Revista Brasileira de Zootecnia
© 2009 Sociedade Brasileira de Zootecnia
ISSN 1516-3598 (impresso)
ISSN 1806-9290 (*on-line*)
www.sbz.org.br

R. Bras. Zootec., v.38, p.93-97, 2009 (supl. especial)

Ant colony algorithm for analysis of gene interaction in high-dimensional association data

Romdhane Rekaya[1,2,3], Kelly Robbins[1]

[1] Department of Animal and Dairy Science.
[2] Department of Statistics.
[3] Institute of Bioinformatics University of Georgia, Athens, GA 30602 USA.

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

Key Words: genome, simulation, SNP

Algoritmo colônia de formigas para análise de interação gênica em dados de associação de alta dimensão

RESUMO - Nos últimos anos muita atenção tem sido dada ao uso de polimorfismos de nucleotídeos simples (SNP) para mapeamento fino do genoma, visando identificar mutações efetivas em características de interesse; todavia, muitos estudos focam apenas os efeitos marginais dos marcadores, ignorando as potenciais interações entre genes. Estudos de simulação tem mostrado que esta abordagem pode não ser poderosa o suficiente para detectar loci importantes quando interações entre genes estão presentes. Vários estudos tem examinado potenciais interações gênicas, porém focando um pequeno número de marcadores SNP. Devido ao proibitivo custo computacional para modelar interações em estudos envolvendo um grande número de SNP's, precisam ser desenvolvidos métodos que considerem potenciais interações gênicas, de uma forma computacionalmente eficiente. Este estudo adota a abordagem de um mecanismo de aprendizagem, adaptando o algoritmo de otimização colônia de formigas (ACA), combinado com regressão logística em função dos haplótipos e genótipos, para estudos de associação envolvendo grande número de marcadores SNP. O método proposto é comparado à análise de haplótipos, implementado usando uma janela deslizante (SW/H), e a associação de genótipos de lócus único (RG). Cada algoritmo foi avaliado usando uma característica binária simulada usando um modelo epistático e dados genotípicos do HapMap ENCODE. Os resultados mostram que o ACA superou o SW/H e RG em todos os cenários de simulação, produzindo aumentos substanciais no poder de detectar regiões genômicas associadas com características simuladas.

Palavras-chave: genoma, simulação, SNP

Introduction

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). Unfortunately, due to the high density of SNP maker maps, it is computationally infeasible to examine all possible interactions. As a result

Correspondências devem ser enviadas para: rrekaya@uga.edu

studies examining gene interactions tend to focus on a small number of SNP, previously identified as having strong marginal associations.

While this approach has shown some success, simulation studies conducted by Marchini et al. (2005) and Pickrell et al. (2007) showed that, in the presence of several types of gene interactions, there is reduced power to detect causative loci with models estimating only marginal effects. Using an exhaustive search of all two-way interactions, Marchini et al. (2005) achieved greater power to detect causative mutations when compared to models estimating only marginal effects. However, 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 an then tested for interactions in the subsequent stage [Marchini et al., 2005]. Such an approach represents a compromise that could 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.

Given that the examination of all possible SNP interactions is computationally infeasible with dense SNP marker maps covering large regions of a genome, an alternative approach must be considered. One such approach would be to view the identification of groups of interacting SNP as an optimization problem, for which several algorithms have been developed. These algorithms are designed to search large sample spaces for globally optimal solutions and have been applied to wide range of problems (Shymygelska & Hoos et al., 2005; Kreiger et al., 2000; 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. Kooperberg et al. (2006) utilized an optimization algorithm, referred to as simulated annealing (SA), to examine interaction effects; however, only 32 SNP were considered in the model selection process. For studies involving hundreds or even thousands of SNP, efficient algorithms are needed to search the sample space for optimal solutions.

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

For this study a modified ACA, enabling the use of permutation testing for global significance, was combined with 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.

Material and Methods

Logistic regression

Groups of SNP markers were evaluated based on haplotype and genotype effects estimated as log odds ratios (*lor*) using logistic regression (LR). The log odds ratio lor_i is modeled as:

$$lor_i = \ln(\frac{p_i}{1-p_i}) = X_i\beta \qquad (1)$$

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

The link function of the log odds ratio with the binary response y_i gives the following equations:

$$p_i(y_i = 0) = \frac{1}{1+\exp(X_i\beta)} \text{ and } p_i(y_i = 1) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)} \qquad (2)$$

Marginal effects model

The genotype and haploype association methods were implemented using R functions developed by Gonzalez et al. (2007) and Sinnwell & Schaid (2005), respectively. 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 average of all haplotypes containing a given SNP.

Ant colony algorithm

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

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

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

Using the PDF as defined in equation (4), each of *j* artificial ants selects a subset S_k of *n* SNP from the sample space S containing all SNP. Given the relationship between adjacent SNP, ants randomly change SNP selections following a multinomial distribution, with changes being limited to the three adjacent SNP on either side of the originally selected SNP marker. The pheromone level of each feature *m* in S_k is then updated according to the performance of S_k as:

$$\tau_m(t+1) = (1-\rho) * \tau_m(t) + \Delta\tau_m(t) \qquad (4)$$

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 SNP *m* was not selected by any of the artificial ants.

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 (4), 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 the amount of pheromone deposited on all other SNP markers and can vary wildly from permutation to permutation. One obvious solution to this problem would be to use the average accuracy of all 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:

$$P_m(t) = \frac{\tau_m(t)^{\alpha}\tau 2_m(t)^{\alpha 2}\eta_m^{\beta}}{\sum_{m=1}^{nf}\tau_m(t)^{\alpha}\tau 2_m(t)^{\alpha 2}\eta_m^{\beta}} \qquad (5)$$

where $_m(t)$ 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 containing genotypes for SNP *m*; and η_m, α, β are as previously described. For the current study α and α2were set to 1, β 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 (4) and $\tau 2_m(t)$ was updated using the following equation:

$$\tau 2_m(t+1) = [t * \tau_m 2(t) + \Delta\tau_m 2(t)]/(t+ns) \qquad (6)$$

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 containing genotypes for SNP *m*, which 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.

Data simulations

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 1. The loci of the causative mutations were selected at random; with the frequencies of the causative mutations being .58 and .6. Although these frequencies might be considered high, it was necessary to restrict selection to SNP with mutant allele frequencies greater than .5. This was done to insure a reasonable simulated disease incidence of 15%.

Table 1 - Relative risk for simulated trait[a]

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

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

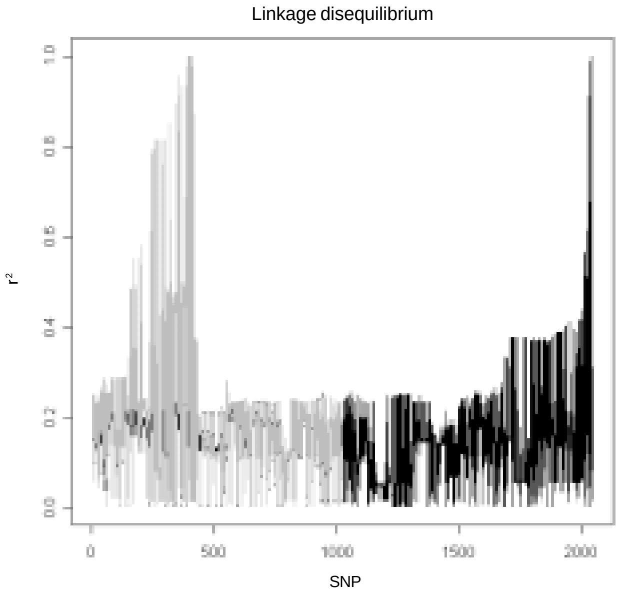

Figure 1 - Plots of each marker's linkage disequilibrium (LD) with the two causative mutations. The light grey line represents LD with the causative mutation located at position 409. The black line represents LD with the causative mutation located at position 2041.

A plot illustrating the LD of all SNP with the two causative mutations is shown in Figure 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.

Permutation testing was used to access global significance for all models used in the study. Statuses were random 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 permutation 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.

Results

Estimates of power for the three methods can be found in Table 2. 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 tended 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 LR 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.

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 the simulated trait by ACA/G/D and RG/D were plotted. 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 ($\alpha = 0.05$) 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 likely imparted a portion of the epistatic effect on these SNP, resulting in significant associations with the simulated traits.

Table 2 - Power calculations[a]

	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.

Conclusions

In the presence of simulated epistasis, the proposed optimization methodology obtained substantial increases in power over models accounting for only marginal effects, 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 methodologies capable of efficiently modeling gene interactions, such as the model proposed in this study, could yield superior performance detecting important SNP markers for complex traits controlled by interacting loci.

Literature Cited

BARENDSE, W.; HARRISON, B.E.; HAWKEN, R.J. et al. Epistasis between Calpain 1 and its inhibitor Calpastatin within breeds of cattle. **Genetics**, v.176, p.2601-2610, 2007.

BENJAMINI, Y.; YEKELI, D. Quantitative trait loci analysis using the false discovery rate. **Genetics**, v.171, p.783-790, 2005..

COUTINHO, A.M.; SOUSA, I.; MARTINS, M. et al. Evidence for epistasis between SLC6A4 and ITGB3 in autism etiology and in the determination of platelet serotonin levels. **Human Genetic**, v.121, p.243-256, 2007.

DING, Y.P.; WU, Q.S.; SU, Q.D. Multivariate calibration analysis for metal porphyrin mixtures by an ant colony algorithm. **Analytical Sciences**, v.21, p.327-330, 2005.

DORIGIO, M.; GAMBARDELLA, L.M. Ant colonies for the travailing salesman problem. **BioSystem**s, v.43, p.73-81, 1997.

GONZALEZ, J.R.; ARMENGOL, L.; SOLE, X. et al. SNPassoc: an R package to perform whole genome association studies. **Bioinformatics**, v.23, n.5, p.644-645, 2007.

HUGOT, J.P.; CHAMAILLARD, M.; ZOUALI, H. et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. **Nature**, v.411, p.599-603, 2001.

KOOPERBURGE, C.; BIS, J.C.; MARCIANTE, K.D. et al. Logic regression for analysis of the association between genetic variation in the rennin-angiotensin system and myocardial infarction or stroke. American Journal of Epidemiology, v.165, p.334-343, 2006.

KREIGER, M.J.B.; BILLETER, J.B.; KELLER, L. Ant-like task allocation and recruitment in cooperative robots. **Nature**, v.406, p.992-995, 2000.

MARCHINI, J.; DONNELLY, P.; CARDON, L.R. Genome-wide stregies for detecting multiple loci that influence complex diseases. **National Genetics**, v.37, p.413-417, 2005.

PICKRELL, J.; CLERGET-DARPOUX, F.; BOURGAIN, C. Power of genome-wide association studies in the presence of interacting loci. **Genetic Epidemiology**, 2007. [Epub ahead of print].

RESSOM, H.W.; VARGHESE, R.S.; ORVISKY, E. et al. Peak selection from MALDI-TOF mass spectra using ant colony optimization. **Bioinformatics**, v.23, n.5, p.619-626, 2007.

ROBBINS, K.R.; ZHANG, W.; REKAYA, R. et al. Ant colony optimization for feature selection in high dimensionality data sets. **Journal of Mathematics Applied in Medicine and Biology**, 2007. (Accepted).

SHYMYGELSKA, A.; HOOS, H.H. An ant colony optimization algorithm for the 2D and 3D hydrocarbon polar protein folding program. **BMC Bioinformatics**, v.6, p.30, 2005.

SINNWELL, J.P.; SCHAID, D.J. **Haplostats**: statistical analysis of haplotypes with traits and covariates when linkage phase is ambiguous. R package version 1.2.2. 2005. (CD-ROM).

WOON, P.Y.; KAISAKI, P.J.; BRAGANCA, J. et al. Aryl hydrocarbon receptor nuclear translocator-like (BMAL1) is associated with susceptibility to hypertension and type 2 diabetes. **Proceedings of the National Academy of Science**, v.104, n.36, p.14412-14417, 2007.