S07(T)-PP-05

Large scale dataset to improve and validate the prediction of lactoferrin content using milk mid-infrared spectrometry

Hélène Soyeurt[1], Frédéric Dehareng[2], Sinead McParland[3], Marion Calmels[4], Nicolas Gengler[1], Clément Grelet[2]

[1]TERRA research and teaching centre, Gembloux Agro-Bio Tech / University of Liège, Gembloux, Belgium

[2]Valorisation of agricultural products, Walloon Research Centre, Gembloux, Belgium

[3]Animal & Grassland, Teasgasc, Moorepark, Ireland

[4]Research and development, Seenovia, Saint-Berthevin, France

Lactoferrin (LF), a glycoprotein with interesting immunological properties, is measured in milk using ELISA kit which limits its use for management and breeding purposes. Past investigations highlighted the feasibility of predicting an indicator of LF using milk mid-infrared (MIR) spectra. Nearly 7,000 samples were analyzed to allow a deeper insight into the potential accuracy of such a trait. LF was measured using two different ELISA kits on 2,654 and 3,965 milk samples. From the first dataset, the best partial least squares regression (N= 2,587) gave a 10 fold cross-validation (cv) R^2 equal to 0.61 ± 0.03 with a cross-validation root mean square error (RMSEcv) of 160 ± 7.8 mg/L. When this equation was applied on the second dataset, R^2 was equal to 0.20 with RMSE of 164 mg/L (N=3,965). The best equation obtained using the second dataset (N=3,689) provided R^2cv of 0.47 ± 0.03 with RMSEcv of 80 ± 2.6 mg/L. Validation R^2 based on the first dataset (N=2,602) was

0.46 with RMSEv of 198 mg/L. RMSE and R^2 provided by the 2 equations were clearly different. However, in both cases, the highest contents of LF were badly predicted using both equations and the smallest contents of LF gave negative predictions. When the data were combined together, the best equation used 6,059 samples and gave R^2cv of 0.63 ± 0.04 and RMSEcv of

99.42 ± 2.82 mg/L. Those statistics are representative of the dataset used which was unbalanced with mainly LF contents lower than 400 mg/L. So, 3 different calibration sets were performed using more balanced LF contents based on 14 classes of 100 mg/L of LF. The number of low LF records influenced the statistics. Lower was the number of records used for the smallest contents of LF, higher was the ability of the equation to predict the highest contents of LF. Finally, the best chosen equation was the one using 100 samples for the 5 first classes of 100 mg of LF per L of milk, all samples in the following classes were kept. Based on those 3 simulations, the regression used 989 samples with a LF content ranged from 6 to 1318 mg/L. Average calibration and cv R^2 were of 0.72 and 0.68. Calibration and cv RMSE were of 81 and 169 mg/L. A validation performed on the remaining samples (N=5,293) gave a R^2 of 0.41 with a RMSE of 157 mg/L. Even if highest LF contents were better predicted using balanced datasets, RMSE stayed largely higher than the ones observed for the smaller contents. A hypothesis explaining those observations could be that the milk spectra can be different for the samples having effectively LF content higher than 600 mg/L. Therefore the use of LF class dependent models could be a potential interesting improvement. A second conclusion is based on the differences observed for the two used datasets. RMSE was largely different suggesting a reference method error requiring the definition of standard procedure for LF measurement.

Keywords: lactoferrin, milk, infrared