Provisional chapter

Ant Colony Algorithm with Applications in the Field of Genomics

R. Rekaya, K. Robbins, M. Spangler, S. Smith,
E. H. Hay and K. Bertrand

Additional information is available at the end of the chapter

1. Introduction

Ant colony algorithms (ACA) were first proposed by Dorigo *et al*. (1999) to solve difficult optimization problems, such as the traveling salesman, and have since been extended to solve many discrete optimization problems. As the name would imply, ACA are derived from the process by which ant colonies find the shortest route to a food source. Real ant colonies communicate through the use of chemicals called pheromones which are deposited along the path an ant travels. Ants that choose a shorter path will transverse the distance at a faster rate, thus depositing more pheromone. Subsequent ants will then choose the path with more pheromone creating a positive feedback system. Artificial ants work as parallel units that communicate through a cumulative distribution function (CDF) that is updated by weights, determined by the "distance" traveled on a selected "path", which are analogous to the pheromones deposited by real ants (Dorigo *et al*. 1999, Ressom *et al*. 2006). As the CDF is updated, "paths" 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. In the specific application of feature selection, the "path" chosen by an artificial ant is a subset of features selected from a larger sample space, and the "distance" traveled is some measure of the features performance.

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 deal-

ing 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). For ACA, 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 (Dorigo 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.

2. General presentation of ant colony algorithm

The ACA employs artificial ants that communicate through a probability density function (PDF) that is updated at-each iteration with weights or "pheromone levels", which are analogous to the chemical pheromones used by real ants. The weights can be determined by the strength of the association between selected feature and the response of interest. Using the notation in [Dorigo and Gambardella, 1997; Ressom et al., 2006], the probability of sampling feature m at time t is defined as:

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

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

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:

$$\tau_m(t+1) = (1-\rho) * \tau_m(t) + \Delta\tau_m(t) \tag{2}$$

where ρ is a constant between 0 and 1 representing the rate at which the pheromone trail evaporates; $\Delta\tau_m(t)$is the change in pheromone level for feature m based on the sum of accuracy of all S_k containing SNP m, and is set to zero if feature m was not selected by any of the artificial ants.

Although the general idea of the ACA is simple and intuitive, its application to solve real world applications requires some good heuristics in defining the pheromone functions and their updating. In this chapter, we are presenting three applications of the ACA in the field of genetics and genomics based on previously published research by our group [Robbins et al., 2007, Robbins et al., 2008; Spangler et al., 2008; Rekaya and Robbins, 2009; Robbins et al., 2011]. Specific implementation details for each application are added in the appropriate sections of the chapter.

2.1. Ant colony algorithm for feature selection in high dimension gene expression data for disease classification

The idea of using gene expression data for diagnosis and personalized treatment presents 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).

In this section the ACA was implemented using the high-dimensional GCM data-set (Ramaswamy et al., 2001), containing 16,063 genes and 14 tumor classes, with very limited prefiltering, 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.

A.1 Latent variable model: A Bayesian regression model was used to predict tumor type in the form of a probability $p_{ic}(y_{ic}=1)$, with $y_{ic} = 1$ indicating that sample i is from tumor class c. The regression on the vector of binary responses 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:

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

where X_{ic} corresponds to row i of the design matrix X_c for tumor class c. The link function of the expectation of the liability $\boldsymbol{X}_{ic}\boldsymbol{\beta}_c$ with the binary response y_{ic} was constructed via a probit model (West, 2003) yielding the following equations:

$$p_{ic}(y_{ic}=1) = \boldsymbol{\Phi}(\boldsymbol{X}_{ic}\boldsymbol{\beta}_c) \text{ and } p_{ic}(y_{ic}=1) = 1 - \boldsymbol{\Phi}(\boldsymbol{X}_{ic}\boldsymbol{\beta}_c)$$

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

A.2 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.

A.3 Ant colony optimization: The general ACA presented in the previous section was used. The prior information, η_{mc} , was assumed as:

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

where $\boldsymbol{f_c}$is a vector of all fold change values for tumor class c; $\boldsymbol{t_c}$is a vector of all t-statistic values for tumor class c; and $\boldsymbol{pt_c}$ is a vector of all penalized t-statistic values for tumor class c. After several trail runs the parameters α and β were set to 1 and.3 respectively.

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 following equation (2).

The procedure can be summarized in the following steps:

1. Each ant selects a predetermined number of genes.
2. 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 (VD) is used at any point in the ACA).
3. 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.
4. 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.
5. The accuracy for each tumor class c is calculated as:

$$acc_c = \frac{\sum_{i=1}^{nc} \Phi(\mathbf{P_{ic}}\beta_c) / nc + \sum_{i=1}^{nr} 1 - \Phi(\mathbf{P_{ic}}\beta_c) / nr}{2} \quad (3)$$

where $\boldsymbol{P_{ic}}$contains principle component values for sample i for tumor class c; $\boldsymbol{\beta_c}$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.

6. 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).

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 select based on the level of pheromone trail deposited on each feature.

A.4 GCM data set: The data set contained 198 samples collected from 14 tumor types: BR (breast adenocarcinoma), Pr (prostate adenocarcinoma), LU (lung adenocarcinoma), CO (colorectal adenocarcinoma), LY (lymphoma), BL (bladder transitional cell carcinoma), ML (melanoma), UT (uterine adenocarcinoma), LU (leukemia), RE (renal cell carcinoma), PA (pancreatic adenocarcinoma), OV (ovarian adenocarcinoma), ME (pleural mesothelioma), and CNS (central nervous system). The unedited data set contained the intensity values of 16063 probes generate using Affymetrix high density oligonucleotide microarrays, and calculated using Affymetrix GENECHIP software (Ramaswamy et al, 2001). Following the thresholding of intensity values to a minimum value of 20 and a maximum value of 16000, a log base 2 transformation was applied to the data set. Genes with the highest expression values being less than two times the smallest were removed, leaving 14525 probes for analysis.

A.5 Results and discussions: The GCM data set has been a benchmark to compare the performance of classification and feature selection algorithms. Table 1 shows the best prediction accuracies obtained by methods used in this study and several previous studies (GASS (Lin et al., 2006), GA/MLHD (Ooi and Tan, 2003), MAMA (Antonov et al., 2004), and GA/SVM (Liu et al., 2005)) 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%, 40%, and 16.6% in accuracy over the FC/LVM, T/LVM, and PT/LVM methods of feature selection, respectively.

	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)	64.8	____	____
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

[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, t-test, and penalized t-test values.

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

Due to its poor performance, the confusion matrix of predictions using T/LVM is not included, but matrices for the predictions obtained by the ACA/LVM, FC/LVM, and PT/LVM using the GCM split can be found in Tables 2-4. These tables show that the ACA/LVM performs as good 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). The ACA/LVM also achieved 100% prediction accuracy for 10 of the 14 tumor classes, as compared to only 7 and 8 when using FC/LVM or PT/LVM, respectively.

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

To further evaluate performance, each of the feature selection algorithms was tested using four additional random splits of the data. The best classification accuracies obtained for each algorithm can be found in Table 5. The ACA/LVM algorithm yielded the best prediction accuracies for all replicates, with increases in accuracies ranging from 6.7% to 14% over the best accuracies obtained by filter methods. When looking at the three filter methods it can be seen that the best method varied depending on the replication. These findings are in agreement with Jefferey et al. (2006).

Replication	1	2	3	4	5
ACA/LVM	90.7	83.3	79.6	81.5	88.9
FC/LVM	79.6	77.8	68.5	72.2	75.9
PT/LVM	77.8	77.8	66.7	68.5	81.5
AVG[b]/LVM	79.6	70.4	70.4	70.4	83.3

[a] Split used by Ramaswamy et al 2001; [b]Weighted average of scaled fold change (FC), t-test (PT), and penalized t-test values (PT).

Table 5. Classification accuracies using several feature selection methods

Due to a lack of any good criterion for determining an objective cut-off value for the rank based methods, several values were used and evaluated. Since the use of fewer features is desirable from a biological standpoint, an upper limit of 50 genes per tumor class was imposed on all methods. Table 6 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 10 of the 14 tumor classes, the ACA/LVM selects fewer genes than the rank based methods.

The performance of the ACA/LVM model was superior, not only to the filter based methods used in this study, but also several reported results using the GCM data set. The ACA/LVM consistently yielded superior accuracies using fewer genes than the filter based methods, for which ranks varied with each replication. 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 in which truncation points were somewhat arbitrary. The objective selection criteria and robustness of the ACA, within the confines of the GCM data set, 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.

	BR	PR	LU	CO	LY	BL	ML	UT	LE	RE	PA	OV	ME	CNS
ACA	3.4	4.8	2	7.8	6.6	19.6	4.6	7.6	3.2	16	14.6	17.2	5	5.6
FC	18	18	18	18	18	18	18	18	18	18	18	18	18	18
PT	14	14	14	14	14	14	14	14	14	14	14	14	14	14
Average[a]	18	18	18	18	18	18	18	18	18	18	18	18	18	18

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

Table 6. Number of genes selected for each tumor type using ACA and other feature selection methods.

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 to the predictive power of an ensemble of genes. Additionally, depending on the method of truncation, the reduced gene list could be highly redundant (Lin et al., 2006; Shen et al., 2006), further reducing the informativeness of pre-selected genes. 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.

For LU tumors, the ACA identified two genes capable of classifying LU tumor samples with 100%, 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 100% 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.

2.2. The use of the ant colony algorithm for the detection of marker associations in the presence of gene interactions

With the advent of high-throughput, cost effective genotyping platforms, there has been much focus on the use of high-density single nucleotide polymorphism (SNP) genotyping 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]. Given the high density of SNP maker maps, examining all possible interactions is seldom possible computationally. As a result, studies examining gene interactions tend to focus on a small number of SNP, previously identified as having strong marginal associations. Using an exhaustive search of all two-way interactions, Marchini et al. achieved greater power to detect causative mutations than when estimating only marginal effects. Due to the high computational cost of this approach, a two-stage model was proposed, in which SNP were selected in the first stage based on marginal effects and then tested for interactions in the subsequent stage [Marchini et al., 2005]. This approach could, however, result in the failure to detect important regions of the genome in the first stage of the model. As such, there is a need for methodologies capable of identifying important genomic regions in the presence of potential gene interactions when large numbers of markers are genotyped.

One 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 a wide range of problems [Shymygelska and Hoos, 2005; Ding et al., 2005]. Through the evaluation of groups of loci efficiently selected from different regions of the genome, optimization algorithms should be able to account for potential interactions.

In this section, a modified ACA, enabling the use of permutation testing for global significance, was combined with logistic regression and implemented on a simulated binary trait under the influence of interacting genes. The performance of the ACA was evaluated and compared to models accounting for only marginal effects.

B.1 Logistic regression: Groups of SNP markers were evaluated based in haplotype genotype effects 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:

$$lor_i = \ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{X}_i\boldsymbol{\beta} + e_i \tag{4}$$

where P_i = probability ($y_i = 1$) and X is a matrix containing indicator variables for the haplotypes formed from the selected SNP. Groups of SNP markers with less than two corresponding observations were discarded, and analysis was conducted on all remaining marker groups.

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

$$p_i(y_i=0) = \frac{1}{1+\exp(\boldsymbol{X}_i\boldsymbol{\beta})} \text{ and } p_i(y_i=1) = \frac{\exp(\boldsymbol{X}_i\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}_i\boldsymbol{\beta})} \tag{5}$$

yielding the following relationships:

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

B.2 Marginal effects model: The genotype and haplotype association methods were implemented using R functions developed by [Gonzalez et al., 2007; Sinnwell and Schaid, 2005]. The haplotype analysis was implemented using a sliding window approach which utilizes a window of *k* SNP in width sliding across the genome *h* SNP at a time. Individual SNP scores were determined as the maximum average of all haplotypes containing a given SNP.

B.3 Ant colony algorithm: While the algorithm, in the aforementioned form can be used to subjectively identify markers, it is not well suited for the calculation of permutation p-values. When updating the pheromone function, as previously described in equation (2), the final pheromone levels are relative not only to prediction accuracy, but the number of times a SNP marker is selected. As a result, the amount of pheromone deposited on a feature depends greatly on the amount of pheromone deposited on all other SNP markers and can vary wildly from permutation to permutation. One obvious solution to this problem is to use the average accuracy of all S_k containing genotypes for SNP *m*; however, this approach substantially reduces the ACA's ability to efficiently burn in on good solutions, an attribute needed to detect unknown gene interactions in high-dimension data sets.

To overcome these limitations, a two-layer pheromone function was developed:

$$P_m(t) = \frac{\tau_m(t)^{\alpha} \tau 2_m(t)^{\alpha 2} \eta_m{}^{\beta}}{\sum_{m=1}^{nf} \tau_m(t)^{\alpha} \tau 2_m(t)^{\alpha 2} \eta_m{}^{\beta}} \quad (6)$$

where $\tau_m(t)$ is the first pheromone layer updated using the sum of accuracies for all S_k containing SNP m; $\tau 2_m(t)$is the second pheromone layer updated using the average accuracy of all S_k containing genotypes for SNP m; andη_m, α, $\beta$are as previously described. For the current study, $\alpha$and $\alpha 2$ were set to 1, $\beta$was set to.3 and the prior information (η_m) was the prediction the accuracy of SNP marker m, obtained using logistic regression on genotypes.

The pheromone for $\tau_m(t)$ was updated using equation (2) and $\tau 2_m(t)$ was updated using the following equation:

$$\tau 2_m(t+1) = [t * \tau_m 2(t) + \Delta \tau_m 2(t)] / (t + ns) \quad (7)$$

where t is the iteration number; $\Delta \tau_m 2(t)$is the change in pheromone level for feature m based on the sum of accuracy of all S_k containing genotypes for SNP m, and is set to zero if feature m was not selected by any of the artificial ants; and ns is the number of times SNP m was selected at iteration t. Permutation p-values were calculated using $\tau 2_m(t)$ only.

The procedure can be summarized in the following steps:

1. Each ant selects a predetermined number of SNP markers.
2. Using the selected SNP markers, accuracies are computed using logistic regression on haplotypes or genotypes.
3. The pheromone for each selected group of SNP, S_k, is calculated as:

$$pheromone_k = acc^{(1-acc)} \quad (8)$$

1. The change in pheromone at time t is then calculated using equations (2) and (7).
2. Following the update of pheromone levels according to equations (2) and (7), the PDF is updated according to equation (6) and the process is repeated until pheromone levels have converged.

B.4 Data simulation: Genotype data on 90 unrelated individuals from the Japanese and Han Chinese populations were downloaded from the HapMap ECODE project website. Each simulation scenario was replicated five times using two 500 Kbp regions on chromosome 2, comprising 2047 polymorphic SNP. All SNP haplotypes were assumed to be known without error. The binary disease trait was simulated under a two locus epistatic model as seen in Table 7.

	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

Table 7. Relative risk for simulated trait (relative to the aa/bb genotype)

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%. A plot illustrating the LD of all SNP with the two causative mutations is shown in Fig (1). The plot shows a large peak of high LD with rs2049736 (SNP 409), while the peak of high LD with rs28953468 (SNP 2041) is substantially narrower, and is preceded by a plateau of SNP in moderate LD with rs28953468.

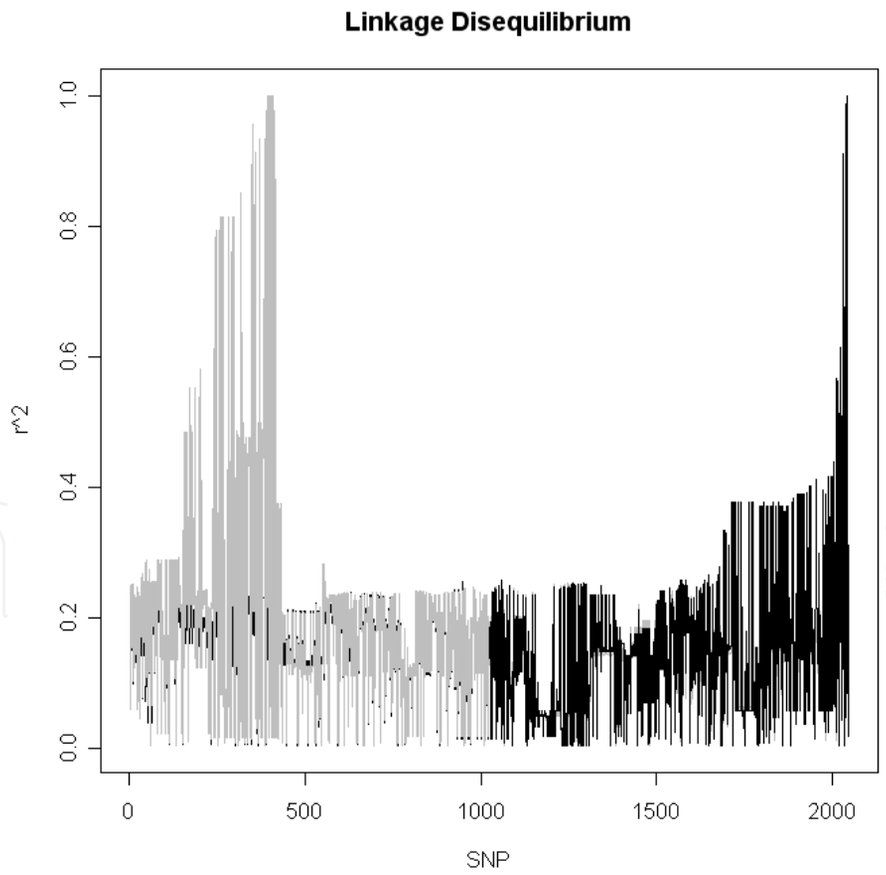

Figure 1. Plots of each marker's linkage disequilibrium (LD) with the two causative mutations. The light grey line represents LD with the causative mutation located at position 409. The black line represents LD with the causative mutation located at position 2041.

Permutation testing was used to access global significance for all models used in the study. Statuses were randomly shuffled amongst subjects, with haplotype effects, genotype effects and association p-values re-estimated for each new configuration of the response variables. The largest estimated haplotype/genotype effect or the smallest haplotype/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 ($r^2 \geq .80$) with a causative mutation.

B.5 Results and discussions: Estimates of power for the three methods can be found in Table 8. Methods employing the ACA showed substantial increases in power when compared to the methods accounting for only marginal effects. Due to the fact that the trait was simulated under a dominance model, analysis of genotypes yielded superior results when compared to haplotype analysis. Despite the inherent advantage of genotype analysis using a dominance model, the ACA using haplotypes (ACA/H) still showed greater power than RG/D in both scenarios. For scenario 2, all models showed a reduction in power; however, the superiority of the ACA methodologies remained constant, with the ACA using LG on genotypes assuming a dominance model (ACA/G/D) yielding 66.7% increase in power for both scenarios when compared to the next best method, RG/D.

	Scenario 1			**Scenario 2**		
	1 locus	2 locus	3 locus	1 locus	2 locus	3 locus
ACA/G/D	___	1.00	0.90	___	0.50	0.40
ACA/G/C	___	0.70	0.80	___	0.40	0.40
ACA/HAP	___	0.60	0.70	___	0.50	0.40
RG/D	0.60	___	___	0.30	___	___
RG/C	0.30	___	___	0.30	___	___
SW/HAP	___	0.10	0.20	___	0.00	0.00

[a] Power was calculated as the proportion of times at least one SNP in high linkage disequilibrium (>.8) with a causative mutations was detected by the model at α=.05 for genome-wide significance

Table 8. Power calculations[a].

Plots of the associative effects, obtained using SW/H, ACA/G/D, and RG/D, are shown in Fig. (2) and (3). When compared to the LD plot (Fig. (1)) all methods show good correspondence for scenario 1, though only the ACA/G/D was able to identify markers for both causative mutations in all replicates. In scenario 2, where the genetic effect was greatly reduced, plots of associative effects tended to be noisier for all models, with the ACA/G/D again showing superior performance, identifying several SNP markers having only moderate LD with causative mutation rs28953468.

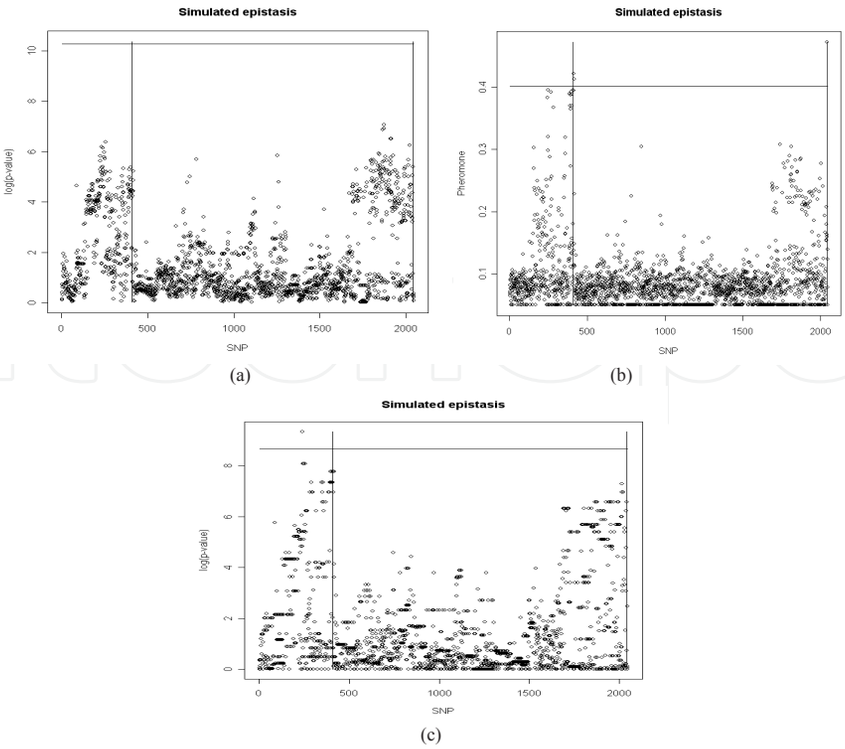

Figure 2. Association plots of SNP markers for the simulated trait under scenario 1. Plots were obtained using 2 SNP haplotypes analyzed by a. SW/LR and b. ACA/LR. Vertical lines represent the position of the two causative mutations, and horizontal lines represent the threshold at which associations are significant at α=. 05

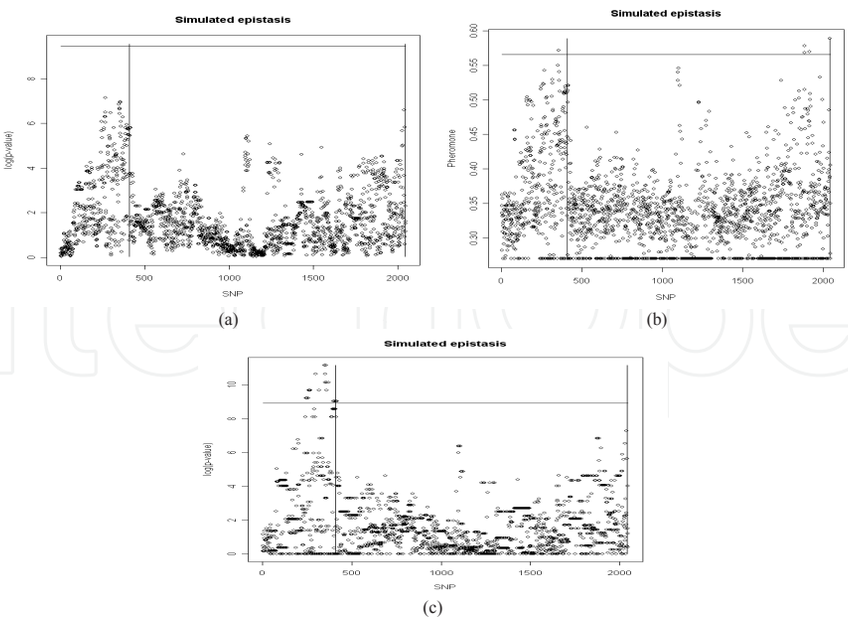

Figure 3. Association plots of SNP markers for the simulated trait under scenario 2. Plots were obtained using 3 SNP haplotypes analyzed by a. SW/LR, b. ACA/LR, and c. RG. Vertical lines represent the position of the two causative mutations, and horizontal lines represent the threshold at which associations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

To determine the effectiveness of the permutation on pheromone levels, the cumulative distribution, based on LD with causative mutations, of SNP identified as being significantly associated with simulated trait by ACA/G/D and RG/D were plotted and can be found in Fig. (4). Despite similarities in the average number of SNP identified by ACA/G/D (15.4) and RG/D (22), the distributions of these SNP, differed substantially. In contrast to RG/D, the ACA/G/D identified a large number of SNP having LD between.35-.45. These SNP corresponded to the broad plateau of SNP in LD with SNP 2041. Unlike RG/D, the ACA/G/D also identified several SNP (5.19%) having less than.10 LD with either of the causative mutations, an unexpected result given the strict family-wise significance thresholds (α=0.05) imposed on all models. Surprisingly, both methodologies identified a large number of SNP having LD of approximately ~.2. Upon closer examination it was found that these SNP had LD of ~. 2 with both causative mutations, likely artifacts of the data resulting from the relatively small sample size. The LD with both causative mutations imparted a portion of the epistatic effect on these SNP, resulting in significant associations with the simulated traits.

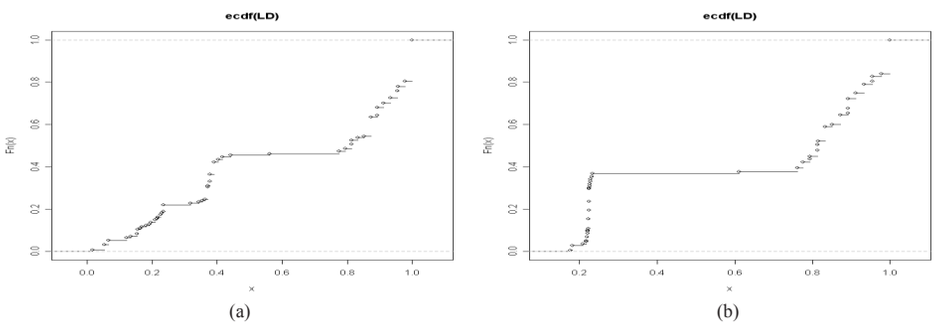

Figure 4. Plot of the cumulative distribution of SNP, identified as have significant associations when using a) ACA/G/D using 2 loci model (5.19%) b) RG/D, based on linkage disequilibrium with the causative mutations

2.3. Ant colony optimization as a method for strategic genotype sampling

Interest in identifying QTL of economic importance for marker-assisted selection (MAS) in livestock populations has increased greatly in the past decade. Yet, it may not be viable to genotype each animal due to cost, time or lack of availability of DNA. A method that would allow for a selected sample (e.g. 5%) of the population to be genotyped and at the same time inferring with high probability genotypes for the remaining animals in the population could be beneficial. By using such a method, fewer animals in a population would be needed for genotyping which would decrease the time and cost of genotyping. Theoretically the problem at hand is simple to solve. If it were possible to evaluate every possible subset of animals equal to the desired size (e.g. 5%) then the optimal solution could be found. However, this is computationally impossible at the current time. Consequently a more feasible solution is needed. An intuitive solution would be one that selects animals based on their relationship with other animals in the pedigree. However, the heterozygosity and the structure of the pedigree play important roles as well. Consequently, the problem is one of optimization.

In the case of genotyping, the ACA should select a subset of animals that, when genotyped, should give an optimal performance in terms of extrapolating the alleles of non-genotyped animals. Therefore, the objectives were to investigate the usefulness of a search algorithm as implemented by Ressom *et al.* (2006) to optimize the amount of information that can be extracted from a pedigree while only genotyping a small portion. The results of the proposed method are compared to other viable methods to ascertain any potential gain. The procedures were tested using simulated pedigrees and actual beef cattle pedigrees of varying sizes and structures.

C.1 Ant colony optimization: The ACA is initialized with all features having an equal baseline level of pheromone which is 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.

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. Upon convergence the optimal subset of features is select based in the level of pheromone trail deposited on each feature.

In the specific case of selecting individuals for genotyping, the features are candidate animals for genotyping from a full or partial pedigree. The pheromone of some feature, m, in the current study was proportional to the sum of an animal's number of mates and number of offspring

$$\tau_m(t) = numoff_m + nummate_m \tag{9}$$

where $numoff_m$ and $nummate_m$ were the number of offspring and number of mates for animal m at time t, respectively. Consequently, the performance of a particular subset, S_k, is determined the by the cumulative sum as described above for each of n animals in the subset.

$$\tau_m(t) = \sum_{m=1}^{n} numoff_m + nummate_m \tag{10}$$

Outside of actual ant colonies, and with regard in particular to the current study, it is difficult to assign a biological explanation to the evaporation rate orρ. Consequently, a relatively small value of 0.01 was chosen in an attempt to reach convergence faster. For each of j artificial ants, a subset of animals was chosen equal to approximately 5% of the pedigree size.

For the five replicates of simulated pedigrees, 100 ants were used for each of 30,000 iterations. The evaporation rate was set equal to 0.01. The criterion used for evaluating candidates was a function of their number of mates and number of offspring. Each animal in the pedigree was randomly assigned to be either homozygous or heterozygous. The probability of an animal being assigned to one of these two groups was dependent on

the allelic frequencies such that if the allele frequencies were assumed to be 0.7/0.3 then approximately 58% of the animals would be categorized as homozygous based off of Hardy-Weinberg Laws of equilibrium. The assignment of homozygous/heterozygous status was performed each iteration. If a selected animal 2was homozygous then his/her number of mates and number of offspring were corrected such that for every homozygous offspring he/she had the number of offspring was corrected accordingly so that the number of offspring only reflected the number of heterozygous offspring. The same correction was done for the number of mates. Similarly, if a selected animal was heterozygous, the number of offspring and the number of mates reflected a count of only homozygous individuals. An animal's probability of being selected was based off of maximizing the corrected sum of the animal's number of offspring and number of mates. The accuracy for evaluating a selected group of animals was proportional to this corrected sum. The uncorrected or original sum of each animal was used as prior information. Selected animals were chosen based off of their cumulative probability were assumed to have known genotypes for the peeling procedure. Simulated allele frequencies of 0.7/0.3 and 0.5/0.5 were used to assign genotypes to the animals in the pedigree.

In the case of the real pedigree the same parameters were used as in the simulated pedigrees with the following exceptions; 100 ants were used for each of 5,000 iterations. The top 1,455 animals out of 29,101 were selected (5% of the total pedigree) based off of their cumulative probability were assumed to have known genotypes for the peeling procedure. In the case of the research beef cattle pedigree, 100 ants were used for each of 20,000 iterations. The top 434 out of 8,688 animals were selected (5% of the total pedigree) based on the same criteria.

C.2 Peeling: Given that genotypes in this study were assigned at random in the population, it is possible to extract additional genotypic information from the pedigree. Animals with missing genotypic information can be assigned one or both alleles given parental, progeny, or mate information. Given this trio of information sources and following an algorithm similar to Qian and Beckmann (2002) and Tapadar et al. (2000), imputation on missing genotypes were made and additional genotypic information was garnered. For the current study it was assumed that there were no errors in the recorded pedigree resulting in all animals having known paternity and maternity. Whenever possible, maternal and paternal alleles were identified based on the inheritance. For the purpose of this study, the first allele was inherited from the sire and the second allele was inherited from the dam. If the parental origin of an allele was unclear, then allele was arbitrarily assigned as either the paternal or maternal allele.

After the peeling process, the number of animals with one or two alleles known was computed. This was done by simply counting the number of animals that were assigned either one or two alleles based on the peeling procedure described above. The percentage of alleles known based on the peeling procedure (AK_P) was then computed as follows:

$$AK_P = \left(\frac{(n_1 \times 2) + n_2}{n_a \times 2} \right) \times 100, \tag{11}$$

where n_1 and n_2 were the number of animals with 2 and 1 allele(s) known and n_a was the total number of animals in the population. Furthermore, $n_1$and n_a were multiplied by two since each animal has two alleles.

At the end of the peeling process those animals that had either one or two alleles known were retained for further analysis to determine the remaining unknown alleles in the population. In other words, those animals having one or two known alleles were used as prior information in the Gibbs sampling procedure for determining the remaining unknown alleles in the population.

C.3 Gibbs sampling: After the known alleles were determined by the peeling process described above, these alleles were used as prior information in the Gibbs Sampler to assign genotypes to the remaining animals in the population. For the base population animals, the unknown allele(s) were randomly sampled given the frequency of alleles in the population and the assumption of Hardy-Weinberg equilibrium. Unknown alleles for non-base population animals were randomly sampled from the parent's genotypes according to Mendelian rules. An equal weight was assumed for inheriting either the first or second allele from a parent. For a non-base population animal that had only one unknown allele, the unknown allele was sampled approximately half of the time from the sire's genotype and the remaining time from the dam's genotype. This was to compensate for incorrect assignment of the known allele as illustrated in the above example.

At the end of the sampling process, a benefit function that described the total number of alleles known in the population was computed. This function was computed from a combination of known alleles and the probability of unknown alleles assigned during the sampling process. In order to be included in the benefit function, an allele in a particular position had to be equal to the true allele of the same position (i.e., *Bb* and *bB* were not equal). The probability of allele$a_{i,j}$, (j = 1 or 2) being assigned as the true allele j for animal i was calculated as:

$$p(a_{i,j}) = \frac{\text{number of times } a_{i,j} \text{ was assigned}}{\text{number of iterations}}. \tag{12}$$

Using $p(a_{i,j})$ and the number of known alleles, the benefit function was then computed as

$$Benefit = n_1 \times 2 + \sum_{i=1}^{n_2}[1 + p(a_{i,j})] + \sum_{i=1}^{n_3}[p(a_{i,1}) + p(a_{i,2})], \tag{13}$$

wheren_1, n_2, and n_3 were the number of animals with 2, 1 or 0 alleles known, respectively, and $p(a_{i,j})$ as previously defined. The percentage of alleles known after the Gibbs sampling process, AK_G, was such that

$$\mathrm{AK_G} = \left(\frac{benefit}{n_a \times 2} \right) \times 100, \tag{14}$$

where *benefit* was the benefit function computed above and n_a was the total number of animals in the population.

During each round of the sampling process only one genotype of a given animal was assigned as the true genotype. Thus, at the end of the sampling process every animal had a probability of having the true genotype,PTG_{ig} , assigned as

$$\mathrm{PTG}_{ig} = \frac{\text{number of times genotype } g \text{ was assigned}}{\text{total number of samples}}, \tag{15}$$

where genotype g was the true genotype for animal$i$. The average probability of the true genotype being identified for every animal in the population (APTG) was computed using the following:

$$\mathrm{APTG} = \frac{\sum_{i=1}^{n_a} \mathrm{PTG}_{ig}}{n_a}, \tag{16}$$

where PTG_{ig} was defined as above and n_a was the total number of animals in the population. In contrast to the benefit function, APTG only required that the animal have the correct genotype—*Bb* was considered the same genotype as *bB*—and therefore was able to compensate for the incorrect allele position and sampling the correct unknown allele.

C.4 Simulation: A simulation using an animal model was carried out to investigate two methods of selecting animals for genotyping and two methods of maximizing the genetic information of the population. A pedigree with four over-lapping generations was simulated. The base population included 500 unrelated animals and subsequent generations consisted of 1,500 animals with a total of 5,000 animals generated. For the simulated pedigrees as well as the real pedigrees, one gene with two alleles was simulated for every animal in the pedigree file. Genotypes of the base population animals were assigned based on allele frequencies. For the subsequent generations, genotypes were randomly assigned using the parent's genotype, where an equal chance of passing either the first or second allele was assumed. Five replicates of the simulated data were generated.

Two different frequencies for the favorable allele were used in the simulation and analyses. The frequencies were 0.30, and 0.50. For the analyses using Gibbs sampling, a total chain length of 25,000 iterations of the Gibbs sampler was run, where the first 5,000 iterations were discarded as burn-in.

C.5 Results of simulated pedigrees: Table 9 presents results of the ACO and alternative methods for analysis of the simulated pedigrees (Spangler 2008). The ant colony optimization method (ACO) appeared to be the most desirable method of those discussed in the current study. Compared to selecting 5% of the animals at random, ACO showed gains in AK_P, AK_G, and APTG ranging from 261.09 to 262.93%, 19.97 to 26.04%, and 23.5 to 29.6%, respectively. As compared to the favorable method of the alternative approaches, selecting males and females based of off the diagonal element of the inverse of the relationship matrix, the increase in AK_P ranged from 4.98 to 5.16%. This gain is due to the amount of animals with both alleles known after the peeling process which was between 20.74 and 21.07% larger in favor of ACO. Admittedly, the gains in AK_G were slight as compared to selecting males and females based of off the diagonal element of A^{-1}, yet ACO still performed better. The increase in APTG ranged from 1.6 to 1.8% in favor of ACO over selecting males and females from their diagonal element.

	ACO		Random		Males		Males and females	
Parameter[b]	(0.30)	(0.50)	(0.30)	(0.50)	(0.30)	(0.50)	(0.30)	(0.50)
No. of animals with								
2 alleles known	811.20	787.20	258.20	259.60	250.00	250.60	670.00	652.00
1 allele known	2,166.80	2,063.00	527.80	485.60	2,939.80	2,793.00	2,262.60	2,152.80
Benefit function	8,055.01	7,550.36	6,713.56	6,007.02	7,943.67	7,401.57	8,019.88	7,497.70
AK_P	37.89	36.29	10.44	10.05	34.40	32.94	36.03	34.57
AK_G	80.55	75.71	67.14	60.07	79.44	74.02	80.20	74.98
APTG	0.63	0.57	0.51	0.44	0.59	0.52	0.62	0.56

[a] Random= 5% selected at random, Males= 5% of males selected from their diagonal element of A^{-1}, Males and females= 2.5% males and 2.5% females selected from their diagonal element of A^{-1}. Numbers in parenthesis are the true allele frequencies used in the simulation. [b] Descriptions of the parameters can be found in equations 5-10

Table 9. Number of animals with one or two alleles known, percentage of alleles known, and probability of assigning the true genotype using other approaches[s]

C.6 Real beef cattle pedigree: Results from the ACO analysis can be found in Table 10 along with results from alternative approaches. The largest gains were seen in AK_P which ranged from 150.00 to 171.62%, 2.95 to 3.04%, and from 1.80 to 1.94% as compared to random selection, selection of males and females from A^{-1}, and selection of males from A^{-1}, respectively. ACO also showed gains in AKG and APTG over random selection between 70.06 and 74.91% and between 14.3 and 15.4%, respectively. Table 3 shows advantages, although slight, of ACO over the methods using the diagonal element of A^{-1} for the parameters of AK_G and APTG.

C.7 Research beef cattle pedigree: Results from the ACO analysis and other approaches using the same pedigree can be found in Table 11. As compared to randomly selecting 5%

of the animals, ACO showed increases in AK_P, AK_G, and APTG ranging from 241.24 to 302.58%, 42.93 to 43.17%, and 20.9 to 38.0%, respectively. Realized gains in AK_P of ACO over selecting males from A^{-1} or males and females from A^{-1} ranged from 8.78 to 10.15%, and 2.04 to 3.40%, respectfully.

The results suggest that ACO is the most desirable method of selecting candidates for genotyping, particularly after peeling (AK_P). From these results it appears that the number of offspring and the number of mates along with the homozygosity of the genotyped animals is critical in the selection process. Consequently, in application it will be critical to have good estimates of allele frequencies prior to implementing the genotype sampling strategy proposed in the current study. Differences in performance of ACO do exist between the pedigrees explored in the current study. This is due to the proportion of sires and dams that have large numbers of offspring and/or mates. In the dairy industry, for example, there may be only a small number of sires in a pedigree but they may all be used heavily as in the case of the simulated pedigrees in the current study. In contrast, a pedigree from the beef industry may have a larger proportion of sires but a large number of them may be used less frequently.

	ACO		**Random**		**Males**		**Males and females**	
Parameter[b]	(0.30)	(0.50)	(0.30)	(0.50)	(0.30)	(0.50)	(0.30)	(0.50)
No. of animals with								
2 alleles known	1,767.00	1,706.00	1,505.00	1,501.00	1,473.00	1,470.00	2,086.00	1,999.00
1 allele known	11,451.00	10,382.00	2,508.00	2,144.00	11,756.00	10,607.00	10,376.00	9,398.00
Benefit function	34,977.61	32,547.06	20,569.53	18,609.00	34,876.62	32,282.40	34,005.21	31,456.36
AK_P	25.75	23.70	9.48	8.84	25.26	23.28	24.99	23.02
AK_G	60.10	55.92	35.34	31.97	59.92	55.47	58.43	54.05
APTG	0.45	0.40	0.39	0.35	0.44	0.39	0.44	0.40

[a] Random= 5% selected at random, Males= 5% of males selected from their diagonal element of A^{-1}, Males and females= 2.5% males and 2.5% females selected from their diagonal element of A^{-1}. Numbers in parenthesis are the true allele frequencies used in the simulation. [b] Descriptions of the parameters can be found in equations 5-10.

Table 10. Number of animals with one or two alleles known, percentage of alleles known, and probability of assigning the true genotype using other approaches from a real beef cattle pedigree [a]

	ACO		Random		Males		Males and females	
Parameter[b]	(0.30)	(0.50)	(0.30)	(0.50)	(0.30)	(0.50)	(0.30)	(0.50)
No. of animals with								
2 alleles known	975.00	720.00	452.00	458.00	438.00	439.00	1,082.00	751.00
1 allele known	5,101.00	4,009.00	847.00	682.00	5,525.00	4,132.00	4,747.00	3,768.00
Benefit function	13,916.18	11,990.71	9,719.53	8,284.42	14,113.18	12,017.80	13,743.44	11,848.01
AK_P	40.58	31.36	10.08	9.19	36.84	28.83	39.77	30.33
AK_G	80.09	68.15	55.94	47.68	81.22	69.16	79.09	68.19
APTG	0.69	0.52	0.50	0.43	0.69	0.51	0.68	0.52

[a] Random= 5% selected at random, Males= 5% of males selected from their diagonal element of A^{-1}, Males and females= 2.5% males and 2.5% females selected from their diagonal element of A^{-1}. Numbers in parenthesis are the true allele frequencies used in the simulation. [b] Descriptions of the parameters can be found in equations 5-10.

Table 11. Number of animals with one or two alleles known, percentage of alleles known, and probability of assigning the true genotype using other approaches from a real beef cattle research pedigree[a]

Furthermore, pedigrees from field data or from research projects will also have innate structural differences. Research projects may be limited by the size of the population and thus only use a small number of sires. In this scenario it would also be possible for higher rates of inbreeding and larger numbers of loops in a pedigree due to a large number of full sibs.

In the current study, the simulated pedigrees are composed of approximately 10% sires, while the large beef cattle pedigree and the small research beef cattle pedigree contain approximately 16 and 7% sires, respectively. Intuitively, as the proportion of sires goes up, the number of offspring per sire goes down. This explains the similarity of the results between the simulated pedigrees and the small research pedigree. Thus, it is expected that the ACO algorithm will be far superior to other alternatives when very small (few hundred animals) pedigrees are considered or in situations where more than 5% of animals are genotyped due to reduction in animal with large diagonal elements in A^{-1}.

Ant colony optimization offers a new and unique solution to the optimization problem of selecting individuals for genotyping. The heuristics used in the current study such as the number of ants, number of iterations, and the evaporation rate are unique only to the pedigrees used in the current study. Each pedigree will offer a different structure and thus require a different set of parameters.

3. Conclusions

When applied to the 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 accura-

cies without the need to pre-select a small percentage of genes. Furthermore, the ant colony algorithm was able to identify small subsets of features with high predictive abilities and biological relevance. In the presence of simulated epistasis, 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. Although the ACA methods identified more SNP markers that could be construed as false positives, the use of a more stringent threshold eliminated the problem without greatly reducing the advantage of the ACA, in terms of power, when compared to other methods. The results of this study provide compelling evidence that the ACA is capable of efficiently modeling complex biological problems, such as the model proposed in this study.

Author details

R. Rekaya[1,2,3*], K. Robbins[4], M. Spangler[5], S. Smith[1], E. H. Hay[1] and K. Bertrand[1]

*Address all correspondence to: rrekaya@uga.edu

1 Department of Animal and Dairy Science, The University of Georgia, Athens, Greece

2 Department of Statistics, The University of Georgia, Athens, Greece

3 Institute of Bioinformatics, The University of Georgia, Athens, Greece

4 Dow AgroSciences, Indianapolis, IN, USA

5 Animal Science Department, University of Nebraska, Lincoln, NE, USA

References

[1] Akiyama,T. and Y Kawasaki (2006) Wnt signaling and the actin cytoskeleton *Oncogene*, 25, 7538-7544.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Dorigo M., Di Caro G. & Gambardella L.M. (1999) Ant algorithms for discrete optimization. Artificial Life 5, 137–72.

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

[10] 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.

[11] 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

[12] Hong,J. and S. Cho (2006) Efficient huge-scale feature with speciated genetic algorithm *Pattern Recognition Lett*., 27, 143-150.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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

[19] 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.

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

[21] Qian D. & Beckmann L. (2002) Minimum-recombinant haplotyping in pedigrees. American Journal of Human Genetics 70, 1434–45.

[22] 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 *PNAS,* 98, 15149-15154.

[23] Rekaya, R, K. Robbins. (2009). Ant colony algorithm for analysis of gene interaction in high-dimensional association data. Revista Brasileira de Zootecnia. doi: 10.1590/S1516-35982009001300011.

[24] 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.

[25] Robbins, K. R., Zhang, W., R. Rekaya, and J. K. Bertrand. 2007. The use of the ant colony algorithm for analysis of high-dimension gene expression data sets. 58th Annual Meeting of the European Association for Animal Production (EAAP):167.

[26] Robbins, K. R., Zhang, W., J. K. Bertrand, R. Rekaya. 2008. Ant colony optimization for feature selection in high dimensionality data sets. Math Med Biol. 24(4):413-426.

[27] Robbins K, K. Bertrand, and R. Rekaya. 2011. The use of the ant colony algorithm for the detection of marker associations in the presence of gene interactions. International Journal of Bioinformatics Research, 2:227-235.

[28] 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.

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

[30] 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.

[31] Sinnwell, J. P. and D. J. Schaid. 2005. haplo.stats: Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous. R package version 1.2.2.

[32] Spangler, M. L., K. R. Robbins, J. K. Bertrand, M. MacNeil, and R. Rekaya. 2008. Ant colony optimization as a method for strategic genotype sampling. Animal Genetics 40: 308 – 314.

[33] Subramani,P., Sahu,R. and S. Verma, Feature selection using Haar wavelet power spectrum *BMC Bioinformatics,* 7:432.

[34] Tapadar P., Ghosh S. & Majumder P.P. (2000) Haplotyping in pedigrees via a genetic algorithm. Human Heredity 50, 43–56.

[35] West M. (2003) Bayesian factor regression models in the "Large p, Small n" paradigm, *Bayesian Statistics*, 7, 723-732.

[36] 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.

IntechOpen