Ant colony optimization as a method for strategic genotype sampling

M. L. Spangler[1*], K. R. Robbins*, J. K. Bertrand[*], M. MacNeil§, and R. Rekaya[2*†‡],
[*]Animal and Dairy Science Department, [†]Department of Statistics, [‡]Institute of Bioinformatics, University of Georgia, Athens, Georgia 30602-2771; and §USDA-ARS, Fort Keogh Livestock and Range Research Laboratory, Miles City, MT 59301

INTRODUCTION

A method that could select a 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. 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. Ant colony algorithms (ACA) were proposed by Dorigio et al. (1999) to solve difficult optimization problems such as the traveling salesman. 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. 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 (Dorigio et al. 1999, Ressom et al. 2006). Therefore, the objectives of the current study 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.

MATERIALS AND METHODS

Ant colony optimization: The ACA, as defined by Dorigio et al. (1999) and Ressom et al. (2006), is a group of parallel units with a common memory in the form of a PDF, where the probability of sampling feature *m* at time *t* is defined as:

$$P_m(t) = \frac{(\tau_m(t))^{\alpha} \eta_m{}^{\beta}}{\sum_i (\tau_m(t))^{\alpha} \eta_m{}^{\beta}} \qquad (1)$$

where $\tau_m(t)$ is the amount of pheromone for feature *m* at time *t*; η_m is some form of prior information on the expected performance of feature *m*; α and β are parameters determining the weight given to pheromone deposited by ants and a priori information on the 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 k=1,…,j.

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. Consequently, the performance of a particular subset, S_k, is determined the by:

[1] Current address: University of Tennessee at Martin, Martin, Tennessee 38237, U.S.A.
[2] Corresponding author: phone: (706) 542-0949; fax: (706) 583-0274; E-mail: rrekaya@uga.edu.

$$\tau_m(t) = \sum_{m=1}^{n} numoff_m + nummate_m \qquad (3)$$

A relatively small value of 0.01 was chosen as the evaporation rate 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. 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 dependant on the allelic frequencies. The assignment of homozygous/heterozygous status was performed each iteration. An animal's probability of being selected was based on maximizing the corrected sum of the animal's number of offspring and number of mates. The uncorrected or original sum of each animal was used as prior information. 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 field data 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 on the pheromone deposited by the artificial ants and were assumed to have known genotypes for the peeling procedure. In the case of the research pedigree, 100 ants were used for each of 20,000 iterations. The top 434 out of 8,688 animals were selected based on the same criteria.

Peeling: 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.

After the peeling process, the number of animals with one or two alleles known was computed. The percentage of alleles known based on the peeling procedure (AK_P) was then computed as follows:

$$\mathrm{AK_P} = \left(\frac{(n_1 \times 2) + n_2}{n_a \times 2} \right) \times 100, \qquad (4)$$

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.

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. The probability of allele $a_{i,j}$, (*j* = 1 or 2) being assigned as the true allele j for animal *i* was calculated as:

$$p(a_{i,j}) = \frac{\text{number of times } a_{i,j} \text{ was assigned}}{\text{number of iterations}}. \qquad (5)$$

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

$$Benefit = n_1 \times 2 + \sum_{i=1}^{n_2} [1 + p(a_{i,j})] + \sum_{i=1}^{n_3} [p(a_{i,1}) + p(a_{i,2})], \qquad (6)$$

where n_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, $\mathrm{AK_G}$, was such that

$$\mathrm{AK_G} = \left(\frac{benefit}{n_a \times 2} \right) \times 100, \qquad (7)$$

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

$$\mathrm{PTG}_{ig} = \frac{\text{number of times genotype } g \text{ was assigned}}{\text{total number of samples}}, \qquad (8)$$

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:

$$\text{APTG} = \frac{\sum_{i=1}^{n_a} \text{PTG}_{ig}}{n_a}, \qquad (9)$$

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 and therefore was able to compensate for the incorrect allele position and sampling the correct unknown allele.

Simulation: 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. For the analyses using Gibbs sampling, a total chain length of 25,000 iterations was run, where the first 5,000 iterations were discarded as burn-in.

RESULTS

Table 1. Results from simulated pedigrees[1]

	ACO		A^{-1}	
	True allele frequency		True allele frequency	
Parameter	0.30	0.50	0.30	0.50
No. of animals with				
1 allele known	2,166.80	2,063.00	2,262.60	2,152.80
2 alleles known	811.20	787.20	670.00	652.00
Benefit function	8,055.01	7,550.36	8,019.88	7,497.70
AK_P	37.89	36.29	36.03	34.57
AK_G	80.55	75.71	80.20	74.98
APTG	0.63	0.57	0.62	0.56

[1] Results are the average of 5 replicates; A^{-1} is from Spangler et al., 2007.

Simulated pedigrees: The results can be found in Table 1. Results from using the inverse of the relationship matrix (A^{-1}) are from Spangler et al. (2007). As compared to 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.

Table 2. Results from the field data pedigree[1]

	ACO		A^{-1}	
	True allele frequency		True allele frequency	
Parameter	0.30	0.50	0.30	0.50
No. of animals with				
1 allele known	11,451.00	10,382.00	11,756.00	10,607.00
2 alleles known	1,767.00	1,706.00	1,473.00	1,470.00
Benefit function	34,977.61	32,547.06	34,876.62	32,282.40
AK_P	25.75	23.70	25.26	23.28
AK_G	60.10	55.92	59.92	55.47
APTG	0.45	0.40	0.44	0.39

[1] A^{-1} results are from Spangler et al., 2007.

Field data pedigree: A field data pedigree as described by Spangler et al. (2007) was used to determine the effectiveness of the proposed method in a larger pedigree more representative of what might be encountered in the beef cattle industry. Results can be found in Table 2 along with results from alternative approaches (Spangler et al.,2007). The largest gains were seen in AK_P which ranged from 1.80 to 1.94% as compared to selection of males

from A^{-1}.. Table 2 shows advantages, although slight, of ACO over the methods using the diagonal element of A^{-1} for the parameters of AK_G and APTG.

Table 3. Results are from the research pedigree[1]

	ACO		A^{-1}	
	True allele frequency		True allele frequency	
Parameter	0.30	0.50	0.30	0.50
No. of animals with				
1 allele known	5,101.00	4,009.00	4,747.00	3,768.00
2 alleles known	975.00	720.00	1,082.00	751.00
Benefit function	13,916.18	11,990.71	13,743.44	11,848.01
AK_P	40.58	31.36	39.77	30.33
AK_G	80.09	68.15	79.09	68.19
APTG	0.69	0.52	0.68	0.52

[1] A^{-1} results are from Spangler et al., 2007.

Research pedigree: The research pedigree used here has been previously described by Spangler et al. (2007). Results from the ACO analysis can be found in Table 3. Realized gains in AK_P of ACO over males and females from A^{-1} ranged from 2.04 to 3.40%, respectfully.

DISCUSSION

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. 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 and the amount of inbreeding within a given pedigree. 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.

LITERATURE CITED

Dorigo, M., and G. D. Caro, 1999 Ant algorithms for discrete optimization. Artificial Life 5: 137-172.

Qian, D., and L. Beckmann, 2002 Minimum-recombinant haplotyping in pedigrees. Am. J. Hum. Genet. 70: 1434-1445.

Tapadar, P., S. Ghosh, and P. P. Majumder, 2000 Haplotyping in pedigrees via a genetic Algorithm. Hum Hered 50: 43-56.

Ressom, H. W., R. S. Varghese, E. Orvisky, S. K. Drake, G. L. Hortin, M. Abdel-Hamid, C. A. Loffredo, and R. Goldman, 2006 Ant colony optimization for biomarker identification from MALDI-TOF mass spectra. 28th Annual International Conference IEEE Engineering in Medicine and Biology Society (EMBS) SaB03.6.

Spangler, M. L., R. L. Sapp, J. K. Bertrand, M. D. MacNeil, R. Rekaya, 2007 Different methods of selecting animals for genotyping to maximize the amount of genetic information known in the population. J. Anim. Sci. (Submitted).