Utilization of gap-weighted kernels for genomic analysis in a multi-breed beef cattle population.

G. Vander Voort[*], S. Miller[*], Z. Wang[¶], G. Plastow[¶], S. Moore[¶]

Introduction

Two general structural properties of genomic data resulting from genome-wide dense marker maps are a larger number of predictor variables relative to dependent variables and high correlation among some of the predictor variables. Modeling the relationship of genotype to phenotype is also problematic due to limited knowledge of the underlying biological mechanisms and overfitting due to the large number of predictors.

Partial least squares (PLS) has been shown effective in analysis of data with genomic data structural properties (Wold (1994)). Kernel modification of the PLS algorithm has been shown to accomodate more complex models in a computationally efficient manner (Rosipal and Trejo (2001)). Suitability of kernel structure is related to data type (Shawe-Taylor and Cristianini(2004)).

This study was conducted to investigate the application of gap-weighted subsequence kernels utlilizing a PLS algorithm to predict breed structure in a cross-bred cattle population from genomic information.

Material and methods

Phenotypic and Genotypic Data. The 671 cattle born between 1998-2006 used in this study were a result of an ongoing crossbreeding program within the University of Guelph beef herd including a predominantly SimmentalxAngus maternal line with some terminal sire usage. Matings utilized AI breeding with predominantly purbred sires. The average animal composition (%) of the major breeds was Angus (40), Simmental (25), Charolais (10), Piedmontese (9). Gelbvieh (4) and Limousin (1.0). These animals were genotyped at the Alberta Bovine Genomics Laboratory, University of Alberta using the 50K Illumina BeadChip®. From the total of 56,947 SNP across 29 autosomal chromosomes after filtering for minor allele frequency 10% and deviation from Hardy-Weinberg equilibrium 38,745 SNP were utilized.

Statistical Analyses. A Partial Least Squares (PLS) algorithm was used to test the efficacy of alternate kernels in genomic data analysis to predict Angus breed composition as the dependent variable. Cattle were divided into two data sets. The training data set had a sample size of 494 with the balance of 671 being included in the test data set. Individuals were included in the test data set to minimize pedigree connection between the two data sets. The training dataset was used to estimate effects of the design matrix (kernel). Estimates derived in the training data set were used to predict breed proportion in the test data set. Correlation

of observed breed proportion and predicted breed proportion were used for evaluation of model efficacy in prediction.

Using a sparse PLS algorithm (Chung and Keles, (2010)) Angus breed proportion was regressed on number of copies of the second allele of the SNP genotype (0,1,2) over 38,745 SNP. Angus breed proportion was chosen given its dominance. By adjustment of tuning parameters (eta=0.8) a subset of the top 106 significant SNP were selected. This was done to facilitate formation of the gap-weighted kernel.

Two kernels were studied. The first kernel utilized was a polynomial kernel. Training kernel element for animal i and j was $K(i, j)=(dot(X(i, :),X(j, :))+1)^2$, where training kernel matrix element K(i, j) (j=i for diagonal) was the dot product of SNP vectors X(i, :) and X(j, :) of length 106. The kernel for validation in the test dataset for test group animal i and training group animal j was $K\_t(i, j)=(dot(Xt(i, :), X(j, :))+1)^2$, where test kernel element K_t(i, j) was the dot product of SNP vector Xt(i, :) SNP vector X(j, :) (Rosipal and Trejo (2001)).

The second kernel studied was a gap-weighted subsequence kernel. The key idea of the gap-weighted subsequence kernel is the weighting of the occurrence of a subsequence in a vector (string) by the degree of continuity in the input vector. For this study a subsequence of length 3 (sub) along with a weighting factor (λ) of 0.5 (1= no weighting) were utilized. Training and test gap-weighted kernels were constructed using a dynamic programming algorithm of Shawe-Taylor and Cristianini (2004). Training gap-weighted kernel matrix elements, Kgap(i, j) for animal i and j was gap_wt(X(i, :),X(j, :)) where gap_wt(sub=3, λ=0.5) was the subsequence comparison of SNP vectors X(i, :) and X(j, :) of length 106. Likewise, the gap-weighted test kernel element for animal i in test dataset and animal j in training data set was gap_wt(Xt(i, :),X(j, :)) where gap_wt(sub, λ) was the subsequence comparison comparison of SNP vectors Xt(i, :) and X(j, :).

Angus proprtion estimates were calculated with NIPALS kernel PLS (KPLS) algorithm (Rosipal and Trejo (2001)). Breed proportion values were regressed on the training and test kernels previously defined. KPLS estimates were multipled by test kernels to predict Angus breed proportion. Correlation between predicted and observed breed was recorded. Also, as a point of comparison PLS estimates from reduced SNP panel selection shown in Figure 1 were used to predict Angus breed proporton in training and test data.

Results and discussion

Angus breed proportion estimates of the 106 SNP selected with threshold parameter of 0.8 by sparse PLS algorithm (Chung and Keles, (2010)) are shown in Figure 1. Values ranged from -0.02 to 0.027.

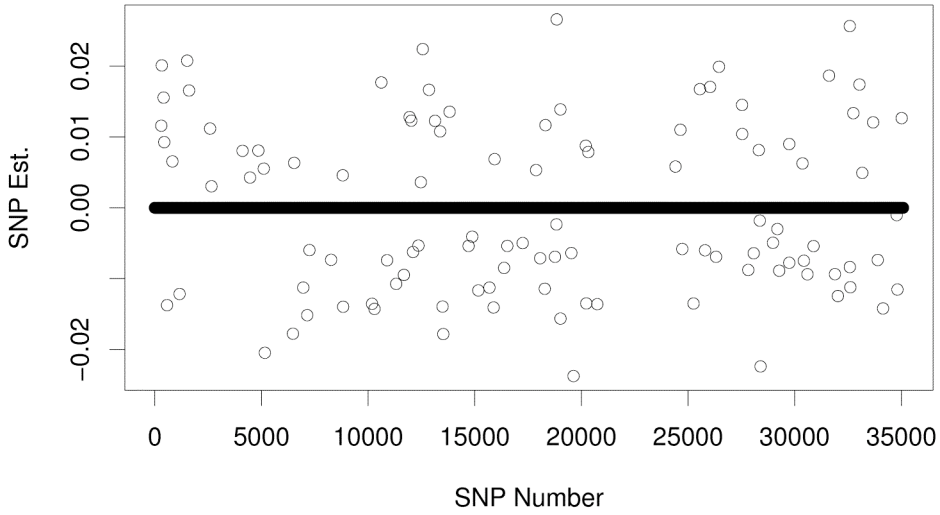

Figure 1. Individual SNP effect estimates predicting Angus breed proportion with a PLS algorithm.

Across all methods used, correlations between predicted and observed Angus breed proportion were always higher in the training datasets compared test data (Table 2). Only KPLS with the gap-weighted subsequence kernel had correlation different from zero.

Table 2. Correlation of observed and predicted Angus breed proportion

	PLS	KPLS(polynomial)	KPLS(gap_wt)
Training set	0.89	0.86	0.41
Test set	0.01	0.008	0.25

Correlation between predicted and observed Angus breed proportion was less than 1 for all methods. The observed value of Angus breed proportion was based on recorded pedigree. Given the effect of Mendelian sampling and crossbreeding over time, and animal registered

as 0.5 Angus based on pedigree may not transmit that exact proportion of Angus alleles to cross-bred offspring.

Kernel PLS utilizing polynomial kernels has been shown to be effective in estimation of genomic effects for traits such as tenderness (Vander Voort et al. (2009)). Gap-weighted kernels have been used in the study of structured data such as DNA sequences to measure pattern differences in insertions or deletions due to mutation (Shawe-Taylor, J. and Cristianini,N. (2004)). However, due to the computational demand of calculating gap-weighted subsequence kernels a smaller SNP panel was studied. Although, this kernel performed better in the test data, lower correlation between predicted and observed Angus breed proportion in training data set indicates this kernel may not be suited in the estimation of genomic effects in smaller SNP panels. Given the breed structure in the data a panel of only 106 SNP may not contain sufficient pattern differences across the 6 breed groups represented in the data.

Conclusion

Based on improved correlation between predicted and observed values in a test data set gap-weighted subsequence based KPLS supports use in analysis of genomic data. However reduced correlation in the training data set indicates further work is needed to study the interaction of data and kernel structure in application of this kernel particularly in the analysis of smaller SNP panels.

References

Chun, H. and Keles, S. (2010). *Journal of Royal Statistical Society, Series B,* 72(1):3-25.

Meuwissen T.H.E., Hayes, B.J, Goddard, M.E. (2001). *Genetics*, 157:1819-1829

Rosipal,R. and Trejo,L. (2001). *Journal of Machine Learning Research,* 2:97-123

Shawe-Taylor,J. and Cristianini,N. (2004). *Kernel Methods for Pattern analysis*. Pages 344-396.

Vander Voort G., Kelly, M., Miller S. *et al. (2009) J. Anim Sci.* 87(E-Supp2):314.

Wold,S. (1994). *Methods and Principles in Medicinal Chemistry van de Waterbeemd Ed. Verlag-Chemie*. Pages 71-112.