Predicting pig digestibility coefficients with microbial and genomic data using machine learning prediction algorithms

Céline Carillier-Jacquin, Vanille Déru, Llibertat Tusell, Alban Bouquet, Laval Jacquin, Hélène Gilbert

Predicting pig digestibility coefficients with microbial and genomic data using machine learning prediction algorithms

C. Carillier-Jacquin[1*], V. Deru[1,2], L. Tusell[3], A. Bouquet[2,4], L. Jacquin[5] and H. Gilbert[1]

[1] GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet Tolosan, France ; [2] France Génétique Porc, 35651 Le Rheu, France; [3]Animal Breeding and Genetics Program, Institute of Agriculture and Food Research and Technology (IRTA), Barcelona, Spain ; [4]IFIP-Institut du Porc, 35650 Le Rheu, France ; [5]CIRAD, BIOS, UMR AGAP, Montpellier, France ; [*]celine.carillier-jacquin@inrae.fr

Abstract

Classical methods as genomic BLUP performs well for genomic prediction of polygenic trait, but does not consider interaction between genes or between genes and other information such as host genetic or microbial data. This study aims at comparing several methods including parametric and machine learning methods to predict digestive coefficient using genomic, microbial and both genomic and microbial information. Considering only microbial data led to the best prediction accuracies for digestive coefficients, whereas considering only genomic data performed worst. BLUP, RKHS and GSVM gave the best prediction accuracies except when combined genomic and microbial data was used. Combining microbial and genomic data did not improve prediction accuracies for all traits and methods considered in this study. Thus, considering microbial information is crucial to predict digestive efficiency and interactions between host genetic and faecal microbial information seem to be limited.

Introduction

Genetic evaluation as a sum of additive trait effects using genomic Best Linear Unbiased Predictor (BLUP) method, as defined by Meuwissen et al. (2001), is known to be more accurate than other methods for polygenic traits in animal genetics (Zhu et al., 2021; Wang et al., 2019). Nevertheless, some non-parametric machine learning methods allow considering interacting genes, or interaction between microbiota and host genetic and major effect of some gene or OTUs (Operational Taxonomic Unit). The objective of this study was to compare parametric approaches, and machine learning methods that can capture such interactions, in terms of prediction accuracy of digestive efficiency considering microbial data, genomic data, or both information. Indeed, in pigs, earlier studies showed that both host genetic and faecal microbial information contribute to digestive coefficients variability (Déru et al, 2021a).

Materials & Methods

Statistical analyses. Corrected phenotypes were predicted using six different methods: BLUP, Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net (EN), Reproducing Kernel Hilbert Space (RKHS), Support vector Machine using Linear (LSVM) and Gaussian (GSVM) kernel. BLUP (aka Ridge regression) and LSVM are linear methods, which can capture additive polygenic effects of SNP. The major difference between BLUP and LSVM lies in the loss function used for measuring the corrected phenotype prediction error i.e. squared euclidean loss versus epsilon-insensitive loss respectively. EN and LASSO are variable selection methods, which allow considering non-polygenic traits or more important

effects of some OTUs. RKHS and GSVM methods are approaches using kernels, which allow considering additive effects and interactions between genes or OTUs. BLUP, LASSO and EN were performed using glmnet R package (Friedman et al. 2010), whereas RKHS and SVM were performed using KRMM (Jacquin et al., 2016) and kernlab R packages (Karatzoglou et al., 2004), respectively. Hyperparameters for RKHS and SVM methods were optimised by cross-validated grid-search as in Jacquin et al. (2016). For LASSO, the regularization parameter was optimised using the cv.glmnet function with its default values (Friedman et al., 2010). For each trait, the six prediction methods were used with genomic data only, microbial data only, and combined genomic and microbial information. Microbial data was included in the model as the centered and standardized log abundancies of each OTU (i.e. matrix of 2,399 OTUs for 1,082 animals). Genomic data was the matrix of 48,919 SNP for 1,082 animals and did not include any pedigree information. Combining genomic and microbial data was considered using the concatenation of the matrix of SNP and the matrix of OTUs (i.e. matrix of 51,318 columns (48,919 SNP + 2,399 OTUs) and 1,082 rows (animals)). Considering the largest amount of SNP compare to the amount of OTUs, combining genomic and microbial information could not be ideal for methods that did not reduce the number of variables. For this reason, we also tested combination of microbial data with genomic data reduced to 10,000 SNP selecting equidistant SNP.

Prediction accuracies. Cross validation was used to estimate prediction accuracy, by splitting randomly our population of 1,082 animals into two sets: a training set (722 animals, i.e. 2/3 of the total population) and a validation set (360 animals). Prediction accuracy was estimated for each trait and predicting method for the validation set, as the Pearson correlation between records corrected for diet and batch and predicted traits. Confidence intervals of the Pearson correlation were computed using cor.test function of psych R package (Revelle, 2021).

Data. Data used for 1,082 Large White male pigs feed with a conventional European diet or a high-fiber diet were available (Déru *et al*., 2021b). Digestibility coefficients of energy (DC_E), and nitrogen (DC_N) were predicted from faecal sampling at 16 weeks of age using near infrared spectrometry (Déru *et al.*, 2021a). Microbial information was obtained from sequencing of the V3-V4 regions of the 16S rRNA contained in the same faecal samples as used for DC predictions. After data curation, the 2,399 operational taxonomic units (OTUs) were kept filtering out OTUs present in less than five samples and with an average abundance lower than 0.001%. Genomic data from the 70K SNP GeneSeek GGP Porcine chip was filtered according to minor allele frequency, call rate, call freq and Hardy-Weinberg equilibrium leading to 48,919 markers available for all pigs.

Results and discussion

Prediction accuracies and confidence intervals estimated in the validation set for DC of nitrogen and energy using the six prediction methods are presented in Table 1. Rankings of the six methods were similar for all DC. Among methods, BLUP, RKHS and GSVM gave better accuracies compared to other methods when considering genomic or microbial information separately. Selection variable methods used in this study (i.e. LASSO and EN) gave the worst prediction accuracies with separate information, and were best when combined microbial and all the genomic information was used. If some genes or OTUs had higher effects on trait than other ones, we expected that these methods outperformed BLUP. We can suppose that all SNP and OTUs seemed to contribute equally to digestive coefficients. LSVM did not performed as well as BLUP, probably due to loss function not adapted to our data.

GSVM and RKHS did not outperform BLUP, which could be explained by only few interactions between host genes, between OTUs and between host genetic and microbial information. The better prediction accuracy obtained for BLUP and RKHS in this study is similar to results obtained in other species for a wide range of traits (Wang *et al.*, 2021; Zhu *et al*., 2019). However, Maltecca *et al*. (2019) have reported similar results with LASSO and RKHS when using microbial data to predict pig growth and carcass traits.

Table 1. Prediction accuracies and confidence intervals for digestive coefficients of energy (E) and of nitrogen (N) with six prediction methods[1]

		BLUP[3]	LASSO[1]	EN[2]	RKHS[4]	LSVM[5]	GSVM[6]
Genomic	N	0.38 [0.3;0.46]	0.29 [0.2;0.38]	0.29 [0.2;0.38]	0.38 [0.29;0.46]	0.33 [0.24;0.42]	0.36 [0.27;0.45]
	E	0.43 [0.35;0.51]	0.42 [0.34;0.5]	0.44 [0.36;0.52]	0.47 [0.39;0.54]	0.40 [0.31;0.48]	0.45 [0.36;0.52]
Microbial	N	0.67 [0.61;0.72]	0.60 [0.53;0.66]	0.60 [0.54;0.66]	0.67 [0.61;0.72]	0.58 [0.51;0.64]	0.65 [0.59;0.70]
	E	0.69 [0.63;0.74]	0.63 [0.57;0.69]	0.64 [0.57;0.69]	0.69 [0.63;0.73]	0.62 [0.56;0.68]	0.67 [0.61;0.72]
Genomic and Microbial	N	0.49 [0.41; 0.56]	0.61 [0.54;0.67]	0.61 [0.54; 0.67]	0.48 [0.40;0.55]	0.47 [0.38; 0.54]	0.48 [0.40; 0.55]
	E	0.51 [0.43;0.58]	0.62 [0.55; 0.67]	0.60 [0.53;0.66]	0.52 [0.45;0.59]	0.49 [0.41;0.56]	0.52 [0.45; 0.59]
Genomic (10k) and Microbial	N	0.57 [0.51;0.64]	0.58 [0.52;0.65]	0.59 [0.53; 0.65]	0.57 [0.49; 0.63]	0.55 [0.49;0.62]	0.56 [0.52;0.65]
	E	0.58 [0.43,0.58]	0.60 [0.55,0.67]	0.59 [0.53,0.66]	0.59 [0.52;0.65]	0.56 [0.41;0.56]	0.59 [0.45;0.59]

[1] LASSO = Least Absolute Shrinkage and Selection Operator; EN = Elastic Net; BLUP = Best Linear Unbiased Predictor; RKHS = Reproducing Kernel Hilbert Space; LSVM = Support Vector Machine using Linear kernel; GSVM = Support Vector Machine using Gaussian kernel

For all traits, considering only microbial data led to the best prediction accuracies, whereas considering only genomic data performed worst. The decrease in prediction accuracies when using only genomic information compared to using microbial data only can be explained by the large part of variance of DC traits explained by microbial information in these traits (Déru *et al*. 2021c). Compared to the use of microbial data only, Pearson correlations decreased when both microbial and genomic data were analysed jointly except for variable selection methods certainly due to much larger amount of genomic data (48,919 SNP) compared to microbial data (2,399 OTUs). In practice, LASSO and EN were able to select more OTUs than SNP in the model, thus prediction accuracies were close to the ones obtained with microbial data only. When reducing genomic data to 10,000 equidistant SNP, other methods (i.e. BLUP, RKHS and SVM) performed similarly to the selection variable methods. This confirms our hypothesis that the highest amount of SNP compared to OTUs did not allow methods to perform as well as using only microbial data, at least in the case of DC traits where microbial information seem to be preeminent on genomic information to explain the traits.

Comparing DC traits, Pearson correlations obtained for DC of nitrogen using only genomic data were lower to the one obtained for DC of energy, which is not the case when considering

only microbial or combined data. This could be explained by a lower contribution of host genetic to DC_N variance, as reported also by Déru *et al.* (2021c) with different approaches.

Conclusion

We conclude that microbial information is crucial to predict digestive coefficients. Machine learning methods did not outperform BLUP, which suggest few interactions between SNP, between OTUs, and between microbial and host genetic information. For further study, it could be interesting to estimate prediction accuracies without filtering out rare OTUs, to evaluate a potential prediction improvement when combing both data. The interest of random forest variable importance methods could also be used to explore genomic and microbial data combination.

Funding

This study was supported by the Feed-a-Gene European H2020 project. The authors acknowledge the breeding companies Axiom and Nucleus for providing animals and the UEPR phenotyping station staff in Le Rheu for animal raising and data recording.

References

Déru V., Bouquet A., Labussière E., Ganier P., Blanchet B. et al. (2021a) J Animal Breed Genet 138 :246-58. Déru V., Bouquet A., Zemb O., Blanchet B., De Almeida M.L. et al. (2021b) Preprint bioRxiv.
Déru V., Tiezzi F., Carillier-Jacquin C., Blanchet B., Cauquil L. et al. (2021c) Proc. of the 72nd EAAP, Davos, Switzerland.
Friedman J., Hastie T., and Tibshirani R. (2010) J. Stat. Softw. 33:1.
Jacquin L., Cao T.V., Ahmadi N. (2016) Front Genet, 7,145.
Karatzoglou A., Smola A., Hornik K., Zeileis A. (2004) J. Stat. Softw. 11, 1–20.
Maltecca C., Lu D., Schillebeeckx C., McNulty N.P., Schwab C., et al. (2019). Sci Rep 9, 6574.
Meuwissen T., Hayes B.J., Goddard M. (2001) Genetics 157: 1819–1829
Revelle W. (2021). psych: Procedures for Psychological, Psychometric, and Personality Research. Available at: https://CRAN.R-project.org/package=psych.
Wang X., Miao J., Chang T., Xia J., An B., et al. (2019) PLoS ONE 14(2).
Zhu S., Guo T., Yuan C., Liu J., Li J., et al. (2021) G3, 11(11).