.000
Overall	0.167	0.167	0.167
		ACA/LR	
Scenario 1	0.833	0.833	0.833
Scenario 2	0.167	0.500	0.333
Overall	0.500	0.667	0.580
		RG	
Scenario 1	_____	_____	0.333
Scenario 2	_____	_____	0.170
Overall	_____	_____	0.250

[a] Power was calculated as the proportion of times at least one SNP in high linkage disequilibrium (>.9) with a causative mutations was detected by the model at α=.05 for genome-wide significance.

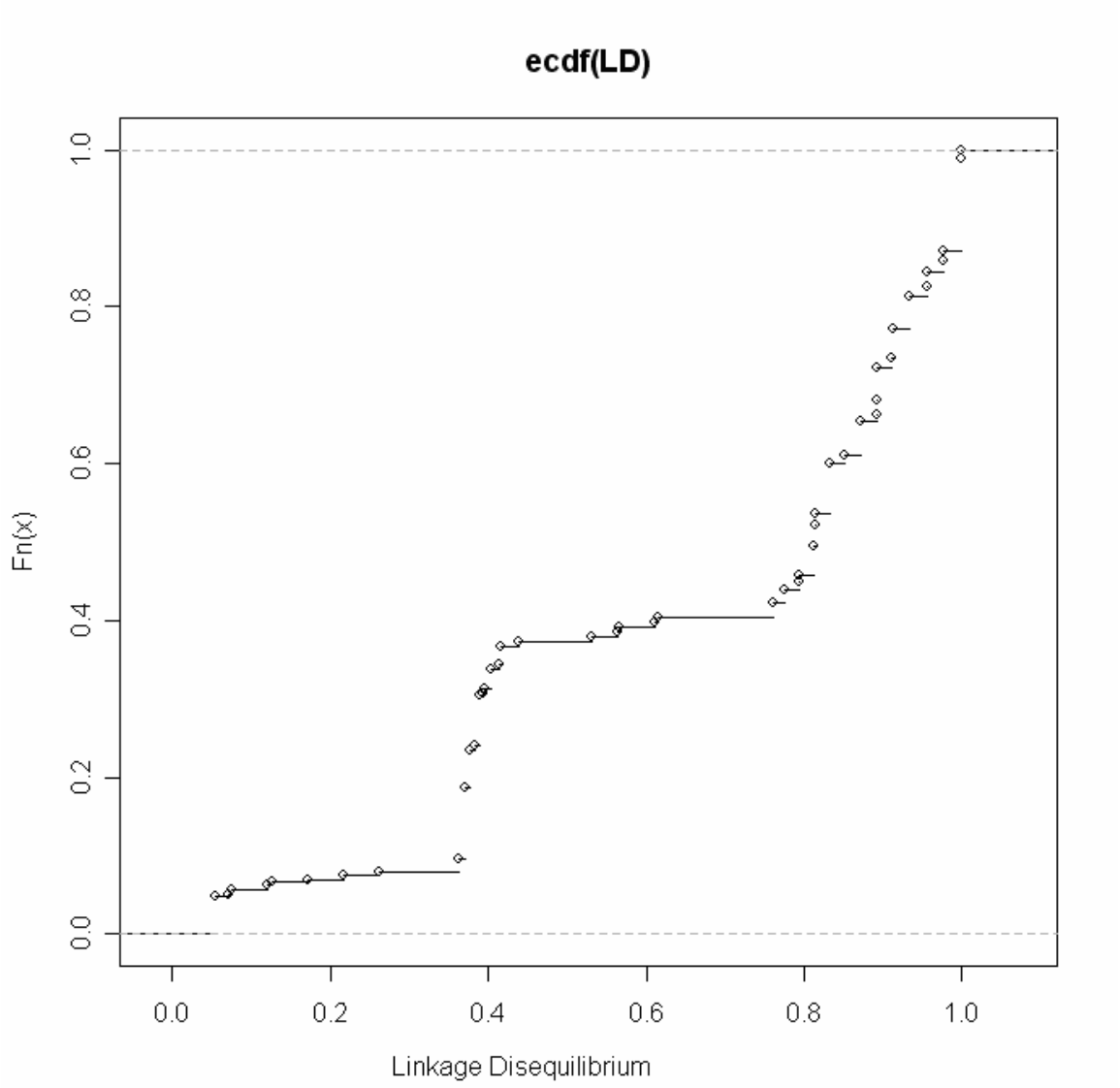


Figure 5.1. Plot of the cumulative distribution of SNP, identified as have significant associations when using ACA/LR, based on linkage disequilibrium with the causative mutations.

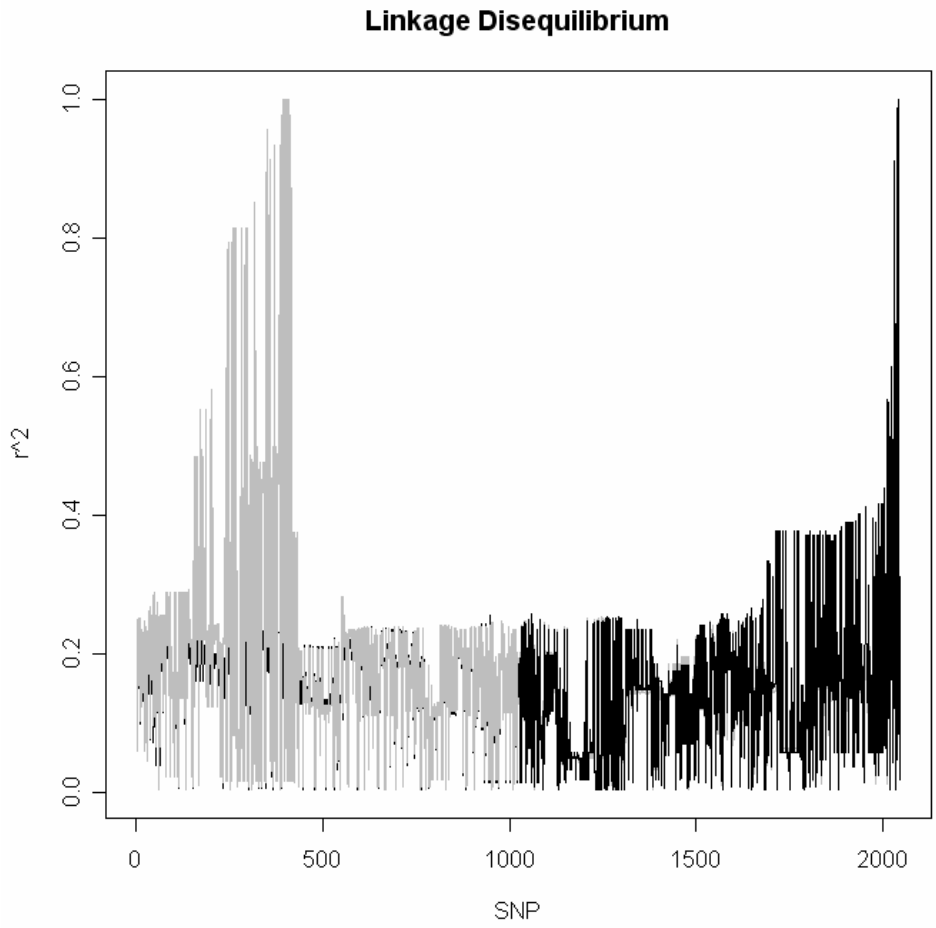
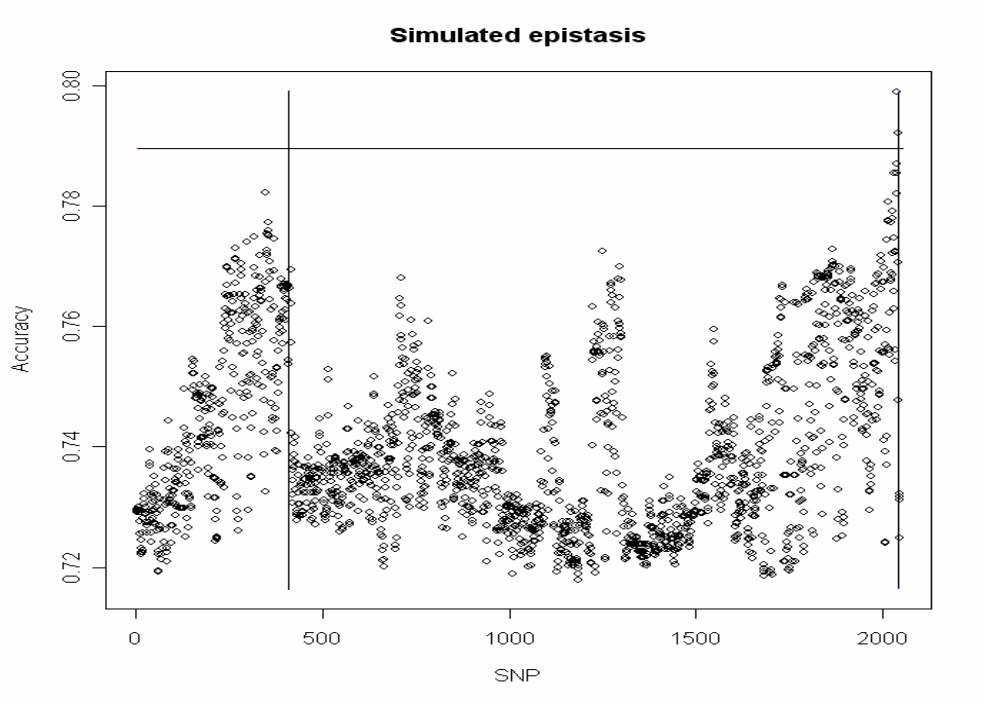
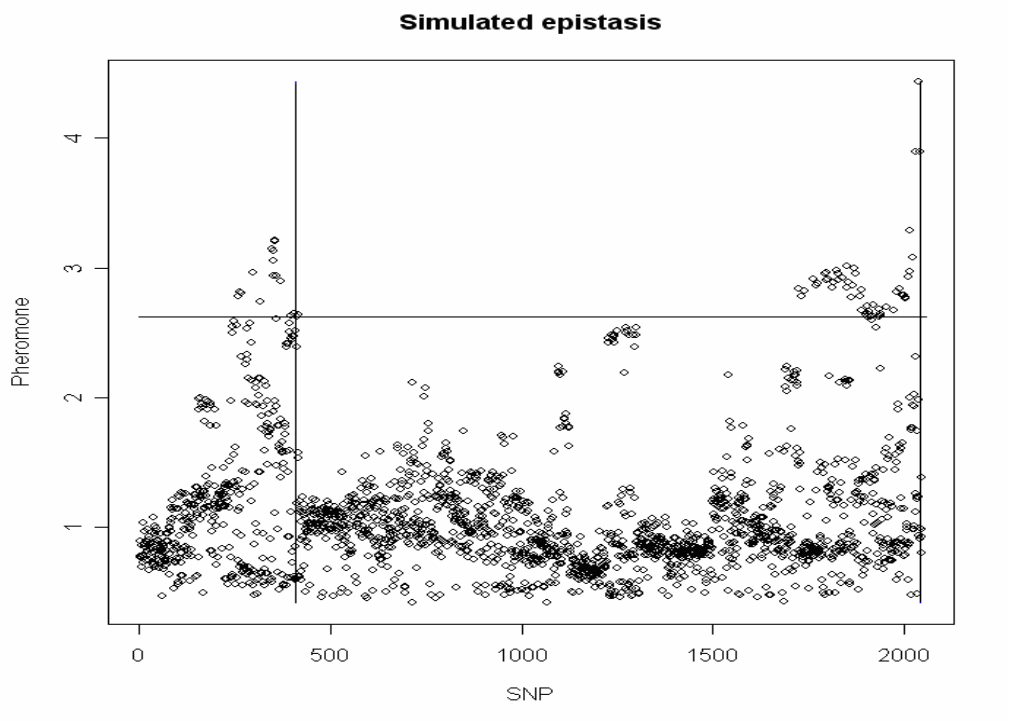


Figure 5.2. Plots of each marker's linkage disequilibrium (LD) with the two causative mutations. The light grey line represents LD with the causative mutation located at position 409. The black line represents LD with the causative mutation located at position 2041.

a.


Simulated epistasis
Accuracy
0.72
0.74
0.76
0.78
0.80
0
500
1000
1500
2000
SNP

b.


Simulated epistasis
Pheromone
1
2
3
4
0
500
1000
1500
2000
SNP

c.

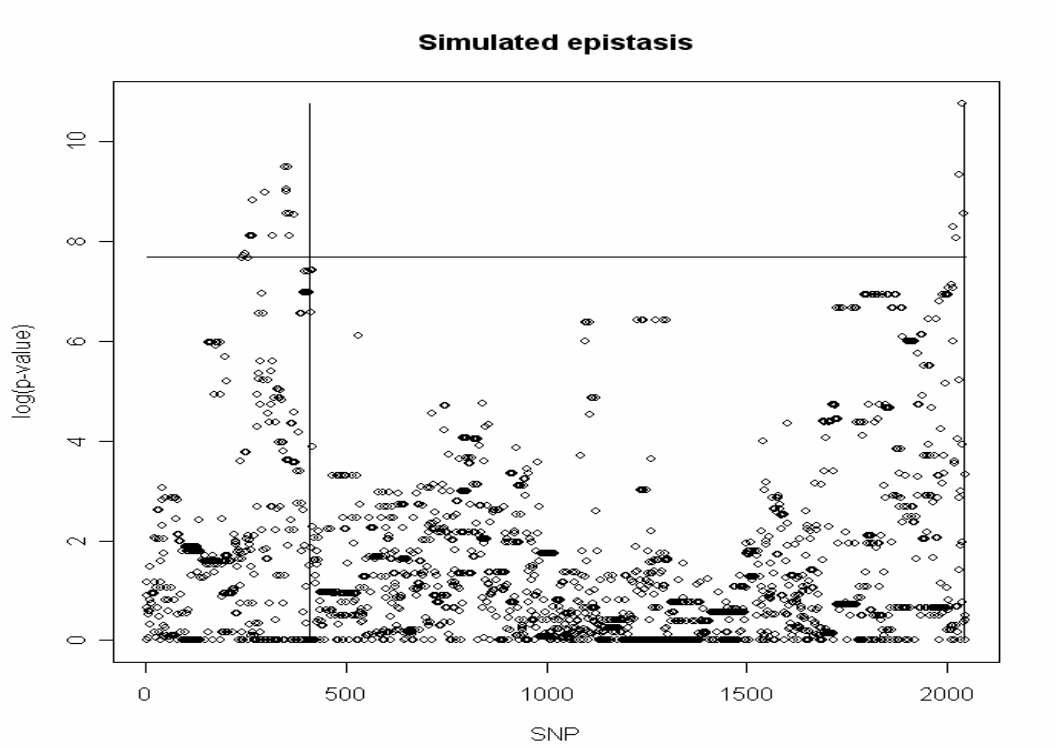
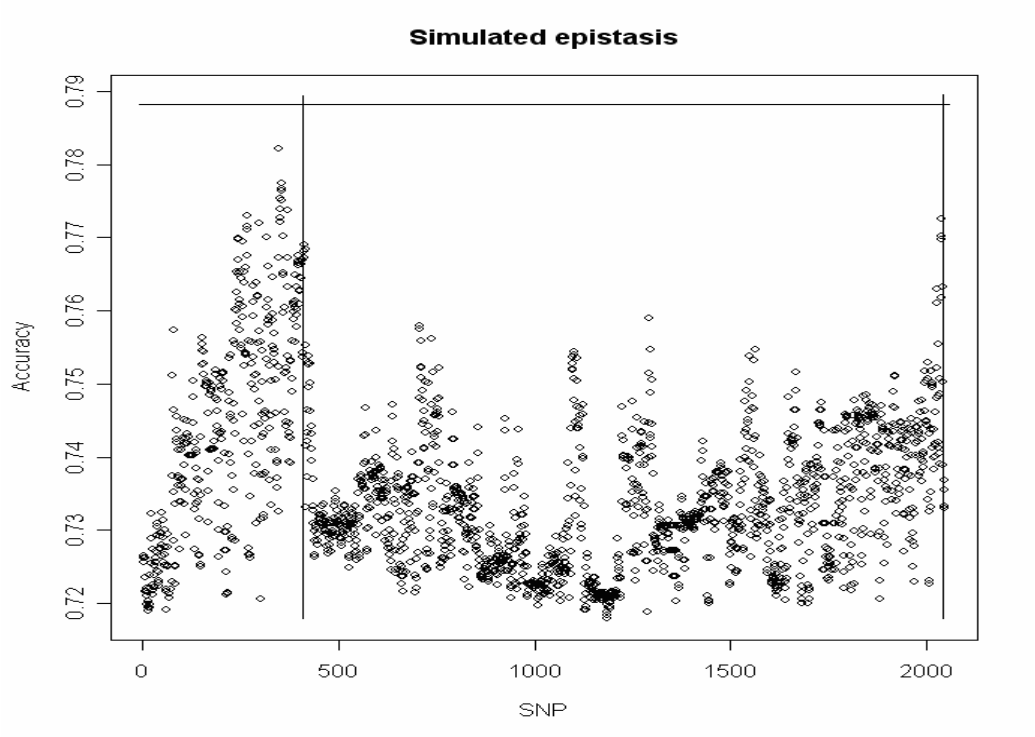
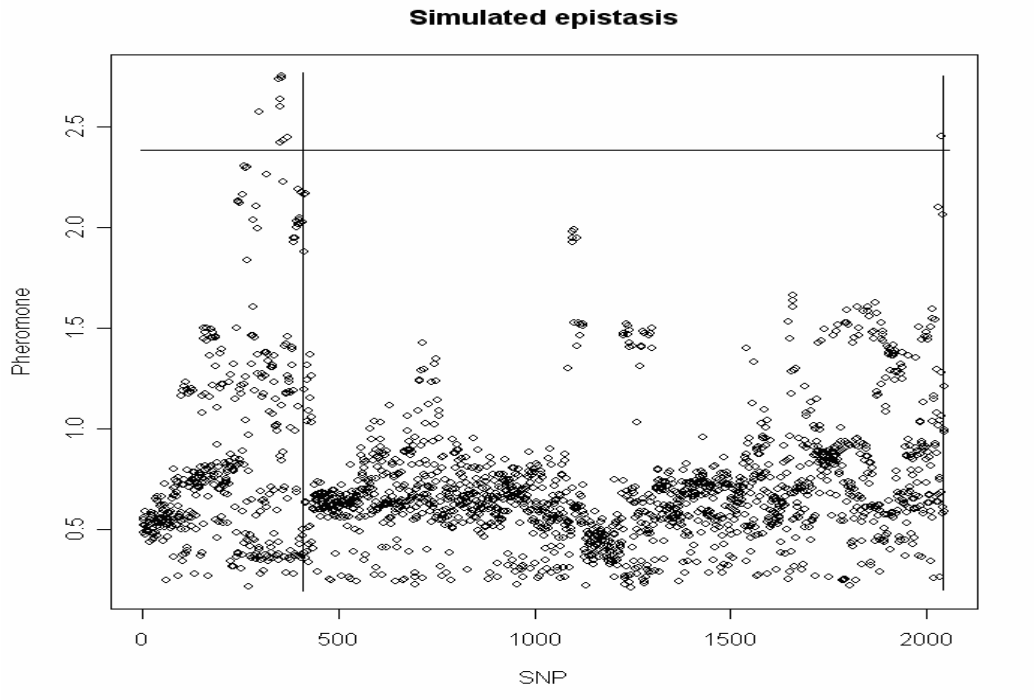


Figure 5.3. Association plots of SNP markers for the simulated trait under scenario 1. Plots were obtained using 3 SNP haplotypes analyzed by a. SW/LR and b. ACA/LR. Vertical lines represent the position of the two causative mutations, and horizontal lines represent the threshold at which associations are significant at α=.05.

a.


Simulated epistasis
Accuracy
0.72
0.73
0.74
0.75
0.76
0.77
0.78
0.79
0
500
1000
1500
2000
SNP

b.


Simulated epistasis
Pheromone
0.5
1.0
1.5
2.0
2.5
0
500
1000
1500
2000
SNP

c.

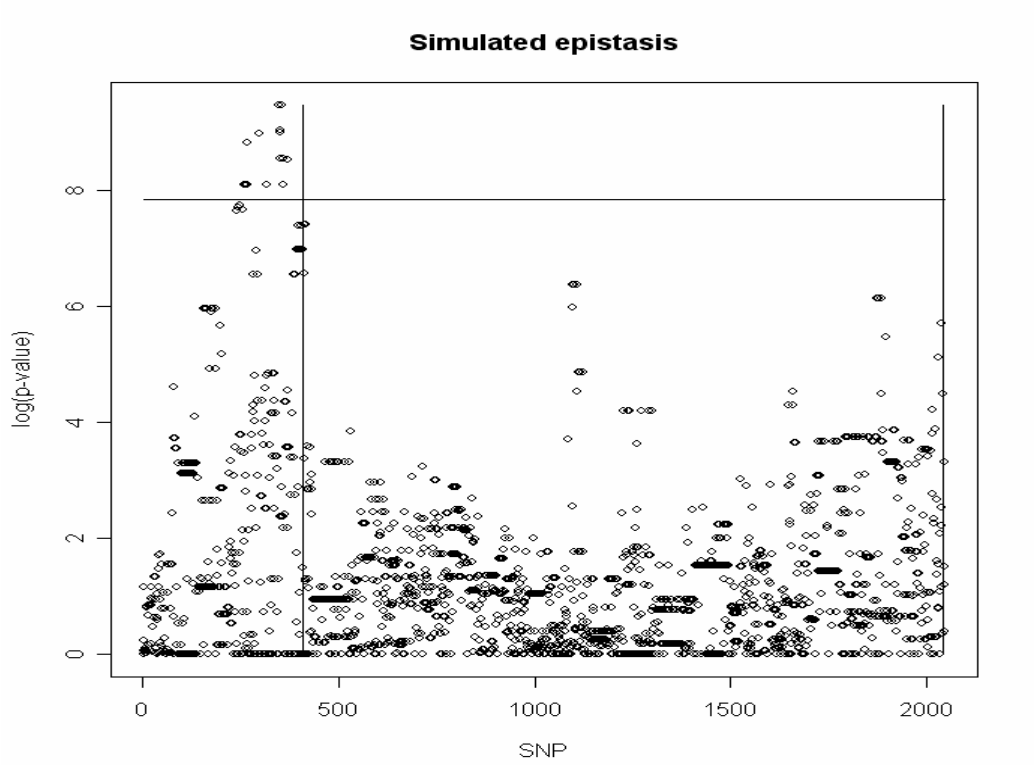


Figure 5.4. Association plots of SNP markers for the simulated trait under scenario 2. Plots were obtained using 3 SNP haplotypes analyzed by a. SW/LR, b. ACA/LR, and c. RG. Vertical lines represent the position of the two causative mutations, and horizontal lines represent the threshold at which associations are significant at α=.05.

CHAPTER 6

CONCLUSIONS

This study has shown that the use of methodologies capable of accounting for misdiagnosis and modeling complex data structures in a computationally efficient manner can increase prediction accuracy and the power to detect informative features. For diseases in which clinical markers are ineffective for diagnosis, the misclassification algorithm can provide a means to identify potentially misdiagnosed individuals. Given the results obtained using an Alzheimer's disease data set, it is clear that the use of genomic information to identify misdiagnosed individuals is necessary to train effective machine learning algorithms and identify important genomic features.

Results of studies involving the analysis of high-dimension genomic data sets demonstrated the effectiveness of the ant colony algorithm for modeling complex genomic data structures. When applied to several high-dimensional data sets, the ant colony algorithm achieved higher prediction accuracies than all other feature selection methods examined. In contrast to previous applications of optimization algorithms, the ant colony algorithm yielded high accuracies without the need to pre-select a small percentage of genes. Furthermore, the ant colony algorithm was able to identify small subsets of genes related to both tissue of origin and neoplasia, demonstrating the algorithms ability to identify highly predictive and biologically relevant genes in data sets with large numbers of features.

For or analysis of marker associations in the presence of simulated epistasis, models accounting only for marginal effects had little power to detect SNP in high linkage disequilibrium with causative mutations. In contrast the proposed method, utilizing ant colony

optimization to account for gene interactions, obtained high power when the epistatic effect was strong. When the interaction effect was relatively weak all models showed reduced power; however the ant colony method showed substantially greater power than all other models examined. Clearly for high-dimension association studies, methodologies capable of efficiently modeling gene interactions, such as the model proposed in this study, should yield superior performance in terms of the power to detect SNP in linkage disequilibrium with causative mutations.

APPENDIX A

GENES SELECTED FROM THE GCM DATA SET BY THE ANT COLONY ALGORITM

Table A.1. Probes selected in at least 2 replicates using ACA for breast adenocarcinomas.

Probe set ID	Affymetrix gene annotation	Proportion of pheromone[a]	Number of times probe was selected
RC_AA287268_at	EST: zs49h12.s1 NCI_CGAP_GCB1 Homo sapiens cDNA clone IMAGE:700871 3', mRNA sequence. (from Genbank)	0.0040184686	2
RC_AA598684_s_at	EST: ae49a02.s1 Stratagene lung carcinoma 937218 Homo sapiens cDNA clone 950186 3', mRNA sequence. (from Genbank)	0.0039910622	2

[a]Relative weight given to probe by pheromone function, averaged across all replicates.

Table A.2. Probes selected in at least 2 replicates using ACA for prostate adenocarcinomas.

Probe set ID	Affymetrix gene annotation	Proportion of pheromone[a]	Number of times probe was selected
S39329_at	KLK1 Kallikrein 1 (renal/pancreas/salivary) {alternative products}	0.0181056833	5
HG2261-HT2351_s_at	Antigen, Prostate Specific, Alt. Splice Form 2	0.0175960061	5
RC_AA176975_s_at	Human prostatic secretory protein 57 mRNA, complete cds	0.0171006689	4
M34376_s_at	MSMB Beta microseminoprotein (prostate secreted)	0.0165141994	3
M24902_at	ACPP Acid phosphatase, prostate	0.0159989983	2

[a]Relative weight given to probe by pheromone function, averaged across all replicates.

Table A.3. Probes selected in at least 2 replicates using ACA for lung adenocarcinomas.

Probe set ID	Affymetrix gene annotation	Proportion of pheromone[a]	Number of times probe was selected
M24461_at	PULMONARY SURFACTANT-ASSOCIATED PROTEIN B PRECURSOR	0.0193028881	5
M68519_rna1_at	Pulmonary surfactant-associated protein SP-A (SFTP1) gene	0.0172827178	5

[a]Relative weight given to probe by pheromone function, averaged across all replicates.

Table A.4. Probes selected in at least 2 replicates using ACA for colorectal adenocarcinomas.

Probe set ID	Affymetrix gene annotation	Proportion of pheromone[a]	Number of times probe was selected
M10050_at	HBG2 Hemoglobin gamma-G	0.0196498774	5
M29540_at	CARCINOEMBRYONIC ANTIGEN PRECURSOR	0.019266758	5
AB006781_s_at	Galectin-4	0.0185816009	5
M35252_at	TUMOR-ASSOCIATED ANTIGEN CO-029	0.0176392264	5
X83228_at	LI-cadherin	0.0175583835	4
X68314_at	GPX2 Glutathione peroxidase 2, gastrointestinal	0.016547709	4
M18728_at	NCA Non- specific cross reacting antigen	0.0164756214	3
RC_AA100719_s_at	Non- specific cross reacting antigen	0.0149580588	3

[a]Relative weight given to probe by pheromone function, averaged across all replicates.

Table A.5. Probes selected in at least 2 replicates using ACA for lymphoma.

Probe set ID	Affymetrix gene annotation	Proportion of pheromone[a]	Number of times probe was selected
M27394_s_at	B- lymphocyte cell- surface antigen B1 (CD20)	0.0172119231	5
X12530_s_at	CD20 RECEPTOR	0.0170145245	5
M89957_at	2- CD79B antigen (immunoglobulin-associated beta)	0.0164564949	5
X07203_at	CD20 RECEPTOR	0.0164357185	5
M89957_at	IGB Immunoglobulin- associated beta (B29)	0.0163790303	5
U77180_at	EBI1- ligand chemokine	0.01509933	4

[a]Relative weight given to probe by pheromone function, averaged across all replicates.

Table A.6. Probes selected in at least 2 replicates using ACA for bladder transition cell carcinoma.

Probe set ID	Affymetrix gene annotation	Proportion of pheromone[a]	Number of times probe was selected
X52426_s_at	KRT13 Keratin 13	0.0121916633	5
U43901_rna1_s_at	37 kD laminin receptor precursor/p40 ribosome associated protein gene	0.0101222846	4
U03754_f_at	Major histocompatibility complex, class I, A	0.0098449775	3
T80746_s_at	Ferritin, light polypeptide	0.0095760173	4
AFFX-HUMGAPDH/M33197_M_at	AFFX- HUMGAPDH/M33197_M_at (endogenous control)	0.0094508355	4
V00478_s_at	Actin, beta	0.009376579	4
AFFX-HUMGAPDH/M33197_M_at	2- Glyceraldehyde-3- phosphate dehydrogenase	0.0092263344	4
RC_AA135095_at	Homo sapiens Sox- like transcriptional factor mRNA, complete cds	0.008699325	4
RC_AA496882_s_at	Eukaryotic translation initiation factor 3, subunit 4 (delta, 44kD)	0.0084978563	3
M13560_s_at	PROBABLE PROTEIN DISULFIDE ISOMERASE ER-60 PRECURSOR	0.008175368	3
Y07755_at	S100A2 gene, exon 1, 2 and 3	0.0081133016	
X03689_s_at	mRNA fragment for elongation factor TU (N- terminus)	0.0080484947	3
D49824_s_at	HLA- B null allele mRNA	0.0079409828	3
M21142_cds2_s_at	Guanine nucleotide- binding protein G-s-alpha-3 gene extracted from Human guanine nucleotide- binding protein alpha-subunit gene (G-s-alpha	0.007882917	3
L11566_at	RPL18 Ribosomal protein L18	0.0074147932	3
X00351_f_at	ACTB Actin, beta	0.0069740126	2
AFFX-HSAC07/X00351_at-2	No info for gene	0.0069019349	2
U14970_at	RPS5 Ribosomal protein S5	0.0068663867	2
T48195_s_at	Eukaryotic translation initiation factor 3, subunit 3 (gamma, 40kD)	0.0063847932	2

[a]Relative weight given to probe by pheromone function, averaged across all replicates.

Table A.7. Probes selected in at least 2 replicates using ACA for melanoma.

Probe set ID	Affymetrix gene annotation	Proportion of pheromone[a]	Number of times probe was selected
S73003_s_at	PMEL 17 PROTEIN PRECURSOR	0.0146365037	5
RC_AA176812_at	EST: zp32g12.s1 Stratagene neuroepithelium (#937231) Homo sapiens cDNA clone 611206 3' similar to contains Alu repetitive element;contains element THR repetitive element ;, mRNA sequence. (from Genbank)	0.0141555994	5
X58079_at	S100 alpha protein	0.0077156399	2

[a]Weight given to the probe/ the sum of all weights, then averaged across all replicates.

Table A.8. Probes selected in at least 2 replicates using ACA for uterine adenocarcinomas.

Probe set ID	Affymetrix gene annotation	Proportion of pheromone[a]	Number of times probe was selected
X63187_at	HE4 mRNA for extracellular proteinase inhibitor homologue	0.0153718754	5
X03635_at	ESR Estrogen receptor	0.0145918918	4
AA393164_s_at	Mammaglobin 2	0.0134894851	4
M76732_s_at	HOX7 gene, exon 2 and complete cds	0.012567893	3

[a]Relative weight given to probe by pheromone function, averaged across all replicates.

Table A.9. Probes selected in at least 2 replicates using ACA for leukemia.

Probe set ID	Affymetrix gene annotation	Proportion of pheromone[a]	Number of times probe was selected
U22376_cds2_s_at	C- myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds	0.0127216499	5
M11722_at	Terminal transferase mRNA	0.0125944135	5
X89399_s_at	Ins(1,3,4,5)P4- binding protein	0.0123198538	5

[a]Relative weight given to probe by pheromone function, averaged across all replicates.

Table A.10. Probes selected in at least 2 replicates using ACA for renal cell carcinoma.

Probe set ID	Affymetrix gene annotation	Proportion of pheromone[a]	Number of times probe was selected
RC_AA434245_r_at	EST: zw24g05.s1 Soares ovary tumor NbHOT Homo sapiens cDNA clone 770264 3', mRNA sequence. (from Genbank)	0.0111998013	3
U50743_at	Na,K- ATPase gamma subunit mRNA	0.0107642568	5
X92744_at	BETA- DEFENSIN 1 PRECURSOR	0.0086855248	4
X59798_at	CCND1 Cyclin D1 (PRAD1; parathyroid adenomatosis 1)	0.0082718566	3
M63438_s_at	GLUL Glutamate- ammonia ligase (glutamine synthase)	0.0080361106	2
RC_AA401098_f_at	EST: zu50g01.s1 Soares ovary tumor NbHOT Homo sapiens cDNA clone 741456 3'	0.0079398391	4
RC_AA609213_at	EST: af12f09.s1 Soares testis NHT Homo sapiens cDNA clone 1031465 3', mRNA sequence. (from Genbank)	0.007656843	2
AFFX-HUMGAPDH/M33197_M_at	AFFX-HUMGAPDH/M33197_M_at (endogenous control)	0.007595209	2
L05144_at	PCK1 Phosphoenolpyruvate carboxykinase 1 (soluble)	0.0074889417	2
RC_AA465690_s_at	Splicing factor, arginine/serine-rich 11	0.0073764694	2
M30257_s_at	VCAM1 Vascular cell adhesion molecule 1	0.0071865214	2
M87789_s_at	(hybridoma H210) anti-hepatitis A IgG variable region, constant region, complementarity- determining regions mRNA	0.0068022391	2
M16594_at	GSTA1 Glutathione S-transferase A2	0.0067434763	2
M21142_cds2_s_at	Guanine nucleotide- binding protein G-s-alpha-3 gene	0.0067406213	2
C14412_s_at	EST: Human fetal brain cDNA 5'- end GEN-055A09, mRNA sequence. (from Genbank)	0.0062901421	2

M95167_at	SLC6A3 Solute carrier family 6 (neurotransmitter transporter, dopamine), member 3	0.006046546	2
AA071387_at	Homo sapiens hJTB mRNA, complete cds	0.0060210345	2
U51240_at	KIAA0085 gene, partial cds	0.0058962097	2
M21305_at	Alpha satellite and satellite 3 junction DNA sequence	0.0056945015	2
M16279_at	MIC2 Antigen identified by monoclonal antibodies 12E7, F21 and O13	0.0054714429	2
M34516_at	Omega light chain protein 14.1 (Ig lambda chain related) gene, exon 3	0.0052594813	2
HG3543	HT3739_at- Insulin- Like Growth Factor 2	0.0046115439	2
AA422025_s_at	EST: zv26f07.r1 Soares NhHMPu S1 Homo sapiens cDNA clone 754789 5', mRNA sequence. (from Genbank)	0.004338238	2

[a]Relative weight given to probe by pheromone function, averaged across all replicates.

Table A.11. Probes selected in at least 2 replicates using ACA for pancreatic adenocarcinoma.

Probe set ID	Affymetrix gene annotation	Proportion of pheromone[a]	Number of times probe was selected
Z48314_s_at	MUC5B Mucin 5, subtype B, tracheobronchial	0.0133790586	5
M22612_f_at	PRSS1 Protease, serine, 1 (trypsin 1)	0.0127774756	4
AA372630_s_at	Homo sapiens GW112 protein (GW112) mRNA, complete cds	0.0123590937	3
RC_AA100719_s_at	Non- specific cross reacting antigen	0.0103259547	3
X51698_s_at	SPASMOLYTIC POLYPEPTIDE PRECURSOR	0.0092080707	3

[a]Relative weight given to probe by pheromone function, averaged across all replicates.

Table A.12. Probes selected in at least 2 replicates using ACA for ovarian adenocarcinoma.

Probe set ID	Affymetrix gene annotation	Proportion of pheromone[a]	Number of times probe was selected
AA393164_s_at	Mammaglobin 2	0.0137589273	5
M34516_at	Omega light chain protein 14.1 (Ig lambda chain related) gene, exon 3	0.0122839038	4
M63438_s_at	GLUL Glutamate- ammonia ligase (glutamine synthase)	0.0094549222	4
RC_AA456055_at	EST: aa03f02.s1 Soares NhHMPu S1 Homo sapiens cDNA clone 812187 3', mRNA sequence. (from Genbank)	0.0090919635	3
RC_AA164851_at	EST: zp02c11.s1 Stratagene ovarian cancer (#937219) Homo sapiens cDNA clone 595220 3', mRNA sequence. (from Genbank)	0.0087784512	3
L02321_at	GSTM5 Glutathione S- transferase M5	0.008183824	2
X58079_at	S100 alpha protein	0.007658021	2
T34377_at	EST: EST66741 Homo sapiens cDNA 5' end similar to None. (from Genbank)	0.00695011	
RC_AA236533_s_at	Ecotropic viral integration site 1	0.0068410586	2
M13560_s_at	PROBABLE PROTEIN DISULFIDE ISOMERASE ER-60 PRECURSOR	0.0065232028	2
RC_AA190676_at	EST: zp89g09.s1 Stratagene HeLa cell s3 937216 Homo sapiens cDNA clone 627424 3', mRNA sequence. (from Genbank)	0.0064895641	3
X04470_s_at	RPL32 Ribosomal protein L32	0.0062484829	2
J05582_s_at	MUC1 Mucin 1, transmembrane	0.0058626952	2
U20391_rna6_at	Folate receptor (FOLR1) gene	0.0057476571	
HG4058-HT4328_at	Oncogene Aml1-Evi-1, Fusion Activated	0.0056173073	2
U78525_at	Eukaryotic translation initiation factor (eIF3) mRNA	0.005379787	2
U40434_at	Pre-pro- megakaryocyte potentiating factor	0.0051919395	2
D78361_at	Ornithine decarboxylase antizyme, ORF 1 and ORF 2	0.0043444532	
D14710_at	ATP5A1 ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1, cardiac muscle	0.0042439403	2
J00105_s_at	BETA-2- MICROGLOBULIN	0.0039355229	2

	PRECURSOR		
D49824_s_at	HLA- B null allele mRNA	0.0039298668	2
U04241_at	Homolog of Drosophila enhancer of split m9/m10 mRNA	0.0038872194	2
M21142_cds2_s_at	Guanine nucleotide- binding protein G-s-alpha-3 gene extracted from Human guanine nucleotide- binding protein alpha- subunit gene (G-s-alpha)	0.0034528747	2
D13748_at	EIF4A1 Eukaryotic translation initiation factor 4A (eIF-4A) isoform 1	0.0033250347	2

[a]Relative weight given to probe by pheromone function, averaged across all replicates.

Table A.13. Probes selected in at least 2 replicates using ACA for pleural mesothelioma.

Probe set ID	Affymetrix gene annotation	Proportion of pheromone[a]	Number of times probe was selected
RC_AA150210_at	EST: zl04g08.s1 Soares pregnant uterus NbHPU Homo sapiens cDNA clone 491390 3', mRNA sequence. (from Genbank)	0.0176189042	5
X16662_at	ANX8 Annexin VIII	0.0159638962	5
RC_AA293796_at	EST: zt56g08.s1 Soares ovary tumor NbHOT Homo sapiens cDNA clone 726398 3', mRNA sequence. (from Genbank)	0.015678904	4
U08021_at	Nicotinamide N- methyltransferase (NNMT) mRNA	0.0152920451	4
U03877_at	HEAT SHOCK 70 KD PROTEIN 1	0.0147746168	2
RC_AA419609_at	EST: zv04b06.s1 Soares NhHMPu S1 Homo sapiens cDNA clone 752627 3', mRNA sequence. (from Genbank)	0.0146994704	3

[a]Relative weight given to probe by pheromone function, averaged across all replicates.

Table A.14. Probes selected in at least 2 replicates using ACA for CNS.

Probe set ID	Affymetrix gene annotation	Proportion of pheromone[a]	Number of times probe was selected
RC_D59321_f_at	Homo sapiens mRNA for APCL protein, complete cds	0.0151449024	5
RC_AA460849_at	EST: zx64h08.s1 Soares total fetus Nb2HF8 9w Homo sapiens cDNA clone 796287 3', mRNA sequence. (from Genbank)	0.0123586904	3
RC_AA284767_at	EST: zt21h07.s1 Soares ovary tumor NbHOT Homo sapiens cDNA clone 713821 3', mRNA sequence. (from Genbank)	0.0122941802	4
RC_AA007153_at	EST: 13cDNA40-3.seq Soares infant brain 1NIB Homo sapiens cDNA clone HY18-44 3', mRNA sequence. (from Genbank)	0.0114697837	3
W26436_s_at	Microtubule- associated protein 1B	0.0114056388	3
RC_AA262340_at	EST: zr71g09.s1 Soares NhHMPu S1 Homo sapiens cDNA clone 668896 3', mRNA sequence. (from Genbank)	0.0109292744	3
C14203_s_at	EST: Human fetal brain cDNA 5'- end GEN-037E11, mRNA sequence. (from Genbank)	0.0108949567	2
RC_AA227443_s_at	Kinesin family member 5C	0.0101217607	2

[a]Relative weight given to probe by pheromone function, averaged across all replicates.