
Ridge regression based hybrid genetic algorithms for multi-locus quantitative trait mapping

Bin Zhang and Steve Horvath*

Departments of Human Genetics and Biostatistics,
University of California at Los Angeles,
Los Angeles, CA 90095 7088, USA
E-mail: binzhang@mednet.ucla.edu
E-mail: shorvath@mednet.ucla.edu
*Corresponding author

Abstract: Genetic algorithms (GAs) are increasingly used in large and complex optimisation problems. Here we use GAs to optimise fitness functions related to ridge regression, which is a classical statistical procedure for dealing with a large number of features in a multivariable, linear regression setting. The algorithm avoids overfitting, gracefully handles collinearity and leads to easily interpretable results. We use the method to model the relationship between a quantitative trait and genetic markers in a mouse cross involving 69 F2 mice. The approach will be useful in the context of many genomic data sets where the number of features far exceeds the number of observations and where features can be highly correlated.

Keywords: genetic algorithm; ridge regression; quantitative trait mapping; gene interactions.

Reference to this paper should be made as follows: Zhang, B. and Horvath, S. (2006) 'Ridge regression based hybrid genetic algorithms for multi-locus quantitative trait mapping', *Int. J. Bioinformatics Research and Applications*, Vol. 1, No. 3, pp.261–272.

Biographical notes: Bin Zhang received his PhD in Computer Science from the State University of New York at Buffalo. He was a PostDoc fellow at the University of California, Los Angeles. He is a Research Faculty and Senior Statistician in the Departments of Human Genetics and Biostatistics, University of California, Los Angeles. His primary research interests are in Computational Biology (Gene Coexpression Networks, Systems Biology), Pattern Recognition, Machine Learning and Data Mining.

Steve Horvath completed his PhD in Mathematics from the University of North Carolina in 1995 and a Doctorate Degree in Biostatistics from the Harvard School of Public Health in 2000. He is an Assistant Professor in Human Genetics and Biostatistics at the University of California, Los Angeles in 2000. His research interests include family based allelic association tests (FBAT software), microarray data analysis (e.g., gene coexpression networks), tissue microarray data analysis (e.g., random forest clustering) and machine learning methods that are relevant for genomic data.

1 Introduction

The linear regression model is one of the most important tools for modelling the relationship between a quantitative variable (the outcome or dependent variable) and several features (predictors, independent variables). For example, we were interested in modelling the relationship between 132 genetic markers (trivariate features) and fat mass in 69 mice. There are several standard methods for modelling the relationship between genetic markers and clinical traits in F2 mouse crosses: quantitative trait locus mapping and F2 mouse crosses are reviewed in Silver (1995). Many methods have been proposed to deal with this situation. In contexts where feature interpretation is important, the linear model and least squares regression may be preferable to black box approaches such as neural nets or k -nearest neighbour classifiers. Least squares estimates often have low bias but large variance (Hastie et al., 2001). The prediction accuracy can sometimes be improved by ‘shrinking’ some coefficients using a technique called ridge regression (Hastie et al. 2001). Ridge regression has another major advantage in the context of genomic data: it gracefully deals with multicollinearity by putting size constraints on parameter estimates (Feiveson, 1994).

A problem with least squares estimates involving many features, may be difficulty of interpretation. This can be addressed by selecting a small subset of features that contain the strongest effect. There are several methods for choosing a subset of features. Best subset regression finds for each k , the subset of size k that gives the smallest residual sum of squares, or equivalently the largest R^2 value (defined below). Several algorithms (e.g., the leaps and bounds procedure) make this feasible for situations with dozens of features. Instead of searching through all possible subsets, other strategies seek to find a good path through them, e.g., forward stepwise regression starts with an ‘intercept only’ model and uses a heuristic to sequentially add features. But when dealing with hundreds of features, the local optima found by stepwise approaches may not be global.

Instead of using stepwise regression, we use a genetic algorithm (GA) to maximise a ridge regression based objective function. As a type of heuristic and stochastic optimisation technique invented by Holland in the 1960s, GAs have been successfully applied to large and complex optimisation problems (Holland, 1975; Goldberg, 1989; Lawrence, 1991). Through simulating natural selection and evolutionary processes in biological systems, GAs probabilistically select highly fit individuals as parents for producing offspring in the next generation (*selection*). The offspring are created using a *crossover* operation. As the offspring inherit ‘good’ genes from their parents, they more likely possess better fitness, thus the overall fitness of the new generation will be improved. Such a process together with the third operation of *mutation* is iterated until the overall fitness of the population becomes stable, resulting in a set of nearly optimal solutions. The hope is that when population size and number of iterations become large enough, the solutions presented by GAs are more likely to be globally optimal due to their stochastic characteristics.

Several studies have used GAs for feature selection by defining a suitable objective function for computing the fitness of each individual in a population. In the work of identifying informative features for discriminating normal and tumour samples by Li et al. (2001a, 2001b), a GA was combined with the k -nearest neighbour classification accuracy as fitness function. For the same task, Keedwell et al. developed a neural-genetic hybrid algorithm where objective function was derived from the classification accuracy of an artificial neural network (Keedwell and Narayanan, 2003).

Another approach by Liu et al. (2001) adopted a Golub-Slonim classifier and the fitness was defined as a weighted sum of the classification accuracy and a penalty term for the number of selected features. However, these three GA based selection schemes do not use the standard predictor for continuous outcomes: the linear regression model.

In this study, we propose a novel algorithm for feature selection by coupling GAs with a classical statistical predictor: ridge regression (reviewed in Hastie et al. (2001)). The proposed algorithm is applicable to problems with both discrete and continuous outcomes and protects against overfitting by using the ridge regression penalty term.

2 A ridge regression based island model genetic algorithm

Island model genetic algorithms (IMGA) are effective parallel genetic algorithms. The idea is to split the total population into subpopulations (called islands) and let each subpopulation evolve independently, except for exchanging the elite individuals with the highest fitness at each generation (*called migration*). An IMGA can greatly reduce the probability of premature convergence to a local optimum and all islands can be processed on parallel computers. In the following, we describe the problem of feature selection, the encoding scheme of IMGA, and the proposed objective function.

2.1 Problem description

Let $\{s_j | j = 1, 2, \dots, J\}$ be a set of J samples. The outcome of each sample s_j is denoted by the trait t_j . For each sample, the I-dimensional vector of features (here genetic markers) is denoted by $X^j = \{x_1^j, x_2^j, \dots, x_n^j\}$. One aims to find a subset of features, $\hat{X} = \{x_{k_1}, x_{k_2}, \dots, x_{k_N}\}$, that best predict the trait t . The algorithm yields several feature importance measures that can be used to rank the features.

2.2 Encoding scheme and objective function

For the problem defined above, a genetic algorithm (ga)-chromosome C in the IMGA is defined by the I-dimensional binary vector $C = \{b_1, b_2, \dots, b_I\}$ where each bit indicates ('1' and '0') whether or not the corresponding feature is included in the ga-chromosome. acc indicates the corresponding feature selected (unselected) and the fitness of a chromosome is computed by ridge regression.

To define the fitness function, we use ridge regression (reviewed in Hastie et al. (2001)). As a shrinkage method, ridge regression is an alternative to feature subset selection. Through shrinking the regression coefficients by forcing a penalty on the size of subset, ridge regression is able to smoothly approach the solution with less variance compared with forward and backward stepwise selection methods. Given the trait t and a ga-chromosome C leading to a subset of selected features $X = \{x_{k_1}, x_{k_2}, \dots, x_{k_N}\}$, we fit the following pairwise interaction model:

$$t(C) = \beta_0 + \sum_{i=1}^n \beta_i x_{ki} + \sum_{u=1}^{n-1} \sum_{v>u} \beta_{uv} x_{ku} x_{kv}. \quad (1)$$

Then the ridge regression is to compute the coefficient set $\hat{\beta} = \{\beta_0, \beta_1, \beta_2, \dots, \beta_n\}$ by minimising the penalised residual sum of squares

$$\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{j=1}^J \left(t_j - \beta_0 - \sum_{i=1}^n \beta_i x_{ki}^j - \sum_{u=1}^{n-1} \sum_{v>u} \beta_{uv} x_{ku}^j x_{kv}^j \right)^2 + \lambda \left(\sum_{i=1}^n \beta_i^2 + \sum_{u=1}^{n-1} \sum_{v>u} \beta_{uv}^2 \right) \right\}.$$

Then, the objective (fitness) function is defined as a multiple R^2 value, which is a decreasing function of the residual sum of squares

$$R^2(C) = 1 - \frac{\sum_{j=1}^J (t_j - \hat{\beta}_0 - \sum_{i=1}^n \hat{\beta}_i x_{ki}^j - \sum_{u=1}^{n-1} \sum_{v>u} \hat{\beta}_{uv} x_{ku}^j x_{kv}^j)^2 + \Delta}{\sum_{j=1}^J (t_j - \bar{t})^2}, \quad (2)$$

where, $\Delta = \lambda \left(\sum_{i=1}^n \beta_i^2 + \sum_{u=1}^{n-1} \sum_{v>u} \beta_{uv}^2 \right)$, $\bar{t} = 1/J \sum_{j=1}^J t_j$ is the mean outcome and $\sum_{j=1}^J (t_j - \bar{t})^2$ is the total outcome variation.

The complexity parameter $\lambda \geq 0$ controls the shrinkage. In the case of ordinary least squares regression ($\lambda = 0$), R^2 lies between 0 and 1; $R^2 = 1$ indicates perfect model fit.

2.3 Choice of the ridge complexity parameter λ

Some variant of model selection should be used to choose a value for the ridge complexity parameter λ . In general, the value chosen should be the one associated with the largest R^2 value.

Several methods have been discussed in the literature. The popular choices are ‘leave one out’, crossvalidation, generalised crossvalidation (Golub et al., 1979), and variants of the Bayesian or Akaike information criterion, and bootstrap methods (Efron and Tibshirani, 1993).

In this paper, we explored the use of three classical estimators of λ , which are implemented in the *R* (<http://cran.r-project.org/>) contributed package MASS. These are the modified Hoerl-Kennard-Baldwin (HKB) estimator, the modified L-W estimator (Brown, 1994) and the generalised crossvalidation method (GCV).

We have found that the HKB estimator works well in this application. However, no claim is made that this is the optimal method. Comparing different ways of estimating λ is beyond the scope of this paper.

2.4 Outline of the algorithm

Consider an island model with M islands and P ga-chromosomes within each island. Let μ be the mutation rate and Q be the number of ga-chromosome pairs selected for φ -point crossover. The algorithm is given by the following steps.

Step 0 (Initialisation): Generate a set of $M \times P$ ga-chromosomes, each with at most η selected genes, and randomly divide the set into M equal sized islands. For each island, execute *Step 1* to *Step 5*.

Step 1 (Fitness evaluation): Compute the fitness of each ga-chromosome, based on the objective function (2).

Step 2 (Selection): Probabilistically select the Q pairs of best fitting ga-chromosomes for crossover to produce offspring and remove the Q pairs of least fitting ga-chromosomes (death).

Step 3 (Crossover): Perform the φ -point crossover operation for each of Q pairs of best fitting ga-chromosomes, leading to $Q \times (2^{\varphi+1} - 2)$ offspring.

Step 4 (Mutation): Probabilistically (according to μ) mutate genes in $Q \times (2^{\varphi+1} - 2)$ offspring.

Step 5 (Shuffler): Compute the fitness of each ga-chromosome in the $Q \times (2^{\varphi+1} - 2)$ offspring and take as new generation, the $2 \times Q$ best fitting chromosomes, together with the remaining $P - 2 \times Q$ ones (the Q pairs of least fitting chromosomes have died away).

Step 6 (Migration): The M islands exchange their best fitting ga-chromosomes and Steps 1–5 are repeated for each island until the average fitness value becomes stable.

2.5 Measures of feature importance

We use GAs to define two different measures of feature importance that can be used for feature selection. First, we compute the occurrence frequencies of all features in the final generations of M islands. Second, we make use of the fact that a linear model has been fit to the data. For each feature, we compute the average Wald-statistic Z , which is defined as the ratio of the corresponding parameter estimate and its standard error. In the classical linear model ($\lambda = 0$), one can show that Z has a standard normal distribution under the null hypothesis of no relationship between feature and regression outcome.

3 Experimental results and analysis

3.1 Experimental settings

In this study, we used an F-2 mouse population of 69 mice constructed from two standard inbred strains (Schadt et al., 2003). For each mouse, a quantitative outcome (fatmass) and 132 trivariate features were collected. Specifically, the features were genetic (SNP) markers (values 1,2,3) that were located across the mouse genome. A significant correlation between marker values and fatmass may indicate that the genetic marker is ‘close’ to a gene that affects fatmass. When looking at one marker at a time, variations of this approach are known as single locus quantitative trait locus mapping (QTL mapping) and numerous methods have been described in the literature (Silver, 1995). Here we were interested in modelling the combined effect of several genetic markers. The goal was to find markers that were far apart, possibly on different mouse-chromosomes that would best explain the fatmass of the mouse.

There are several parameters in the proposed hybrid genetic algorithms. As discussed in Section 2.3, we consider the HKB estimate for the ridge complexity parameter. Specifically, we first randomly select a number of subsets (with varying sizes) of markers. Then we compute the HKB estimate for each subset by forming ridge regression of the fatmass over the markers in the set. Our choice of λ is the average of the HKB

estimates. For this data set, the average HKB estimate of λ is around 3.0. To test the stability of the algorithm, we also consider $\lambda = 1.0$ and $\lambda = 5.0$. Another parameter in the ridge regression is the number of selected markers (η). It is biologically implausible that more than five different markers would interact to affect the outcome (Silver, 1995), which is why we restrict our attention to 3 to 5 markers.

Some other parameters involved, are the number of islands (M) and the population size (P) of each island. We are not aware of any method for estimating these island parameters, and the stability of the GAs with regard to M and P should be tested by varying the values of M and P . In our subsequent experiments, we consider $M \in \{10, 15, 20\}$ and $P \in \{200, 400\}$. For each island generation, $Q = 8$ pairs of ga-chromosomes are selected for eight point crossover and the standard mutation rate μ was set to 0.001. As described below, we find that the proposed hybrid GAs are not sensitive to these parameters.

3.2 Sensitivity of the algorithm to parameter choices

Figure 1 shows the importance measurements for all 132 markers under varying parameter settings. As expected, the algorithm is quite robust with regard to the parameter choices.

As the algorithm leads to similar results for different parameter settings, we restrict our attention to an IMGGA with the following parameters: 10 islands, 400 population, $\lambda = 3.0$, and $\eta = 4$ (genetic markers). We found that the average fitness and best fitness converge after 50 generations with a best fitness value of 0.46.

Figure 1 Marker importance measurements from IMGAs with varying parameters

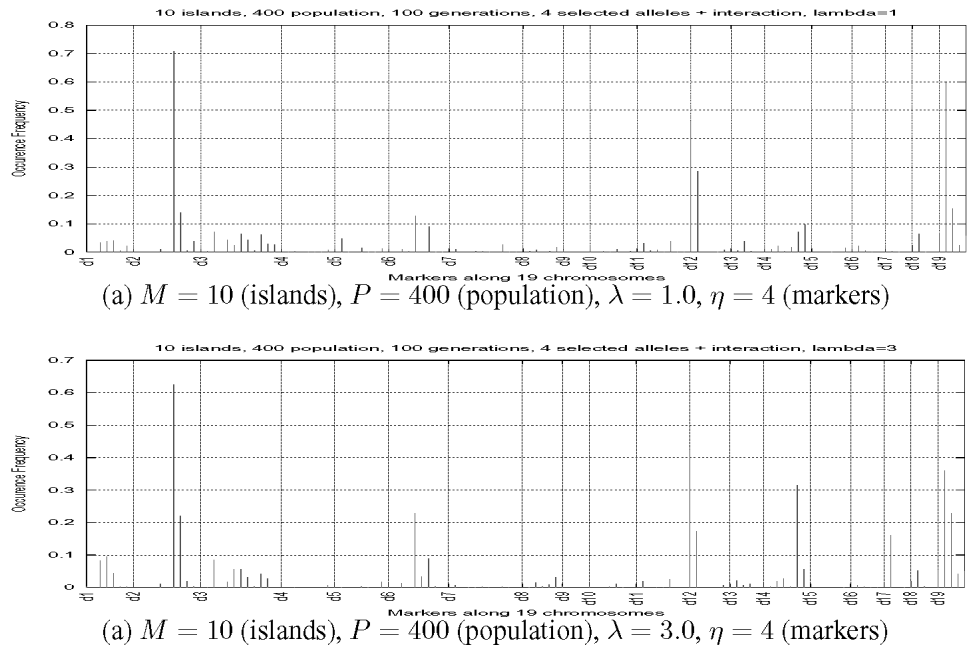
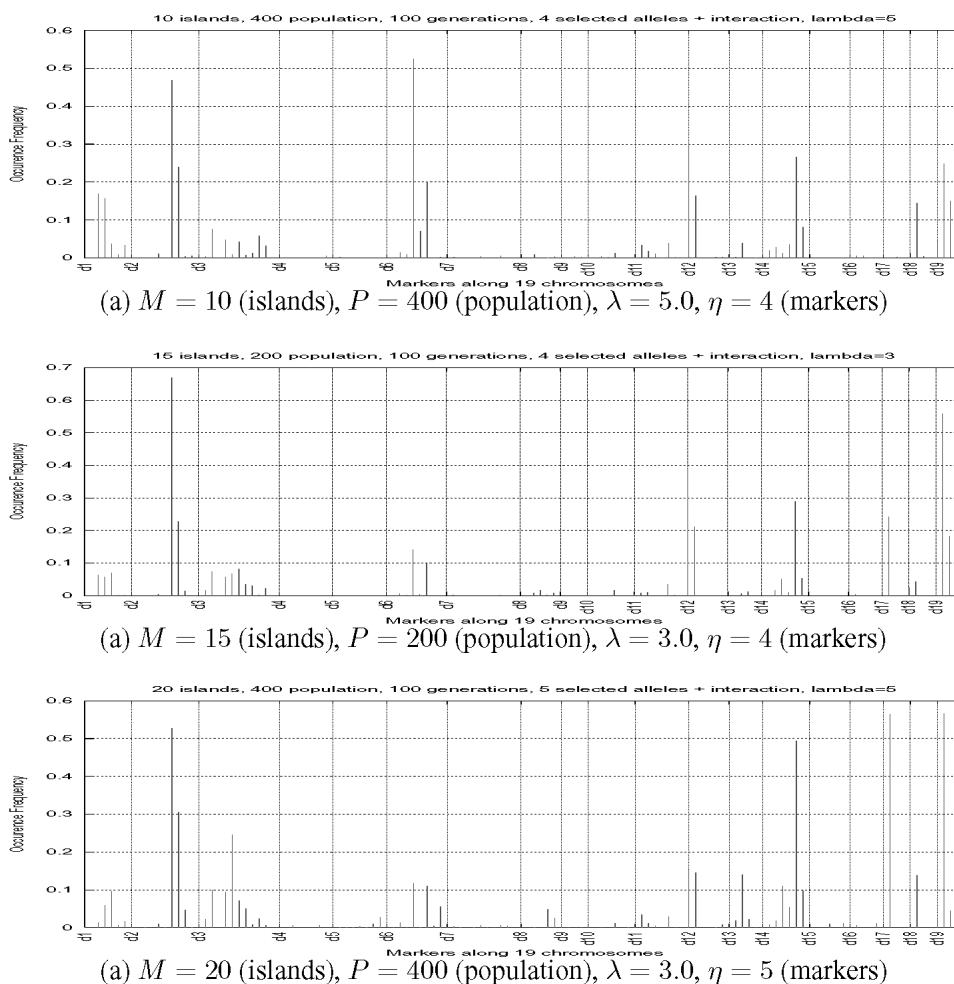


Figure 1 Marker importance measurements from IMGAs with varying parameters (continued)

3.3 Importance measurement of genetic markers

Figure 2 presents the two importance measures for all the 132 markers. We find that nine markers have occurrence frequencies above 0.1 and these markers are listed below in the descendant order of occurrence frequency, *d2m50*, *d12m1*, *d19m63*, *d14m166*, *d6m44*, *d19m8*, *d2m413*, *d12m2*, and *d17m132*.

The marker names indicate that the markers come from the mouse-chromosomes 2, 6, 12, 14, 17 and 19. Previous research by Schadt et al. (2003) already implicated markers on chromosomes 2 and 19 for fatmass. Using the same data, our analysis not only confirms these findings but also implicates five markers on chromosome 6, 12, 17, 14 with similar significance levels. Interestingly, a previous study (Mehrabian et al., 1998) has shown that plasma hepatic lipase (HL) activity and fatmass are associated with marker *d2m50*, which provides indirect evidence that the proposed hybrid genetic algorithm yields biologically meaningful results.

Figure 2 Occurrence frequency and regression Z-score: (a) the occurrence frequencies of all the 132 markers and (b) the occurrence frequencies and the regression Z-scores of the top 15 markers

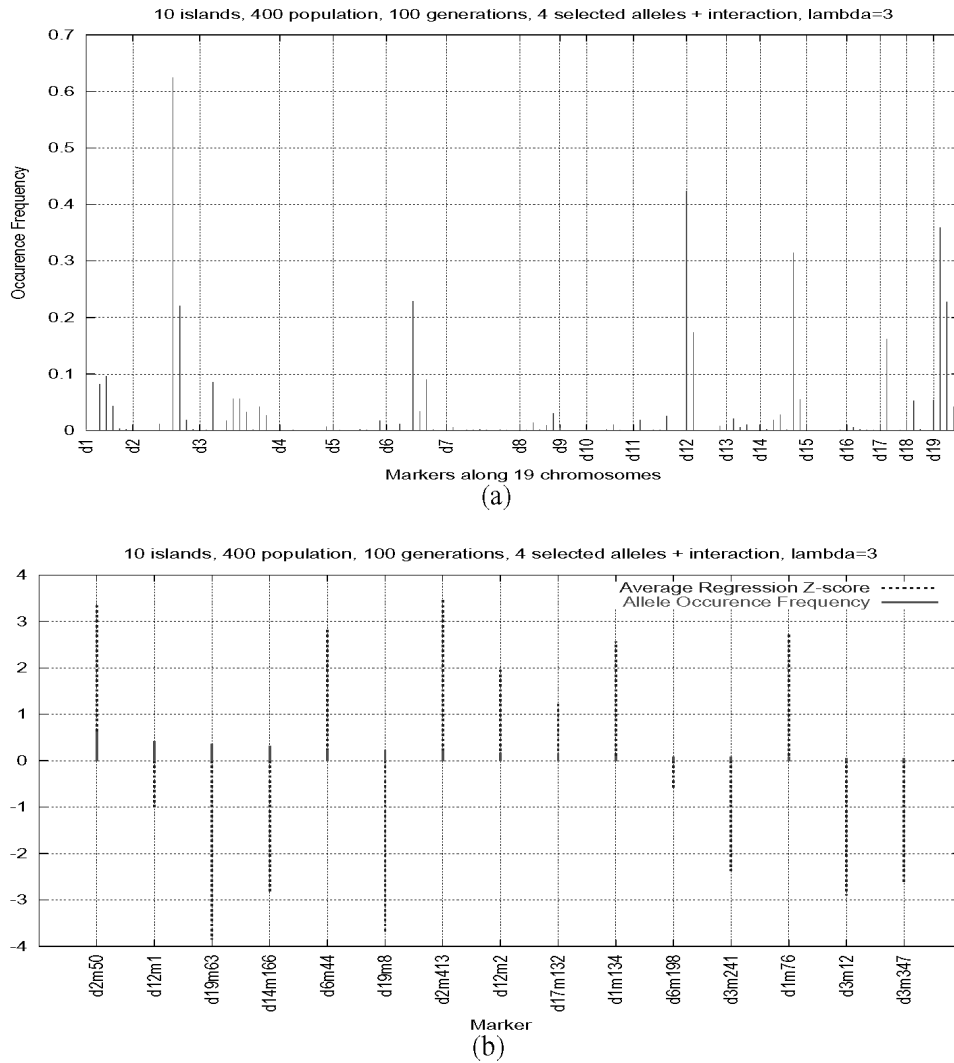


Figure 2 also shows the average Z-score for each marker. It turns out that regression coefficients of 6 out of the top 9 markers are highly significant (the absolute values of the four Z-scores are above 2.5 (corresponding to Wald test p -values smaller than 0.012). Specifically, the most significant markers are $d2m50$, $d12m1$, $d19m63$, $d14m166$, $d6m44$, $d19m8$, and $d2m413$.

3.4 Significant sets of genetic markers

The proposed algorithm does not only determine the importance of the genetic markers but it also reveals a number of significant sets of markers which correspond to near optimal solutions. Since the outcome (fatmass) is a complex trait, many biologists would expect that it is associated with a group of interacting genetic markers, i.e., significant sets of markers may be more interesting to biologists than individual markers.

In the experiment, the final generation contains 12,000 ga-chromosomes, each with 4 ga-genes (markers). The ga-chromosomes (subsets of markers) can be ranked according to their fitness values. We found that the top five sets of markers are: $\{d2m413, d14m166, d17m132, d19m63\}$, $\{d2m50, d3m347, d12m1, d12m8\}$, $\{d2m50, d14m166, d17m132, d19m63\}$, $\{d2m50, d12m2, d14m193, d19m63\}$, and $\{d6m44, d6m198, d19m63, d19m71\}$.

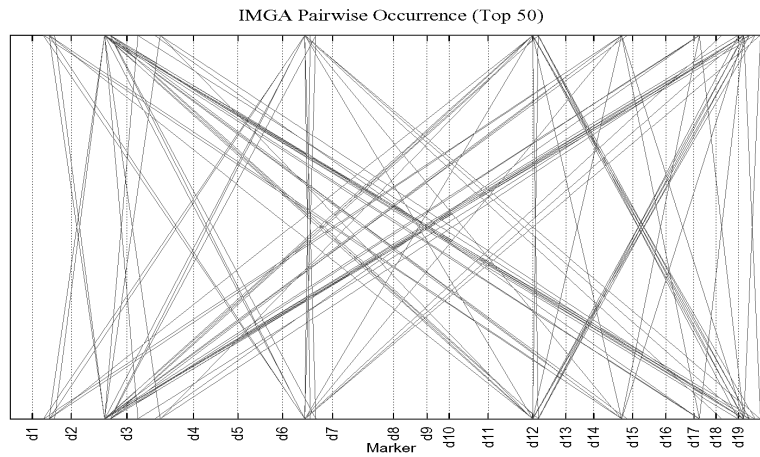
These markers sets can be used in a subsequent (stepwise) linear model fit to elucidate how the markers interact to produce the outcome. This type of postprocessing analysis is descriptive in nature and facilitates the interpretation of the findings. As an example, we report a detailed subsequent analysis of the third most significant set. Table 1 reports the results of using forward stepwise regression with the Akaike's Information Criterion to determine the interaction terms between the four markers. Note that three markers ($d2m50$, $d12m2$ and $d19m63$) are highly significant (p -value ≤ 0.001) predictors. Further, marker $d2m50$ has a highly significant interaction with marker $d12m2$ (p -value ≤ 0.001).

Table 1 Postprocessing analysis involving the markers of the third fittest chromosome (multiple $R^2 = 0.44$). The AIC criterion is used in forward stepwise regression to detect interaction terms. There is evidence for a significant pairwise interaction term between markers $d2m50$ and $d14m193$

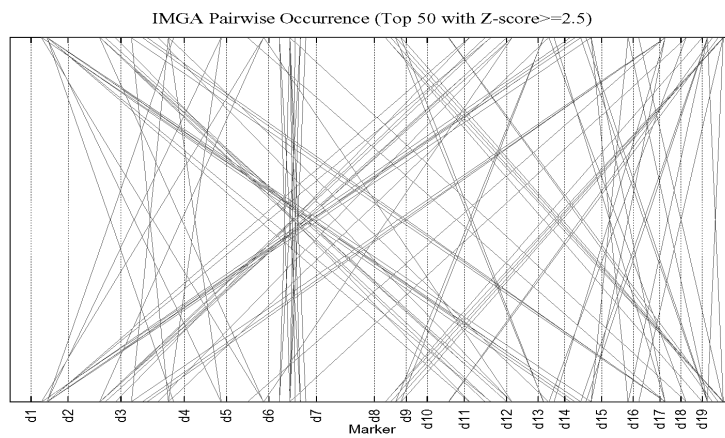
Variable	Estimate	Std. error	z-value	p-value
Intercept	-1.10	0.520	-2.10	3.7e-02
d2m50	1.10	0.230	4.80	9.5e-06
d19m63	-0.23	0.054	-4.20	8.6e-05
d14m193	0.17	0.180	0.96	3.4e-01
d12m2	0.75	0.200	3.80	3.1e-04
d2m50:d14m193	-0.12	0.084	-1.40	1.7e-01
d2m50:d12m2	-0.33	0.096	-3.50	8.8e-04

3.5 Pairwise interaction analysis

From the final generation of the algorithm, we can also compute the cooccurrence frequency of each pair of genetic markers. The top six pairs of markers are ($d2m50$, $d12m1$), ($d2m50$, $d19m63$), ($d2m50$, $d19m8$), ($d12m1$, $d19m63$), ($d12m1$, $d19m8$), and ($d2m50$, $d12m2$). Figure 3 shows the top 50 markerpairs discovered by the algorithm. From the figure, one can find that chromosomes 2, 6, 12, and 19 have the many interactions with each other.

Figure 3 The top 50 most important markerpairs discovered by the IMGA algorithm

From the experiments, we also examine the effect of interaction terms in the fitness function. Figure 4 shows the top 50 markerpairs with significant interaction (the absolute value of the Wald test Z-score for the interaction term is bigger than 2.5, corresponding to a false positive rate of 0.012). We find clear evidence for significant interactions between the markers.

Figure 4 The top 50 markerpairs with significant interaction

4 Conclusion

We developed a novel hybrid algorithm for feature selection by combining island model genetic algorithms with ridge regression, which is a classical, penalised likelihood approach. The new algorithm has been implemented for a continuous outcome but can easily be applied to class outcomes by using a suitable penalised likelihood as fitness function. We used the algorithm to identify sets of genetic markers that influence fatness in mice but we have also applied it to gene expression microarray data (involving thousands of features) from the same mice (unreported data).

As discussed earlier, many GA-based feature selection algorithms involving class outcomes are based on the crossvalidation error rate as fitness function. In this paper, we focused on a simple and fundamental prediction method (least squares regression) for *continuous* outcomes since only 69 observations were available and ease of interpretation was essential. By using a ridge regression based objective function we avoided problems with fitting the linear model due to the low sample size and multicollinear (highly correlated) features. We have demonstrated that the ridge regression based hybrid genetic algorithm is an effective tool for exploring interaction effects between genetic markers. Future research should systematically compare this approach with that of others (Carlborg et al., 2000; Ljungberg et al., 2002, 2004).

Acknowledgement

We are grateful for valuable conversations with Dr. Ashish Ghosh, Machine Intelligence Unit Indian Statistical Institute and Joerg Zimmermann, University of Bonn, Germany. This work was supported in part by the Scripps Program Project grant #1U19AI063603-01.

References

- Brown, P.J. (1994) *Measurement, Regression and Calibration*, Oxford University Press, Oxford, UK, ISBN: 0198522452.
- Carlborg, O., Andersson, L. and Kinghorn, B. (2000) 'The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci', *Genetics*, Vol. 155, No. 4, pp.2003–2010.
- Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*, Chapman and Hall, CRC, Boca Raton, Florida, USA, ISBN: 0412042312.
- Feiveson, A.H. (1994) 'Finding the best regression subset by reduction in nonfull-rank cases', *SIAM Journal on Matrix Analysis and Applications*, Vol. 15, No. 1, pp.194–204.
- Goldberg, D. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Publishing Company, New York.
- Golub, G., Heath, M. and Wahba, G. (1979) 'Generalised cross-validation as a method for choosing a good ridge parameter', *Technometrics*, Vol. 21, No. 2, pp.215–223.
- Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, ISBN: 0387952845.
- Holland, J. (1975) *Adaptation in Natural and Artificial Systems*, The University of Michigan Press, 2nd ed., 1992, MIT Press, ISBN: 0262580969.
- Keedwell, E. and Narayanan, A. (2003) *Genetic Algorithm for Gene Expression Analysis*, Evo Workshops 2003, Essex, UK, April 14–16, pp.76–86.
- Lawrence, D. (1991) *Handbook of Genetic Algorithms*, van Nostrand Reinhold, New York.
- Li, L., Darden, T.A., Weinberg, C.R., Levine, A.J. and Pedersen, L.G. (2001a) 'Gene assessment and sample classification for gene expression data using a genetic algorithm/*k*-nearest neighbor method', *Combinatorial Chemistry and High Throughput Screening*, Vol. 4, No. 8, pp.727–739.
- Li, L., Weinberg, C.R., Darden, T.A. and Pedersen, L.G. (2001b) 'Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method', *Bioinformatics*, Vol. 17, No. 12, December, pp.1131–1142.

- Liu, J., Iba, H. and Ishizuka, M. (2001) 'Selecting informative genes with parallel genetic algorithms in tissue classification', *Genome Informatics*, Vol. 12, pp.14–23.
- Ljungberg, K., Holmgren, S. and Carlborg, O. (2002) 'Efficient algorithms for quantitative trait loci mapping problems', *Journal of Computational Biology*, Vol. 9, No. 6, pp.793–804.
- Ljungberg, K., Holmgren, S. and Carlborg, O. (2004) 'Simultaneous search for multiple qtl using the global optimization algorithm direct', *Bioinformatics*, Vol. 20, No. 12, March, pp.1887–1895.
- Mehrabian, M., Wen, P-Z., Fislser, J., Davis, R.C. and Lusic, A.J. (1998) 'Genetic loci controlling body fat, lipoprotein metabolism, and insulin levels in a multifactorial mouse model', *Journal of Clinical Investigation*, Vol. 101, No. 11, June, pp.2485–2496.
- Schadt, E.E. *et al.* (2003) 'Genetics of gene expression surveyed in maize, mouse and man', *Nature*, Vol. 422, March, pp.297–302.
- Silver, L.M. (1995) *Mouse Genetics*, Oxford University Press, Oxford, UK, ISBN: 0-19-507554-4.