Noname manuscript No.
(will be inserted by the editor)

Developing a parsimonius predictor for binary traits in sugar beet (*Beta vulgaris*)

Filson Nazzarini · Simone Marini · Piergiorgio Stevanato · Nelppo Biscicari

Received: 05 August 2014 / Accepted:

Abstract Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed.

Keywords binary traits · genomic predictions · parsimonious predictor · sugar beet

1 Introduction

The primary goal of breeding schemes in farm animals and crops is generally to increase the agricultural output. Production traits are typically quantitative continuous variables (e.g. milk yield in dairy cattle, or per hectare yield in maize and rice). Many traits of importance in plant and animal breeding follow nonetheless a discrete categorical distribution, both ordered (e.g. calving ease in cattle, grain texture in rice) and unordered (e.g. grain pigmentation in rice, coat colour in cattle). A special case is that of binomial traits, which can take up only two different values, like disease resistance/susceptibility or

Filippo Biscarini and Nelson Nazzicari contributed equally to the work.

F. Biscarini
Fondazione Parco Tecnologico Padano
Tel.: +123-45-678910
E-mail: filippo.biscarini@tecnoparco.org

S. Marini
second address

N. Nazzicari
Fondazione Parco Tecnologico Padano
Tel.: +123-45-678910
E-mail: nelson.nazzicari@tecnoparco.org

P. Stevanato
address

presence/absence of a morphological characteristic. Annual bolting (flowering) behaviour and root vigor are examples of binomial traits of agronomic importance in sugar beet (*Beta vulgaris*).

Advances in biotechnology and genomics, and the advent of high-density molecular markers (especially sinlge-nucleotide polymorphisms, SNP) genotyping have led to the emergence of molecular breeding [6].

One exciting application of molecular breeding are genomic predictions ([5]): something on genomic predictions in general. Genomic predictions in sugar beet already done both for continuous ([4,9]) and binary ([2]) traits.

The concept of parsimony: when many possible predictors are available, it is useful to select a subset to limit analysis cost and time. Moreover: use the minimun necessary information set, occam razor ([3]), and so forth.

A model need to be simpler than the data the it fits/explains (e.g. knn with k=1)

Given two models that fit the data, the simplest has to be chosen (Occam's razor)

As the technology advances, and available predictors grow, not only the prediction precision becomes important, but also the actual cost must be considered.

Sugar beets in particular: we work on root vigor [2].

In this paper we propose statistical methods to highlight and select the most useful predictors given a set. We started on real world data and validated our approach on a XXX dataset. We found that it is possible to strongly reduce the dimension of the predictors set and still achieve high performance.

2 Material and methods

2.1 Plant material and SNP genotypes

Root vigor. Available data. SNP technology used, imputation.
Copypaste from other articles. Dataset description. Text with citations [8] and [7].

2.2 Predictor development procedure

A two-step approach was adopted for the construction of a parsimonious predictor for root vigor.

- a ranker to rank the various available predictors (SNPs in our case). We used the BOSS algorithm - this is an iterative step. we progressively reduced the predictors set, taking away the laest useful predictor and applying to the resulting subset a ridge logistic regression apprach. Thus, we obtained as many performances estimation as the number of original predictors.

2.2.1 Rank of predictors

This explain the BOSS algorithm

2.2.2 Selection of predictors and classification method

We take one predictor out at each iteration You put the model formula for ridge logistic regression

2.2.3 Predictive ability

Cross validation: how many times, what fractions. Explanation of error rate and other parameters (ROC?)

2.3 Comparison with another method to rank predictors

Another ranker: why use one, and its description. P value and SNP effect (as it is done in GWAS)

2.4 Software

R, weka, perl.

3 Results

Possible charts: - Precision as a function of the number of predictors. - Breakdown of two types of error.

If possible: probability of assignment as a function of predictors, maybe with ROC curve? Maybe in discussion?

4 Discussion

General overview why error rates are not evenly distributed? Reminder: it works very well because of LD and H2

4.1 SNP effects

Manhattan plot with BOSS weights and weights from the other articles, somehow compared (same chart? two charts? only ten best?).

Do the peaks make sense from the biological perspective?

4.2 Relative performance of rankers

why using Pvalues and not other standard rankers (e.g. backward stepwise selection)? Because of the specific nature of the problem

4.3 Genotyping strategies and applications to breeding

genotyping strategies: Costs, possible technologies (gbs, snp chip, macroarrays), implications

applications to breeding: why is it important root vigor early detection. Other binomial traits (e.g. disease resistance) May be applied to bolting (another trait which exhibits binomial distribution), which has been shown to be controlled by multiple genes and influenced by environmental factors ([1]).

Extensions to multinomial traits? Examples?

5 Conclusions

Concluding remarks

Acknowledgements Do we need to ack somebody? (projects?) This research was financially supported by the Marie Curie European Reintegration Grant "NEUTRADAPT".

References

1. Abou-Elwafa, S., Büttner, B., Kopisch-Obuch, F., Jung, C., Müller, A.: Genetic identification of a novel bolting locus in *Beta vulgaris* which promotes annuality independently of the bolting gene *B.* Molecular Breeding **29**, 989–998 (2012)
2. Biscarini, F., Stevanato, P., Broccanello, C., Stella, A., Saccomani, M.: Genome-enabled predictions for binomial traits in sugar beet populations. BMC Genetics **18**(5), 1–9 (2014)
3. Chaitin, G.: The limits of reason. Scientific American **294**(3), 74–81 (2006)
4. Hofheinz, N., Borchardt, D., Weissleder, K., Frisch, M.: Genome-based prediction of test cross performance in two subsequent breeding cycles. Theoretical and Applied Genetics **125**(8), 1639–1645 (2012)
5. Meuwissen, T., Hayes, B., Goddard: Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**(4), 1819–1829 (2001)
6. Moose, S.P., Mumm, R.H.: Molecular plant breeding as the foundation for 21st century crop improvement. Plant physiology **147**(3), 969–977 (2008)
7. Saccomani, M., Stevanato, P., Trebbi, D., McGrath, J.M., Biancardi, E.: Molecular and morpho-physiological characterization of sea, ruderal and cultivated beets. Euphytica **169**(1), 19–29 (2009)
8. Stevanato, P., Broccanello, C., Biscarini, F., Del Corvo, M., Sablok, G., Panella, L., Stella, A., Concheri, G.: High-throughput rad-snp genotyping for characterization of sugar beet genotypes. Plant Molecular Biology Reporter pp. 1–6 (2013)
9. Würschum, T., Reif, J.C., Kraft, T., Janssen, G., Zhao, Y.: Genomic selection in sugar beet breeding populations. BMC genetics **14**(1), 85 (2013)