Comparison of BLUP and Reproducing kernel Hilbert spaces methods for genomic prediction of breeding values in Australian Holstein Friesian cattle

K.V. Konstantinov[*†] *and B.J. Hayes*[*]

Introduction

Traditionally breeding values for economically important traits in dairy cattle in Australia have been calculated by using phenotypic and pedigree information. Meuwissen et al. (2001) introduced genomic selection as a selection strategy based on direct genomic values (DGV) predicted from massive marker information, e.g. single nucleotide polymorphism (SNP). Word wide application of genomic selection has started recently (Harris et al., 2008; Van Raden, 2008, Hayes et al. 2009).

A variety of methods have been suggested in the literature for the estimation of DGVs. Those methods can be classified as parametric and nonparametric or semi-parametric methods. Considerable effort has been made worldwide in comparison the accuracies of those methods (Hayes et al. 2009; Moser et al., 2009; Verbyla et al., 2009; Gonzales-Recio et al., 2009). Although most of the studies reported very similar correlations between genomic and traditional breeding values, there were trait specific differences in correlations among the methods.

So far very few studies (de los Campos et al., 2009) have reported results of genomic breeding values calculated from semi-parametric methods such as Reproducing Kernel Hilbert Spaces (RKHS) using real data with large number of SNP.

The objective of this study was to compare the accuracy of DGV from parametric methods, namely SNP-BLUP (Mozer et al., 2009) to the accuracy of DGV from RKHS using field data on six traits in Australian Holstein Friesian data.

Material and methods

Phenotypic and Genomic data. Daughter trait deviations (DTD) were calculated for milk, fat and protein yield and for fertility, overall type and survival. The number of sires in training set varied between 1094 to 1858. Sires were required to have at least 20 progeny from inclusion in the data set. The validation set included 257 bulls from years 2007, 2008 and 2009 which were offspring from sires with records in the training set. The number of progeny for the bulls in the validation set varied from 25 to 130, 1 to 112, 2 to 60 and 2 to 132 for yield, fertility, overall type and survival, respectively.

[*] Biosciences Research Division, Department of Primary Industries Victoria, 1 Park Drive, Bundoora 3083, Australia

[†] K.V. Konstantinov acknowledges the financial support from Australian Dairy Herd Improvement Scheme.

The SNP data contained Australian Holstein Friesian bulls genotyped for the Illumina Bovine 50K array. After editing, 43115 SNPs remained in the training set. Details of the quality control applied to the SNP data are described by Hayes et al. (2009).

Statistical analyses. ***SNP-BLUP***. The model currently in use by ADHIS for estimation of genomic breeding values is as follows: $y = 1_n \mu + Xb + e$, (1).

where y is a vector of phenotypes (DTDs); μ is a unknown constant; 1_n is a vector of ones; X is a design matrix allocating records to the marker effects; The (co)variance matrix of the SNP effects is assumed to be $I\lambda$; e is a vector of random residuals with (co)variance matrix = $I\sigma_e^2$. The system of equations based on model (1) was solved iteratively by Preconditioned Conjugate Gradient Method (PCG). A 10-fold cross-validation procedure with golden segment search using the training data set was performed to locate the best λ within a given range.

RKHS.This method was first introduced by Gianola and van Kaam (2008) and used by Gonzalez-Recio et al. (2009). A RKHS regression is a semi-parametric approach that allows inference regarding functions, e.g., genomic breeding values, without making strong prior assumptions. Following Gianola and van Kaam (2008) and Gonzalez-Recio et al. (2009) the model for calculation DGVs can be formulated as follows: $y = 1\mu + K_h\alpha + e$, (2)

where y is a vector of phenotypes (DTDs); μ is a unknown constant;1 is a vector of ones; K_h is a positive definite matrix of kernels, possibly dependant on bandwidth parameter (h); α is a vector containing non-parametric coefficients that are assumed to be normally distributes as $\alpha \sim N\left(0, K_h^{-1}\sigma_a^2\right)$, with σ_a^2 representing a reciprocal of a smoothing parameter $(\sigma_a^2 = \lambda^{-1})$. The residuals were assumed to be normally distributed with co-variance matrix R, with $R = N^{-1}\sigma_e^2$, where $N = \{n_{ij}\}$ is a diagonal matrix with elements equal to the number of progeny per sire and σ_e^2 is the residual variance.

Selecting a kernel is the most critical stage in applying kernel-based algorithms (Gianola and van Kaam, 2008; Gonzalez-Recio, 2009). In this research a Gaussian kernel was used. Using the training set, the parameter h was tuned by generalised cross-validation procedure as described by Wahba et al. (2007) and Gianola and van Kaam (2008). In this case the system based on model (2) was solved by matrix inversion. The distancies between genotypes were calculated as described by Gonzalez-Recio (2009) and (Gianola and van Kaam, 2008). After tuning the system (2) was solved by PCG method.

With SNP-BLUP procedure the phenotypes were predicted as: $\hat{y} = [1 \quad X]\hat{b}$. With RKHS the phenotypes were predicted as $\hat{y} = \hat{\mu}1 + K^*(h)\hat{\alpha}$, where a row of $K^*(h)$ had the form $k_t^{'*} = \{K_h^*(x_t - x_j)\}$ with $K_h^*(x_t - x_j)$ being the kernel between the genotype of sire *t* in the validation set and sire *j* in the training set. The same *h* parameter that was tuned with the training set was used. The proxy for accuracy of the methods was the correlation of the DGV, Australian Breeding Values (ABV) or DTD on the predicted values in the

validation set. The regression of DGV on ABV or DTD is also given. Fortran 95 and C++ programs were developed for computing RKHS regressions.

Results and discussion

Estimates of correlations and regression coefficients between predicted genomic breeding values and ABV (DTD) are summarised in table 1. The results from SNP-BLUP are similar to those obtained by Hayes et al. (2010) where the authors compared 8 different methods for calculation of genomic breeding values. There were slight differences in milk, fat and overall type estimates. In another Australian study Verbyla et al.(2009) also found somewhat higher correlations for protein yield and fertility by using Bayesian methods. However, those authors used different dependant variables (de-regressed ABV) in their models and different set of bulls in the training and validation sets.

The correlations between DGVs and DTDs for milk, fat and protein yield are similar to those between DGVs and ABVs but considerably lower for fertility, overall type and survival. This could be due to the fact that for those traits not all bulls from validation set had DTDs. For example only 190, 209 and 243 bulls had DTD for overall type, fertility and survival, respectively.

When genotype distancies were calculated as in Gonzalez-Recio et al, (2009) SNP_BLUP outperformed RKHS for all traits. The correlations were significantly lower for all traits studied.

When the genotypes distancies were calculated as in Gianola and van Kaam (2008), RKHS performed better compared to SNP-BLUP. For the traits milk, fat and protein yield the differences were small but in favour of RKHS. Significant differences were observed for fertility, overall type and survival. The correlations for fertility from RKHS were almost double the correlations from SNP-BLUP (table 4). Their values are comparable with the value of 0.54 reported by Verbyla et al. (2009) from another Australian study. All correlations for fertility reported by Hayes et al. (2010) obtained from 8 different methods were lower that those from the present study. Again the correlations involving DTDs were lower compare to those with ABV especially for traits other than production. Irrespectively of that RKHS still produced higher correlations for most traits.

Table 1: Correlations and regression coefficients between predicted DGV and ABV (DTD) in the validation data set for the two methods [α]

Trait	Method	Correlations		Regression
		Pearson	Spearman	$\hat{b}$
Milk kg	SNP-BLUP	0.571 (0.533)	0.561 (0.523)	1.224 (1.118)
Fat kg		0.552 (0.509)	0.536 (0.497)	1.512 (1.160)
Protein kg		0.519 (0.487)	0.505 (0.471)	0.918 (1.034)
Fertility		0.344 (0.035)	0.316 (0.049)	0.328 (0.318)
Overall type		0.536 (0.134)	0.581 (0.158)	1.205 (0.617)
Survival		0.351 (0.029)	0.333 (0.138)	1.614 (0.273)
Milk kg	RKHS[1]	0.466 (0.427)	0.451 (0.411)	1.353 (1.323)
Fat kg		0.486 (0.447)	0.480 (0.443)	1.100 (1.029)
Protein kg		0.399 (0.359)	0.394 (0.353)	1.146 (1.059)

Table 1: Cont.				
Fertility		0.365 (0.085)	0.344 (0.116)	0.743 (0.948)
Overall type		0.495 (0.100)	0.521 (0.149)	1.127 (0.841)
Survival		0.225 (0.023)	0.220 (0.063)	1.143 (0.207)
Milk kg	RKHS[2]	0.596 (0.555)	0.575 (0.531)	0.776 (0.737)
Fat kg		0.554 (0.510)	0.543 (0.499)	1.294 (1.129)
Protein kg		0.522 (0.488)	0.509 (0.475)	1.049 (1.184)
Fertility		0.609 (0.105)	0.601 (0.148)	0.998 (0.936)
Overall type		0.591 (0.178)	0.621 (0.224)	1.327 (0.737)
Survival		0.478 (0.061)	0.493 (0.185)	1.518 (1.184)

[α]Estimates in parenthesis are those obtained from correlation (regression) analyses using DTDs

[1]Distances between genotypes calculated as in Gonzalez-Recio et al. (2009)

[2]Distances between genotypes calculated as in Gianola and van Kaam (2009)

Conclusion

The results from this study clearly show that measuring genomic differences (non-Euclidean distances) in different ways yields different predictive ability of RKHS method. To be effective in predicting future phenotypes a proper kernel function must be used. The issue of tuning the bandwidth parameter still needs to be considered carefully and MCMC algorithms maybe required to obtaining good estimates. Also RKHS performed better when sires in the validation set had parents in the training set (results not shown). SNP-BLUP had similar predictive ability as in other studies (Hayes et al., 2010; Mozer et al., 2009) irrespectively of the fact that different tuning procedure was used in the present study.

References

De los Campos, G., Gianola, D., Weigel, K.A. and Rosa, G.J.M. (2009). In *Proc SGLPGE*, posters.

Gianola, D. and Johannes B. C. H. M. van Kaam (2008). *Genetics*, 178:2289–2303.

Gonzalez-Recio, O., Gianola, D., Rossa, G.J.M., Weigel, K.A. and Kranis, A. (2009*). Genet .Sel. Evol.*, 41/1:1–10.

Harris, B.L., Johnson, D.L., and Spelman, R.J. (2009). *Interbull Meeting, Niagara Falls, USA*.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J. and Goddard, M. E. (2009*). J. Dairy Sci.*, 92:433–443.

Hayes, B.J., Crump, R.E., Hichey, Bowman, P.J., Tier, B., Verbyla, K., Usai, Chamberlain, A.J., Pryce, J. Khatkar, M.S., Thompson, Goddard, M.E. and Raadsma, H.W. (2010*). In prep.*

Klara Verbyla, Hayes, B. J. and Goddard, M. E. (2009). *Genet. Res. Camb.*, 91:307–311.

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

Mozer, G., Tier B., Crump, R. E., Khatkar, M. and Raadsma, H. W. (2009). *Genet. Sel. Evol*., 41:56.

Van Raden, P.M., Van Tassel, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D. Taylor, J.F. and Schenkel, F. (2008). J. Dairy Sci., 91:4414-4423.

Wahba, G. (2007*). Tech. Report 1136, Dept. Stat. Univ. Wisconsin, 1300 Univ. Ave., Madison, WI 53706*.