Using machine learning to predict feed intakes of meat sheep from animal traits and ruminal microbiota

Quentin Le Graverand, Christel Marie-Etancelin, J L Weisbecker, Annabelle Meynadier, D Marcon, Flavie Tortereau

Using machine learning to predict feed intakes of meat sheep from animal traits and ruminal microbiota

Q. Le Graverand[1*], C. Marie-Etancelin[1], J.L. Weisbecker[1], A. Meynadier[1], D. Marcon[2], F. Tortereau[1]

[1] GenPhySE, Université de Toulouse, INRAE, ENVT, 24 Chemin de Borde-Rouge - Auzeville CS 52627, F-31326 Castanet-Tolosan, France; [2] INRAE, Experimental Unit P3R, Domaine de la Sapinière, F-18390 Osmoy, France; [*]quentin.le-graverand@inrae.fr

Abstract

Animal traits, such as body weights, and rumen microbiota composition have been proposed as feed intake predictors. The present study assessed what are the best predictors out of animal traits, metabarcoding data or a combination of both. Predictions were carried out with sparse Partial Least Squares Regression (sPLSR), Support Vector Regression (SVR) and Random Forest Regression (RFR). With all three approaches, best feed intake predictions were obtained with animal traits only. The generalizability of models to animals of an independent year was assessed: negative (<-0.1) to high (>0.8) correlations between actual and predicted feed intakes were obtained. Finally, estimated breeding values (EBVs) were computed for actual and predicted feed intakes. These EBVs were highly correlated (>0.9) depending on the prediction approach. It mainly varied with proportions of true and predicted feed intakes used during the genetic evaluation.

Introduction

To assess feed efficiency of sheep, expensive feed intake recording is necessary. Thus, most meat sheep breeding companies cannot afford to select for feed efficiency. Predicting feed intake or feed efficiency could be one solution. As reviewed by Pittroff and Kothmann (2001), many sheep feed intake prediction models include live weight. Omics data were also used as predictors, such as 16S data to predict sheep Residual Feed Intake (RFI) with a general linear model (Ellison et al., 2019), or rabbit feed intakes with mixed models and SPLSR (Velasco-Galilea et al., 2021).

We have chosen to focus on the prediction of feed intake since it intervenes in the definition of many complex phenotypes such as feed efficiency or the environmental footprint. Moreover, to our knowledge, there is no comparison between accuracies of feed intake predictions from animal traits or microbiota compositions with a large number of sheep and machine learning.

The first goal of our study is to check whether microbial information improves feed intake predictions with different machine learning approaches. To do so, we compare accuracies of different machine learning approaches (sPLSR, SVR and RFR) applied to several sets of predictors (body weight, body composition traits, and/or ruminal microbiota composition). Then, the second objective is to see if a machine learning model could be generalized to new animals raised a different year. Finally, the last objective is to assess if a genetic selection with machine learning predictions is feasible.

Materials & Methods

Animal husbandry and traits

Overall 277 Romane male lambs were reared at the INRAE Experimental Unit P3R between 2018 and 2020. Animals originated from two experimental divergent lines selected for a decreased or increased RFI. Lambs are part of the second generation of selection in 2018 (n=103) and the third generation in 2019 (n=101) or 2020 (n=73). The divergence represented 1.9 genetic standard deviations during the third generation. After weaning, lambs were

accustomed to an *ad libitum* diet with low-energy concentrates. When animals were approximately 3 months old, feed intakes were recorded during six weeks. Average Daily Feed Intake (ADFI) was computed as the daily feed intake mean over six weeks. At the beginning and end of the trial period, body weights were recorded, then average daily weight gain and growth capacity (estimated body weight at 145 days old) were calculated. Back ultrasound measurements enabled to estimate the back-fat thickness and muscle depth at the trial end.

Ruminal microbiota composition

At the end of the six-weeks trial, ruminal fluid was collected with a medical gastric tube and DNA was extracted from samples. The V4-V5 region of 16S rRNA gene was amplified and then sequenced at the Genomic and Transcriptomic Platform (INRAE, Toulouse, France). Sequences were processed, cleaned and clustered into Operational Taxonomic Units (OTUs) with FROGS tools and pipeline (Escudié et al., 2018). OTUs were pre-filtered according to their relative abundance: only OTUs representing at least 0.005% of all sequences were kept. OTUs zero counts were imputed with Bayes-Laplace multiplicative replacement (Martín-Fernández et al., 2015) before applying the Centered LogRatio (CLR) transformation. CLR coordinates were filtered to retain OTUs with less than 90% imputed values each year. Finally, the 496 retained OTUs coordinates were adjusted for the age, the year, the pen and sequencing plate effects with a robust MM linear regression (Maechler et al., 2017).

Machine learning

Three different approaches of machine learning were carried out to predict ADFI: sparse Partial Least Squares (sPLSR, mixOmics R package), support vector (SVR, e1071 R package) and random forest (RFR, randomForest R package) regressions. Different sets of prediction features were considered: animal traits (body weights, growth traits and body composition), or adjusted 16S CLR coordinates. Those feature sets were used separately or were combined together by concatenation for ADFI predictions. Successively, animals of one year were used as a testing set. To avoid a bias, 2018 will not be used as a testing set because 2020 lambs are the progeny of 2018 individuals. Animals of the two remaining years were used as training/validation sets: 10-fold cross-validations repeated 50 times were used to tune the hyperparameters and maximize the coefficient of determination (R^2) between predicted and actual ADFI phenotypes. Models were then fitted without the testing dataset. Next, ADFI of testing set animals were predicted with each predictor set and machine learning model. Pearson correlations between predicted and actual ADFI phenotypes were computed. Within one testing set, significance of differences between machine learning approaches and features sets were tested with Dunn and Clark's z test (Dunn and Clark, 1969). P-values were adjusted with Bonferroni's procedure.

Breeding values estimation

Breeding values were estimated with PEST (Groeneveld et al., 1990), considering a feed intake heritability of 0.28 (Tortereau et al., 2020). Two sets of populations were used to estimate breeding values: an entire Romane population named E (born from 2009 to 2020), with 6,419 animals in the pedigree including 1,900 with ADFI records; one subset population named S (2018 to 2020), with 4,102 animals in the pedigree including 277 with records. The model included the fixed effects of year, pen, early life traits (litter size, suckling method), sex and body weight as a covariate. EBVs of actual ADFI and EBVs of predicted phenotypes are estimated. EBVs were only estimated for phenotypic ADFI predicted with accurate strategies ($R^2 > 0.7$).

Results

Comparison of different predictors and machine learning approaches for phenotypic ADFI

Table 1 details accuracies of ADFI predictions for the three machine learning models carried out with 16S data, animal traits and a concatenation of both. Considering correlations between actual and predicted ADFI phenotypes, there was no significant difference between sPLSR, SVR and RFR accuracies whatever the testing set or the predictors.
With all testing sets and machine learning techniques, correlations were significantly lower when only 16S data was used as the feature set. Whatever the testing set and the machine learning model, combining animal traits and 16S data together as predictors did not significantly increased correlations.

Table 1. Pearson correlations between predicted and actual ADFI phenotypes of testing set animals.

		Testing sets	
Models	Features	2019	2020
sPLSR	16S	-0.116[a]	0.191[a]
	Animal	0.763[b]	0.810[b]
	Animal+16S	0.726[b]	0.818[b]
SVR	16S	-0.040[a]	0.351[a]
	Animal	0.766[b]	0.802[b]
	Animal+16S	0.763[b]	0.817[b]
RFR	16S	0.031[a]	0.312[a]
	Animal	0.777[b]	0.773[b]
	Animal+16S	0.709[b]	0.737[b]

[a,b] Within one testing set, correlations with no common letter significantly differ (adjusted P<0.05, Dunn and Clark's z test).

Relationship between EBVs of predicted and actual ADFI

The quality of predicted ADFI EBVs is presented in Table 2. Within one machine learning approach and regardless of the testing set, correlations between EBVs are significantly higher when a full set of records (E) is used during the genetic evaluation compared to a partial set (S).

Table 2. Pearson correlations between predicted and actual ADFI estimated breeding values of testing set animals.

		Testing sets and populations[1]			
		2019		2020	
Models	Features	S	E	S	E
sPLSR	Animal	0.681[a]	0.868[c]	0.843[a]	0.954[b]
	Animal+16S	0.698[ab]	0.876[cd]	0.852[a]	0.956[b]
SVR	Animal	0.624[b]	0.814[d]	0.848[a]	0.962[b]
	Animal+16S	0.650[ab]	0.817[cd]	0.839[a]	0.956[b]
RFR	Animal	0.731[ab]	0.891[cd]	0.863[a]	0.952[b]
	Animal+16S	0.698[ab]	0.880[cd]	0.848[a]	0.959[b]

[1] Populations differ in the number of records used for EBVs estimation. S: subset population; E: entire population.
[a,b] Within one testing set (2019 or 2020), correlations with no common letter significantly differ (adjusted P<0.05, Dunn and Clark's z test).

In 2020, no significant difference could be found between machine learning approaches. However, when only animal traits were used as 2019 ADFI predictors, sPLSR performed significantly better than SVR, with RFR having intermediate performances. Finally, the

combination of animal traits and 16S data as predictors did not significantly improve correlations between actual and predicted ADFI EBVs.

Discussion

Rumen microbiota composition alone lead to poorly accurate ADFI predictions with sPLSR, SVR and RFR, which was also observed with sPLSR carried out with rabbit caecal microbiota compositions (Velasco-Galilea et al., 2018). Predictions made with animal traits reached higher performances, whatever the machine learning model. The integration of animal traits and microbiota compositions did not improve predictions. Therefore, our results do not advocate for the use of metabarcoding data to predict feed intake. However, the microbiota composition could be more relevant when one is trying to predict feed efficiency. Ellison et al. (2019) managed to predict RFI of 20 animals with 16S data and achieved a correlation of 0.71 between actual and predicted feed efficiency. In the future, another comparison will be done with RFI to assess the potential of 16S and 18S data predictors and determine the relevance of archaeal, bacterial, fungal and protozoal compositions.

Depending on testing sets and predictors, no consistent difference was found between the three modelling approaches when it comes to predict ADFI and then estimate its EBVs. Only SVR performances seemed to be less stable across testing sets when predicted ADFI EBVs were compared to actual ADFI breeding values.

The moderate or high correlations achieved between EBVs of predicted and actual ADFI suggests that predictions could be used for genetic selection of feed intake. Further research in the phenotyping strategy should be carried out, especially about the ratio between predicted and recorded phenotypes. In the present study only two testing sets were used to assess performances of machine learning predictions. Those testing sets were constituted as independent cohorts, *i.e* with animals raised a different year than the training set. When predictors are sensitive to environmental factors, such as microbiota compositions, the presence of contemporaneous animals in both training and testing sets should improve predictions.

References

Dunn O. J., and Clark V. (1969). J. Am. Stat. Assoc. 64(325):366-377.

Ellison M. J., Conant G. C., Lamberson W. R., Austin K. J., Van Kirk E., et al. (2019). J. Anim. Sci. 97(7) :2878–2888. https://doi.org/10.1093/jas/skz170

Escudié F., Auer L., Bernard M., Mariadassou M., Cauquil L., et al. (2018). Bioinformatics 34(8):1287–1294. https://doi.org/10.1093/bioinformatics/btx791t

Groeneveld E., Kovač M., and Wang T. (1990). In: World Congress on Genetics Applied to Livestock Production. Edinburg, UK, pp. 488–491.

Maechler M., Rousseeuw P., Croux C., Todorov V., Ruckstuhl A., et al. (2017). Robustbase: basic robust statistics R package.

Martín-Fernández J. A., Hron K., Templ M., Filzmoser P., and Palarea-Albaladejo J. (2015). Stat. Modelling. 15(2):134–158. https://doi.org/10.1177/1471082X14535524

Pittroff W., and Kothmann M. M. (2001). Livest. Prod. Sci. 71(2-3):131–150. https://doi.org/10.1016/S0301-6226(01)00218-4

Tortereau F., Marie-Etancelin C., Weisbecker J. L., Marcon D., et al. (2020). Animal 14(4):681-687. https://doi.org/10.1017/S1751731119002544

Velasco-Galilea M., Piles M., Ramayo-Caldas Y., and Sánchez J. P. (2021). Sci. Rep. 11(1):1–18. https://doi.org/10.1038/s41598-021-99028-y

This work received funding from the French funding agency (ANR) (GrassToGas project - ERA-GAS n° 39413) and from the European Union's Horizon 2020 research and innovation program under the Grant Agreement No. 772787 (SMARTER).