

MACHINE LEARNING APPROACHES FOR THE ANALYSIS OF COMPLEX, HIGH-DIMENSION GENOMIC DATA

K. R. Robbins, J. K. Bertrand, and R. Rekaya

ABSTRACT

The use of genomic technology has the potential to provide invaluable insight into the mechanisms of many important production and fitness traits in the animal production industry. Unfortunately this information comes at a cost, in terms of the high-dimensions and complex structure of genomic datasets. Due to the large number of features in many genomic datasets, explicit modeling of gene interactions is often infeasible. As such, models are often nested within genes, assuming independence between gene effects. Given that genes operate in interacting networks, such assumptions can limit the power to detect important gene variants. To eliminate the need for simplifying assumptions, a machine learning algorithm, referred to as the ant colony algorithm (ACA), has been adapted for analysis of high-dimension genomic data. The performance of the ACA was compared to several standard methodologies using both simulated and real genomic datasets. Methodologies were evaluated based on their performance in identifying highly predictive biomarkers and causative mutations under the influence of genetic interactions. Results show that the ACA achieved substantially better performance in all scenarios evaluated.

INTRODUCTION

In the last decade there has been a rapid expansion of new technologies enabling the efficient collection of large quantities of genomic data in important animal production species. These high-throughput technologies, capable of capturing thousands of data points on single nucleotide polymorphisms (SNP), gene expression, protein expression, and metabolomic data, have ushered in the ‘omic’ era of quantitative biology. Unfortunately the high-throughput nature of these technologies has proven to be a double-edged sword, as the high-dimensions and complex structure of these datasets have made effective modeling and extraction of useful information difficult. Due to the complex interactions among genes and their products, large numbers of parameters must be estimated from datasets with relatively small replication. As a result, models are often applied that analyze each gene effect independently or account for only marginal effects of markers, ignoring important epistatic relationships (Wofinger et al. 2001; McClurg et al., 2006). Though these methods can be effective in identifying genes with significant marginal contributions they do not take into account the predictive power of a feature when grouped with other important features in a classifier (Shen et al., 2006). As such, these methods may not be suitable for applications in animal breeding, where the ultimate goal is to identify genomic features that can accurately predict offspring performance.

For such applications machine learning and optimization algorithms may be better suited than nested models. These methodologies require no explicit modeling of data structures, but rely on simple algorithms that are often based on natural processes that are capable of implicitly modeling data structures. Given the potential benefits of machine

learning approaches, the goal of the current study was to develop machine learning methodologies capable of efficiently identifying biologically relevant and highly predictive gene expression profiles and single nucleotide polymorphisms in the presence of complex genomic data structures.

MATERIALS AND METHODS

Data

Two datasets were used in this study. This first dataset (D1) contained the expression of 14,525 genes collected on 198 subjects (Ramaswamy et al., 2001). Each subject fell into one of fourteen different categorical phenotypes. The second dataset had 2047 SNP genotyped on 90 individuals from the HapMap ENCODE project. Phenotypes for D2 were simulated as a binary trait under the influence of two, randomly selected, interacting SNP. All simulations were replicated three times using the parameters found in Table 1.

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.

Biomarker Identification

For D1 several nested models, utilizing t-test (T), fold change (FC), and penalized t-test statistics (PT), were used to select genes for phenotype prediction. In the case of the SNP data, publicly available software was used to test for significant genotype effects (WG) (Gonzalez et al., 2007). Due to its computational efficiency, a machine learning algorithm, referred to as the ant colony algorithm (ACA) was modified for analysis of all genomic datasets. The performance of the ACA was compared to the nested models based on prediction accuracy (D1) and power to identify predictive genomic regions (D2). Prediction accuracy was determined using cross validation with 144 subjects used to train a classifier and 54 subjects used for validation. Power calculations for D2 were estimated as the proportion of times at least one SNP, in linkage disequilibrium > .9 with a causative mutation, was detected as being significant when controlling for a family-wise error rate of 0.05 using permutation testing.

Ant Colony Algorithm

Artificial ants work as parallel units that communicate through a probability density function (PDF) that is updated by weights or “pheromone levels”, in this case determined by the performance of the selected features in classifying samples (Dorigio and Gambardella,

1997; Resson et al., 2006), where the probability of sampling feature m at time t is defined as:

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

where $\tau_m(t)$ is the amount of pheromone for feature m (out of a total of nf features) 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, respectively. For this study the prior information (η_m) was determined as the average FC, T, and PT scores for gene expression data and the marginal effects of SNP for marker data.

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 selects 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 . The change in pheromone, $\Delta\tau_m(t)$, was calculated using prediction accuracies obtained using prohibit and logit latent variable models for gene expression and marker data, respectively.

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 rates by subsequent artificial ants which, in turn, deposit more “pheromone”, thus leading to a positive feedback system similar to the method of communication observed in real ant colonies. Upon convergence, the optimal subset of features is selected based on the level of pheromone deposited on each feature, and p-values computed using permutation testing for gene expression and marker data, respectively.

RESULTS AND DISCUSSION

Table 2 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), and MAMA (Antonov et al., 2004)). The proposed ACA yielded substantial increases in accuracies over all other methods, with a 6.5% increase in accuracy over the next best results (Antonov et al., 2004). When compared to the nested methods, the ACA achieved increases

of 13.9%, 40%, and 16.6% in accuracy over the FC, T, and PT methods of feature selection, respectively.

Table 2. Accuracy (%) of tumor class predictions using ant colony algorithm (ACA) and several previously published methods.

Method	Prediction Accuracy (%)
ACA(14525 ^a)	90.7
FC(14525)	79.6
T(14525)	64.8
PT(14525)	77.8
GASS(1000)	81.5
GA/MLHD(1000)	76.0
MAMA	85.2

^aNumber of genes selected prior to the implementation of feature selection algorithm.

By evaluating the prediction performance of groups of genes, rather than individual gene effects, the ACA reduced collinearity in the select genes, enabling higher prediction accuracies to be obtained using fewer selected biomarkers. Furthermore, due to computational efficiency of the ACA, truncation of the dataset prior to implementation was not needed. This was in contrast to other machine learning algorithms (GASS, GAMLHD) which require the removal of the gene expression values of 13,000 genes prior to implementation in order to converge to good solutions.

Power calculations for analysis of D2 using ACA and WG can be found in Table 3. In scenario 1, where the gene effects were the strongest, the ACA performed very well with an estimated power of .83 averaged across three replicates. In contrast, WG showed poor performance with a power of only .33. For scenario 2, in which gene effects were reduced by 50%, both models showed substantial decreases in power; however, the ACA retained superior performance with an estimated power of .33, as compared to the power of .17 obtained using WG. These substantial increases in power demonstrate the effectiveness of the ACA mechanism in accounting for gene interactions. This allowed the ACA to detect causative mutations with no significant marginal effects, as identified by WG.

Table 3. Power calculations for ACA and WG^a.

	2 SNP Haplotypes	3 SNP Haplotypes	Average
		ACA	
Scenario 1	0.83	0.83	0.83
Scenario 2	0.17	0.50	0.33
Overall	0.500	0.67	0.58
		WG	
Scenario 1	_____	_____	0.33
Scenario 2	_____	_____	0.17
Overall	_____	_____	0.25

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

When looking at the effect of haplotype size on the performance of ACA (Table 3), it can be seen that there was no effect for scenario 1, while three SNP haplotypes performed best in scenario 2. These findings are important as the true number of interacting SNP will be unknown for real data applications. As such it is important that the ACA be robust with respect to the number of SNP selected by the ACA versus the true number of interacting SNP. While these results suggest some level of robustness, as two SNP haplotypes were not superior to three SNP haplotypes, it is clear that the number of SNP selected by the ACA can impact performance; therefore, an approach in which several haplotype sizes are examined may ensure that the optimal performance of the algorithm is achieved.

CONCLUSIONS

The superior performance of the modified ant colony algorithm, when applied to both gene expression and genetic marker datasets, demonstrates the potential benefits of machine learning approaches to the analysis of high-dimension genomic datasets. The increases in power obtained when using the ant colony algorithm resulted from the fact that the algorithm is capable of modeling the true structure of the datasets without the need for simplifying assumptions, as required by many traditional methodologies. This advantage is particularly important in genomic datasets where data structures are often complex and unknown.

LITERATURE CITED

- 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.
- Dorigio, M. and L. M. Gambardella. 1997. Ant colonies for the travelling salesman problem. *BioSystems*. 43:73-81.
- Golub, T. R. 2001. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci.* 98:15149-15154.
- 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
- Lin, T., Liu, R., Chen, C., Choa, Y. and S. Chen. 2006. Pattern classification in DNA microarray data of multiple tumor types. *Pattern Recognition*. 39:2426-2438.
- McClurg, P., Pletcher, M. T., Wiltshire, T. and A. I. Su. 2006. Comparative analysis of haplotype association mapping algorithms. *BMC Bioinformatics*. 7:61.
- 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.

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

Shen, R., Ghosh, D., Chinnaiyan, A. and Z. Meng. 2006. Eigengene-based linear discriminant model for tumor classification using gene expression microarray data. *Bioinformatics*. 22:2635-2642.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P, Afshari, C. and R. S. Paules. 2001. Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol*. 8(6):625-637.