2009

Ant colony optimization as a method for strategic genotype sampling

Matthew L. Spangler
University of Nebraska-Lincoln, mspangler2@unl.edu

K. R. Robbins
University of Georgia

J. Keith Bertrand
University of Georgia, adshead@uga.edu

M. MacNeil
USDA-ARS

R. Rekaya
University of Georgia, rrekaya@uga.edu

doi:10.1111/j.1365-2052.2008.01835.x

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*,[‡,§]

*Animal and Dairy Science Department, University of Georgia, Athens, GA 30602-2771, USA. [†]USDA-ARS, Fort Keogh Livestock and Range Research Laboratory, Miles City, MT 59301, USA. [‡]Department of Statistics, University of Georgia, Athens, GA 30602-2771, USA. [§]Institute of Bioinformatics, University of Georgia, Athens, GA 30602-2771, USA

Summary

A simulation study was carried out to develop an alternative method of selecting animals to be genotyped. Simulated pedigrees included 5000 animals, each assigned genotypes for a bi-allelic single nucleotide polymorphism (SNP) based on assumed allelic frequencies of 0.7/0.3 and 0.5/0.5. In addition to simulated pedigrees, two beef cattle pedigrees, one from field data and the other from a research population, were used to test selected methods using simulated genotypes. The proposed method of ant colony optimization (ACO) was evaluated based on the number of alleles correctly assigned to ungenotyped animals (AK_P), the probability of assigning true alleles (AK_G) and the probability of correctly assigning genotypes (APTG). The proposed animal selection method of ant colony optimization was compared to selection using the diagonal elements of the inverse of the relationship matrix (A^{-1}). Comparisons of these two methods showed that ACO yielded an increase in AK_P ranging from 4.98% to 5.16% and an increase in APTG from 1.6% to 1.8% using simulated pedigrees. Gains in field data and research pedigrees were slightly lower. These results suggest that ACO can provide a better genotyping strategy, when compared to A^{-1}, with different pedigree sizes and structures.

Keywords ant colony optimization, genotype sampling, search algorithms, simulation.

Introduction

Interest in identifying QTL of economic importance for marker-assisted selection in livestock populations has increased greatly in the past decade. However, it may not be viable to genotype each animal because of cost, time or lack of availability of DNA. A method that could select a subset (e.g. 5%) of the population for genotyping, and at the same time infer the genotypes for the remaining animals in the population with high probability, 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%), the optimal solution could be found. Unfortunately, such an approach is computationally impossible at present, and consequently an optimal solution is needed. Several methods including segregation analysis have been applied to selectively genotype animals in an attempt to reduce genotyping costs (Kinghorn 1999; Macrossan *et al.* 2001). An intuitive approach would be one that selects animals based on their relationship with other animals in the pedigree, such as those suggested by Spangler *et al.* (2008). However, the heterozygosity and the structure of the pedigree also play important roles and therefore must be accounted for in some manner.

Given the limitations of a hard search procedure and the use of animal relationships, an alternative approach, viewing the problem as one of optimization, may be better suited. Although evolutionary algorithms and machine learning have been applied to the issues of group and selective genotyping (Macrossan & Kinghorn 2003a; Kinghorn *et al.* 2006), an optimization technique such as ant colony optimization (ACO) has not been explored. Ant colony algorithms (ACA) were proposed by Dorigo *et al.* (1999) as a

Address for correspondence

R. Rekaya, Animal and Dairy Science Department, University of Georgia, Athens, GA 30602-2771, USA.
E-mail: rrekaya@uga.edu

[1]Present address: University of Nebraska, Lincoln, NE 68583, USA.

Accepted for publication 7 November 2008

means to solve difficult optimization problems such as the travelling salesman problem, 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. In ACA, artificial ants work as parallel units that communicate through a cumulative distribution function (CDF) that is updated by weights, determined by the 'distance' travelled on a selected 'path', which are analogous to the pheromones deposited by real ants (Dorigo *et al.* 1999; Dorigo & Stuetzle 2004; Ressom *et al.* 2007). 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' travelled is some measure of the feature's performance. 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 of the current study were to investigate the usefulness of a search algorithm as implemented by Ressom *et al.* (2007) to optimize the amount of information that can be extracted from a pedigree whilst only genotyping a small portion. The results of the proposed method are compared with 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.

Materials and methods

Overview

A search algorithm was implemented to select candidates for genotyping with preference given to animals that have a large number of offspring and/or mates. The algorithm utilized artificial ants that selected subsets of animals to be genotyped at each iteration. These subsets were then evaluated based on their performance, which was derived by an accuracy function that accounted for their number of mates, number of offspring, and the homozygosity of their mates and offspring. This performance was then added to the pheromone concentration of each animal in the subset. As the pheromone concentration of a particular animal increases, it makes that animal more likely to be chosen by other ants. As the algorithm reaches convergence, ants will 'burn in' on a particular group of animals that have the highest cumulative pheromone concentration. This group of animals, in this case 5% of the pedigree, would then be chosen to genotype. This method of ACO is described in detail below.

Ant colony optimization

The ACA, as defined by Dorigo *et al.* (1999) and Ressom *et al.* (2007), is a group of parallel units with a common memory in the form of a probability distribution function (PDF), where the probability of sampling feature, in this context an animal to be genotyped, m at time t is defined as:

$$P_m(t) = \frac{(\tau_m(t))^{\alpha}\eta_m^{\beta}}{\sum_m (\tau_m(t))^{\alpha}\eta_m^{\beta}}, \quad (1)$$

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.

As a method of foundation sampling, 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. The pheromone level of each feature m in S_k is then updated according to the performance of S_k as:

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

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, $k = 1,\ldots, j$.

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 selected based on 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. In the case where only a subset of animals are genotyped and the remainder are inferred from known genotyped animals, it is logical to choose candidates for genotyping based on some measure of their relationship with other animals. 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.

$$\tau_m(t) = \text{numoff}_m + \text{nummate}_m, \quad (3)$$

where numoff_m and nummate_m were the number of offspring and number of mates for animal m at time t respectively. If any animal with a large sum from equation (3) were to be genotyped, then more knowledge can be gained from the population as a whole due to the relationship (either as a parent or mate) of the genotyped animal and others in the pedigree. Offspring and mates in the equation above were given equal weights. This is because it is possible to infer the missing genotype of an offspring, given knowledge of both parent genotypes, as easily as it is to infer the missing genotype of a mate given the offspring's genotype and the genotype of the other parent.

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.

$$\tau_m(t) = \sum_{m=1}^{n} \text{numoff}_m + \text{nummate}_m. \quad (4)$$

Outside of actual ant colonies, and with regard to the current study, it is difficult to assign a biological explanation to the evaporation rate or ρ. However, the evaporation rate serves as the memory of the algorithm, and a fast evaporation rate will avoid the possibility of accepting local optimums, while a slow evaporation rate will allow for faster convergence. Because of the size and complexity of pedigrees used in the current study, a relatively small value of evaporation rate (0.01) was chosen in an attempt to reach convergence faster. For each of j artificial ants, a subset of animals were 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 a test genotype that was either homozygous or heterozygous. The probability of an animal being assigned to one of these two groups was dependant 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 on Hardy–Weinberg Laws of equilibrium. The assignment of homozygous/heterozygous status was performed at each iteration. If a selected animal was homozygous then his/her number of mates and number of offspring were corrected such that the number of offspring only reflected heterozygous offspring. The same correction was made for the number of mates. These corrections were made with the following rationale: If a selected animal is homozygous then more knowledge can be extracted about missing genotypes if his/her mates/offspring are heterozygous, because it is known with complete certainty what allele the genotyped animal will pass on. 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 on 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 on their cumulative probability and were assumed to have known genotypes for the peeling procedure. Simulated allele frequencies for a single nucleotide polymorphism (SNP) of 0.7/0.3 and 0.5/0.5 were used to assign genotypes to the animals in the pedigree. Admittedly, one could reasonably expect increased performance by using knowledge of the pedigree structure, such that at every iteration and for every ant the chosen set of animals is evaluated based on the amount of genotypic information that can be inferred from genotyped animals. However, this proved to be computationally costly. Additional increases in performance could be expected if selection of an animal is dependent on whether or not a full-sib or other close relative is also selected.

In the case of the field data pedigree (Spangler *et al.* 2008), the same parameters were used as in the simulated pedigrees with the following exceptions: 100 ants were used for each of 5000 iterations. The top 1455 animals 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 (Spangler *et al.* 2008), 100 ants were used for each of 20 000 iterations. The top 434 of 8688 animals were selected (5% of the total pedigree) based on the same criteria.

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 & Beckmann (2002) and Tapadar *et al.* (2000), imputations on missing genotypes were made and additional genotypic information was garnered. Terminal animals, which are parents without known parents themselves and only one offspring, or progeny with only one known parent and no offspring themselves, are temporarily removed (peeled) and all of their genotypic information is transferred to the core of the pedigree, creating another set of terminal animals. This process is repeated until no further genotypic information can be 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 the 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 performed 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:

$$AK_P = \left(\frac{(n_1 \times 2) + n_2}{n_a \times 2}\right) \times 100, \quad (5)$$

where n_1 and n_2 were the number of animals with two and one allele(s) known and n_a was the total number of animals in the population. Due to the assumption of a SNP with two alleles, n_1 and n_a were multiplied by two because each animal has two alleles.

Gibbs sampling

After the known alleles were determined by the peeling process described above, a Gibbs sampler (Fernandez *et al.* 2001; Wang *et al.* 1993; Sorenson *et al.* 1994; Sheehan 2000) was implemented to assign genotypes to the remaining animals in the population using known alleles as prior information. For the base population animals, the unknown allele(s) was(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. Methods of assigning genotype probabilities using segregation analysis without sampling have been described by Thallman *et al.* (2001).

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:

$$p(a_{i,j}) = \frac{\text{number of times } a_{i,j} \text{ was assigned}}{\text{number of iterations}}. \quad (6)$$

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

$$\text{Benefit} = n_1 \times 2 + \sum_{i=1}^{n_2}[1 + p(a_{i,j})] + \sum_{i=1}^{n_3}[p(a_{i,1}) + p(a_{i,2})], \quad (7)$$

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, AK_G, was such that

$$AK_G = \left(\frac{\text{benefit}}{n_a \times 2}\right) \times 100, \quad (8)$$

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

$$PTG_{ig} = \frac{\text{number of times genotype } g \text{ was assigned}}{\text{total number of samples}}, \quad (9)$$

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:

$$APTG = \frac{\sum_{i=1}^{n_a} PTG_{ig}}{n_a}, \quad (10)$$

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 has 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.

Simulation

A simulation using an animal model was carried out to investigate methods of selecting animals for genotyping and methods of maximizing the genetic information of the population. A pedigree with four overlapping generations was simulated. The base population included 500 unrelated animals and subsequent generations consisted of 1500 animals with a total of 5000 animals generated. For the simulated pedigrees as well as the field data and research pedigrees, one SNP 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 favourable 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 5000 iterations were discarded as burn-in.

Results

Simulated pedigrees

The results can be found in Table 1. The ACO method appeared to be the most desirable method of those discussed in the current study. Compared with 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. An intuitive and simplistic method of selecting animals for genotyping would be to select those with larger values for the diagonal element of the inverse of the relationship matrix, because a larger value would indicate more connectedness with other animals in the pedigree. As compared to the favourable method of the alternative approaches, selecting males and females based on 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 number of animals with both alleles known after the peeling process, which was between 20.74% and 21.07% larger in favour of ACO. The increase in APTG ranged from 1.6% to 1.8% in favour of ACO over selecting males and females from their diagonal element.

Field data pedigree

A field data pedigree as described by Spangler *et al.* (2008) was used to determine the effectiveness of the proposed method in a larger pedigree that was 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. 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 2 shows some advantages of ACO over the methods using the diagonal element of A^{-1} for the parameters of AK_G and APTG.

Research pedigree

The research pedigree used here has been previously described by Spangler *et al.* (2008). Results from the ACO analysis can be found in Table 3. 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% respectively.

Discussion

The results suggest that ACO is the more 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

Table 1 Number of animals with one or two alleles known, percentage of alleles known (SD) and probability of assigning the true genotype (SD) from multiple approaches[1] compared to ant colony optimization using simulated pedigrees[2].

	True allele frequency							
	Random		Males		Males and females		Ant colony optimization	
Parameter[3]	0.3	0.5	0.3	0.5	0.3	0.5	0.3	0.5
No. animals with								
2 alleles known	258	260	250	251	670	652	811	787
1 allele known	528	486	2940	2793	2263	2153	2167	2063
Benefit function	6714	6007	7944	7402	8020	7498	8055	7550
AK_P	10.44 (0.007)	10.05 (0.007)	34.40 (0.005)	32.94 (0.005)	36.03 (0.007)	34.57 (0.009)	37.89 (0.006)	36.29 (0.003)
AK_G	67.14 (1.36)	60.07 (0.66)	79.44 (1.31)	74.02 (0.41)	80.20 (1.16)	74.98 (0.42)	80.55 (1.33)	75.71 (0.56)
APTG	0.51 (0.01)	0.44 (0.005)	0.59 (0.02)	0.52 (0.003)	0.62 (0.01)	0.56 (0.002)	0.63 (0.02)	0.57 (0.005)

[1]Results from approaches described by Spangler *et al.* (2008). 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}.

[2]Results are the average of five replicates.

[3]Descriptions of the parameters can be found in equations (5)–(10).

Table 2 Number of animals with one or two alleles known, percentage of alleles known (SD) and probability of assigning the true genotype (SD) from multiple approaches[1] compared to ant colony optimization using a field data pedigree.

	True allele frequency							
	Random		Males		Males and females		Ant colony optimization	
Parameter[2]	0.3	0.5	0.3	0.5	0.3	0.5	0.3	0.5
No. animals with								
2 alleles known	1505	1501	1473	1470	2086	1999	1767	1706
1 allele known	2508	2144	11 756	10 607	10 376	9398	11 451	10 382
Benefit function	20 569	18 609	34 877	32 282	34 005	31 456	34 978	32 547
AK_P	9.48	8.84	25.26	23.28	24.99	23.02	25.75	23.70
AK_G	35.34	31.97	59.92	55.47	58.43	54.05	60.10	55.92
APTG	0.39	0.35	0.44	0.39	0.44	0.40	0.45	0.40

[1]Results from approaches described by Spangler *et al.* (2008). 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}.
[2]Descriptions of the parameters can be found in equations (5)–(10).

Table 3 Number of animals with one or two alleles known, percentage of alleles known (SD) and probability of assigning the true genotype (SD) from multiple approaches[1] compared to ant colony optimization using a research pedigree.

	True allele frequency							
	Random		Males		Males and females		Ant colony optimization	
Parameter[2]	0.3	0.5	0.3	0.5	0.3	0.5	0.3	0.5
No. animals with								
2 alleles known	452	458	438	439	1082	751	975	720
1 allele known	847	682	5525	4132	4747	3768	5101	4009
Benefit function	9719	8284	14 113	12 018	13 743	11 848	13 916	11 991
AK_P	10.08	9.19	36.84	28.83	39.77	30.33	40.58	31.36
AK_G	55.94	47.68	81.22	69.16	79.09	68.19	80.09	68.15
APTG	0.50	0.43	0.69	0.51	0.68	0.52	0.69	0.52

[1]Results from approaches described by Spangler *et al.* (2008). 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}.
[2]Descriptions of the parameters can be found in equations (5)–(10).

the homozygosity of the genotyped animals, are 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. The same could be true in the swine industry, as illustrated by Macrossan *et al.* (2006), where sampling sires only proved more beneficial than sampling both sires and dams, under the assumption that each sire would be mated to 40 females. 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. 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 because of 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 superior to other alternatives when very small (a few hundred animals) pedigrees are considered, or in situations where more than 5% of animals are genotyped because of a reduction in the number of animals with large diagonal elements in A^{-1}.

One assumption of the current study is that the allelic frequencies are known. Macrossan & Kinghorn (2003b) showed that incorrect assumptions of population allele frequencies could alter the performance of segregation analysis in the context of selective genotyping and calculating genotype probabilities for ungenotyped animals, particularly when the assumption is of an extreme frequency (i.e. 01.) and that an assumption of an intermediate frequency is more robust. In the case when a small number of animals, perhaps even one, are used to approximate the population frequencies, then the probability of error is higher. As the number of animals sampled increases, then it is reasonable to assume that the accuracy of the assumed allele frequencies is greater. Spangler *et al.* (2008) used the same pedigrees as used in the current study and explored the differences between using estimated allele frequencies from the sampled animals and assuming that they were known. Due to the fact that the simulated genotypes are randomly assigned in the base population and thus not subject to the effects of artificial selection, the estimated allele frequencies are virtually identical to the true values. The effects of selection over time could impact the ability to sample few animals and accurately determine allelic frequency. However, all methods would be subject to this error and it would be reasonable to assume that ACO would still show advantages over the other methods illustrated by Spangler *et al.* (2008).

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. However, the proposed method was found to be fairly robust with regard to proposed heuristic parameters. Finally, ACO was superior even with the simplistic pheromone function used in this study. The choice of an accuracy function drives the performance of the algorithm and it is possible that more sophisticated functions, which more completely exploit the pedigree structure, could increase performance but may become more computationally costly. This is an area where further research is being conducted.

Acknowledgements

The authors would like to thank the American Gelbvieh Association and the USDA-ARS Livestock and Range Research Laboratory at Miles City, Montana for supplying pedigrees used in the current study. The first author would also like to thank Dr Robyn Sapp for contributions to an earlier manuscript from which the majority of comparisons in the current study are made.

References

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

Dorigo M. & Stuetzle T. (2004) *Ant Colony Optimization*. MIT Press, Cambridge, MA, USA.

Fernández S.A., Fernando R.L., Guldbrandtsen B., Totir L.R. & Carriquiry A.L. (2001) Sampling genotypes in large pedigrees with loops. *Genetics Selection Evolution* **33,** 337–67.

Kinghorn B.P. (1999) Use of segregation analysis to reduce genotyping costs. *Journal of Animal Breeding and Genetics* **116,** 175–80.

Kinghorn B.P., Bastiaansen J.W.M., van der Steen H.A.M., Deeb N., Yu N. & Mileham A.J. (2006) Visually aided interpretation of results from a genome scan. In: *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production,* Communication No. 20-05.

Macrossan P.E. & Kinghorn B.P. (2003a) Cyclic genotyping strategies. II. True and perceived utilities under incorrect allele frequency assumptions. *Journal of Animal Breeding and Genetics* **120,** 312–21.

Macrossan P.E. & Kinghorn B.P. (2003b) A genetic algorithm to investigate genotyping in groups. *Association for the Advancement of Animal Breeding and Genetics* **15,** 43–6.

Macrossan P.E., Kinghorn B.P. & Davis G.P. (2001) Strategies for cost effective DNA testing for the *Thryoglobulin* gene in beef cattle. *Association for the Advancement of Animal Breeding and Genetics* **14,** 309–12.

Macrossan P.E., Southwood O.I. & Kinghorn B.P. (2006) The practical application of group genotyping theory in porcine herds. In: *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production,* Communication No. 24-20.

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

Ressom H.W., Varghese R.S., Orvisky E., Drake S.K., Hortin G.L., Abdel-Hamid M., Loffredo C.A. & Goldman R. (2007) Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics* **23,** 619–26.

Sheehan N.A. (2000) On the application of Markov chain Monte Carlo methods to genetic analysis of complex pedigrees. *International Statistical Review* **68,** 83–110.

Sorenson D., Wang C.S., Jensen J. & Gianola D. (1994) Bayesian analysis of genetic change due to selection using Gibbs sampling. *Genetics Selection Evolution* **26,** 229–49.

Spangler M.L., Sapp R.L., Bertrand J.K., MacNeil M.D. & Rekaya R. (2008) Different methods of selecting animals for genotyping to maximize the amount of genetic information known in the population. *Journal of Animal Science* **86,** 2471–9.

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

Thallman R.M., Bennett G.L., Keele J.W. & Kappes S.M. (2001) Efficient computation genotype probabilities for loci with many alleles: II. Iterative method for large, complex pedigrees. *Journal of Animal Science* **79,** 34–44.

Wang C.S., Rutledge J.J. & Gianola D. (1993) Marginal inferences about variance components in a mixed linear model using Gibbs sampling. *Genetics Selection Evolution* **25,** 41–62.