Paper presented at the 58th Annual Conference of the European Association for Animal Production Dublin, 26.-29.8. 2007, Session code G18.2

A comparison of different regression methods for genomic-assisted prediction of genetic values in dairy cattle

J. Sölkner^{1,2,3}, B. Tier^{1,4}, R. Crump^{1,4}, G. Moser¹, P. Thomson^{1,2} and H. Raadsma^{1,2} ¹ CRC for Innovative Dairy Products, William Street, Melbourne, Vic, 3000, Australia ²University of Sydney, Dairy CRC, Camden, NSW 2570, Australia ³University of Natural Resources and Applied Life Sciences, A-1180 Vienna, Austria ⁴University of New England, AGBU, Armidale, NSW 2351, Australia

johann.soelkner@boku.ac.at



The problem

Predict breeding values of young bulls using SNP information

- Choice of method considering
 - the large number of SNP
 - the problem of overfitting



Approaches

- BLUP/Bayes on haplotypes
- Kernel Regression
- Principal components regression



Comparison of Methods

- Predictive capacity
 - Observations not involved in the estimation step
- Correlation of observed and predicted values



Australian Data

- 1546 bulls,10715 SNP
- Wide range of birth years
- Many traits
 - APR (total merit)
 - Protein yield
 - Overall type (conformation)
 - Fertility
 - Somatic cell count
- Accurate EBV, treated as proxies to TBV



Methods applied

- Ordinary least squares regression using a genetic algorithm for SNP-weighting
- Partial least squares regression
- SNP selection via Least Angle Regression



Regression using a genetic algorithm for SNP-weighting

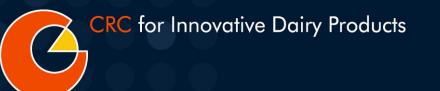
- Apply genetic algorithm for variable selection
- Final regression coefficient of a SNP depends on how often it was in a chosen model and how good these models were
- Internal cross-validation was applied

Partial least squares regression

- Standard tool for dimensionality reduction
- Similar to principal components regression, but including independent AND dependent variables for component selection
- Internal cross-validation was applied

OLSR using least angle regression for variable selection

- Efron et al., 2004
- Model selection algorithm, much less greedy version of traditional forward selection methods
- No internal cross-validation



Comparison for "unseen" data

- Split of data set into training (1346) and test (200) sets
- 5 replicates
- Make a set of 1000 from 5 x 200
- Construct subsets of "contemporary" bulls
 - YOB 1990 1999 \Rightarrow n=587
 - YOB 1997 2001 \Rightarrow n=377



APR (total merit)

	Correlations		
	PLSR	OLSR-GA	LAR (100)
All YOB	0.867	0.801	0.725
1990-1999	0.788	0.695	0.559
1997-2001	0.740	0.623	0.460



Protein yield

	Correlations		
	PLSR	OLSR-GA	LAR (100)
All YOB	0.882	0.781	0.740
1990-1999	0.793	0.577	0.528
1997-2001	0.741	0.500	0.446



Overall type

	Correlations		
	PLSR	OLSR-GA	LAR (100)
All YOB	0.779	0.630	0.651
1990-1999	0.672	0.423	0.441
1997-2001	0.644	0.449	0.442



Fertility

	Correlations		
	PLSR	OLSR-GA	LAR (100)
All YOB	0.720	0.718	0.645
1990-1999	0.564	0.586	0.435
1997-2001	0.618	0.582	0.451



Somatic cell count

	Correlations		
	PLSR	OLSR-GA	LAR (100)
All YOB	0.470	0.425	0.360
1990-1999	0.449	0.385	0.377
1997-2001	0.505	0.446	0.463



2

Conclusions

- Real-life example of genome wide selection
- Regression methods are very useful
- Of the methods tested, with their current settings, PLSR was best

