Short communication: Selecting the most informative mid-infrared spectra wavenumbers to improve the accuracy of prediction models for detailed milk protein content

G. Niero,¹ M. Penasa, P. Gottardo, M. Cassandro, and M. De Marchi

Department of Agronomy, Food, Natural resources, Animals and Environment, University of Padova, Viale dell'Università 16, 35020 Legnaro (PD), Italy

ABSTRACT

The objective of this study was to investigate the ability of mid-infrared spectroscopy (MIRS) to predict protein fraction contents of bovine milk samples by applying uninformative variable elimination (UVE) procedure to select the most informative wavenumber variables before partial least squares (PLS) analysis. Reference values ($n = 114$) of protein fractions were measured using reversed-phase HPLC and spectra were acquired through MilkoScan FT6000 (Foss Electric A/S, Hillerød, Denmark). Prediction models were built using the full data set and tested with a leave-one-out cross-validation. Compared with MIRS models developed using standard PLS, the UVE procedure reduced the number of wavenumber variables to be analyzed through PLS regression and improved the accuracy of prediction by 6.0 to 66.7%. Good predictions were obtained for total protein, total casein (CN), and α -CN, which included α_{S1} - and α_{S2} -CN; moderately accurate predictions were observed for κ -CN and total whey protein; and unsatisfactory results were obtained for β -CN, α -lactalbumin, and β -lactoglobulin. Results indicated that UVE combined with PLS is a valid approach to enhance the accuracy of MIRS prediction models for milk protein fractions.

Key words: Fourier transform infrared spectroscopy, milk protein fraction, uninformative variable elimination, partial least squares regression

Short Communication

In recent decades, milk proteins have attracted the attention of scientific community for several reasons. From a nutritional point of view, milk is considered one of the most suitable sources of protein in the human diet, and both casein and whey protein fractions are

classified as high-quality compounds in relation to human AA requirements, digestibility, and bioavailability (Pereira, 2014). It has been widely demonstrated that milk proteins have positive effect on consumer health due to their biological activity. In particular, milk proteins and several peptides derived from their enzymatic hydrolysis and metabolic processing have shown antibacterial, antiviral, antifungal, and antioxidant activity, both in vivo and in vitro (Mills et al., 2011; Niero et al., 2015). Moreover, adequate milk protein, calcium, and vitamin D intake results in decreased bone remodeling, better calcium retention, and decreased age-related bone loss and fracture risk (Caroli et al., 2011). Finally, protein composition of milk is an important factor for the profitability of the dairy industry, playing a central role in cheese production (Comin et al., 2008; Pretto et al., 2013). Several studies have demonstrated that increased κ -CN content in milk is associated with shorter rennet coagulation time, a firmer curd, and greater cheese yield (Wedholm et al., 2006; Jõudu et al., 2008).

High-performance liquid chromatography is one of the most used techniques to identify and quantify milk proteins, mainly because of its sensitivity, repeatability, and reproducibility; nevertheless this analytical method is time consuming, quite expensive, and it requires skilled operators. For these reasons, the development of cheaper and faster methods for determining protein composition is needed. Mid-infrared spectroscopy (MIRS) is a rapid and cost-effective tool for recording phenotypes at the population level (De Marchi et al., 2014). Results from the scientific literature have indicated that the accuracy of prediction of milk protein composition based on MIRS spectra is generally moderate to low (Bonfatti et al., 2011; Rutten et al., 2011); hence, MIRS-predicted protein composition cannot be used in milk payment systems, but only for selection purposes (Rutten et al., 2011). Recently, Gottardo et al. (2015) obtained a significant improvement of the accuracy of prediction models for milk titratable acidity and calcium content by combining partial least squares (PLS) approach with uninformative variable

Received August 27, 2015.

Accepted November 29, 2015.

¹Corresponding author: giovanni.niero@studenti.unipd.it

elimination (**UVE**) procedure. Therefore, the present study investigated the effect of UVE procedure on the accuracy of MIRS models to predict contents of protein fractions of individual bovine milk samples.

Individual milks of Holstein-Friesian ($n = 63$), Brown Swiss ($n = 26$), and Jersey ($n = 25$) cows from parity 1 to 7 and from 7 to 408 DIM were collected in 4 herds between February and March 2015 during the morning milking. Milks were immediately added with preservative (bronopol, 2-bromo-2-nitropropan-1,3-diol), transferred at 4°C to the laboratory of the South Tirol Dairy Association (Bolzano, Italy), and analyzed for milk chemical composition using a MilkoScan FT6000 (Foss Electric A/S, Hillerød, Denmark). An aliquot of each sample was transferred to the laboratory of the Department of Agronomy, Food, Natural resources, Animals and Environment of the University of Padova (Legnaro, Italy) for milk protein analysis.

Milk samples were added with an aqueous solution of guanidine HCl (6 *M* guanidine-HCl, 0.1 *M* BisTris Buffer, 19.5 *mM* dithiothreitol, 5.37 *mM* sodium citrate) in a proportion of 1:1 (vol/vol) and incubated at room temperature for 1 h to promote proteins solubilization. Samples were then centrifuged for 7 min at room temperature at $16,000 \times g$ to promote the separation of fat. An aliquot of soluble fraction was added with a solution containing 4.5 *M* guanidine diluted in a solvent consisting of water, acetonitrile, and trifluoroacetic acid (94.9:5.0:0.1; vol/vol/vol) in a proportion of 1:4 (vol/vol). Samples were finally filtered with a 0.45- μm filter.

Analysis of protein fractions was carried out using an HPLC station (Agilent 1260 Series; Agilent Technologies, Santa Clara, CA), equipped with a reversed-phase column C8 (Aeris Widepore XBC8, Phenomenex, Torrance, CA; 3.6 μm , 300, 250 $\times$ 2.1 μm i.d.) kept at 70°C, and with a diode array Detector (Agilent 1260 Series, DAD VL+, G1315C), following the method proposed by Maurmayr et al. (2013). This method allowed the detection and quantification of α -CN, handled as sum of α_{S1} - and α_{S2} -CN chromatographic peaks, β -CN, κ -CN, α -LA, and β -LG. Identification and quantification of different milk protein fractions were carried out using internal and external standards and calibration curves of α -CN, β -CN, κ -CN, α -LA, and β -LG (Sigma-Aldrich, St. Louis, MO). In addition, the following traits were derived from protein fractions: total casein (**TCN**), which was calculated as the sum of α -CN, β -CN, and κ -CN; total whey protein (**TWP**), which was the sum of α -LA and β -LG; and total protein (**TP**), which was the sum of TCN and TWP (i.e., of all protein fractions quantified with HPLC).

The MIRS spectra, recorded in 2 batches on all 114 milk samples using MilkoScan FT6000 (Foss Electric A/S), were retrieved from the laboratory of the South

Tirol Dairy Association. After the transformation of spectral data to absorbance using the $\log_{10}$ of the reciprocal of the transmittance, the presence of outliers was checked using the robust Mahalanobis distance procedure. Following this procedure, no outliers were detected. Uninformative variable elimination procedure followed by PLS (Centner et al., 1996; Gottardo et al., 2015) was used to develop calibration models for milk protein fraction contents. Partial least squares regression was performed using the ChemometricsWithR package (Wehrens, 2011), and UVE procedure using a homemade script in R software (R Core Team, 2015). Calibration models were carried out using spectra after discarding wavenumbers related to water noise absorption (1,601 to 1,717, and 3,052 to 5,011 cm^{-1} ; Hewavitharana and van Brakel, 1997). The removal of regions related to water reduced the wavenumbers from 1,060 to 865. Models were built using the full data set and tested with a leave-one-out cross-validation (114 segments). Goodness of fit was measured through the coefficient of determination of leave-one-out cross-validation ($1 - \text{VR}$) and root mean square error in cross-validation. The optimal number of principal components was defined as the minimum number of latent factors to achieve the lowest root mean square error of calibration. The ratio performance deviation (**RPD**) was calculated by dividing the standard deviation of reversed-phase HPLC reference values to the root mean square error in cross-validation, and it was used to test the practical utility of the prediction models (Williams, 1987).

Table 1 shows the descriptive statistics of milk composition predicted by MilkoScan FT6000 (Foss Electric A/S) and casein and whey protein fraction contents measured by reversed-phase HPLC. Means of fat and protein contents were 3.85 and 3.44% (wt/wt), respectively. These values are slightly lower than those reported by Penasa et al. (2014), who compared 3 cattle breeds in mixed dairy herds for predicted milk coagulation properties and composition traits.

Casein and whey protein fraction contents exhibited good degree of variability, with a coefficient of variation between 18.6 (α -CN) and 27.1% (κ -CN), and equal to 15.8 (α -LA) and 48.2% (β -LG), respectively (Table 1); this facilitated the development of prediction models. Total protein (36.79 mg/mL) calculated as the sum of all protein fractions was higher than protein content obtained in the laboratory of the South Tirol Dairy Association (3.44%). The difference can be mainly related to the fact that TP was measured by reverse-phase HPLC on skim milk, whereas protein content was predicted by MIRS on raw milk. Moreover, TP was measured in weight per volume, whereas protein content was expressed as weight per weight; this could

contribute to further explain the difference between TP and protein content.

Concerning casein fractions, the most abundant was α -CN, which averaged 16.15 mg/mL, followed by β -CN and κ -CN, which averaged 10.15 and 6.60 mg/mL, respectively (Table 1). The content of α -CN was very similar and that of β -CN was lower than findings of De Marchi et al. (2009) and Bonfatti et al. (2011), who investigated the effectiveness of MIRS for the prediction of protein composition of Simmental cow milk. The concentration of κ -CN was much greater than values reported in the literature (De Marchi et al., 2009; Bonfatti et al., 2011), probably because of the inclusion of milk from Jersey breed, which is known to have high levels of κ -CN (Auld et al., 2004). Finally, β -LG and α -LA averaged 3.07 and 0.82 mg/mL, respectively (Table 1), which were somewhat lower than results of De Marchi et al. (2009) and Bonfatti et al. (2011). For these reasons, and because the HPLC method used as a reference did not allow the detection of some minor whey proteins (e.g., BSA and immunoglobulins), the casein index of the present study (89.4%) was higher than expected. However, it is worth noting that this value is very close to those (87 to 88%) reported by other authors who investigated the ability of MIRS to predict protein fractions based on reference methods (HPLC, capillary zone electrophoresis) different from Kieldahl (De Marchi et al., 2009; Bonfatti et al., 2011; Rutten et al., 2011).

Fitting statistics of MIRS prediction models for TP and protein fraction contents developed using PLS regression before UVE procedure are shown in Table 2. Models were developed using spectra information (865 wavenumbers) after discarding regions related to water absorption. The most accurate prediction models were those for α -CN, TP, TCN, and κ -CN that reached $1 - VR$ values of 0.83, 0.78, 0.77, and 0.66,

respectively. According to Williams (2003), values of $1 - VR$ between 0.66 and 0.81 underline moderately accurate predictions and values between 0.82 and 0.91 indicate good predictions. The poorest predictions were observed for TWP ($1 - VR = 0.53$), β -CN ($1 - VR = 0.36$), and α -LA and β -LG ($1 - VR = 0.31$). The very low $1 - VR$ for α -LA and β -LG might be partly due to their lower content in milk compared with casein fractions. Regarding RPD, the greatest values were obtained for α -CN (2.43), TP (2.13), and TCN (2.09). According to Karoui et al. (2006), RPD values greater than 2 indicate that predictions are good enough to be considered for analytical purposes.

In general, the prediction of protein fractions by MIRS is quite difficult because spectra information do not allow detecting and discriminating the length of AA chains and secondary and tertiary structure of different protein fractions (Bonfatti et al., 2011). Compared with the present study, Bonfatti et al. (2011) reported the same $1 - VR$ values for TP, TCN, and α -LA, greater $1 - VR$ values for TWP, β -CN, and β -LG, and lower $1 - VR$ values for α -CN and κ -CN. Rutten et al. (2011) obtained moderate to low predictive ability of milk protein composition; however, they expressed protein fractions on a protein and not a milk basis. Overall, comparison among studies is quite difficult because of differences in expression of protein fractions, reference analysis, MIRS equipment, and statistical approaches and software used to develop prediction models. For example, Rutten et al. (2011) used capillary zone electrophoresis as a reference method that is known to be less accurate respect to HPLC in the determination of protein composition (Recio et al., 1997; De Marchi et al., 2014).

The UVE procedure selected 110 (β -CN) to 390 (TCN) informative wavenumbers for protein fraction contents and TP from a total of 865 (Table 2). The

Table 1. Descriptive statistics of milk quality traits determined by mid-infrared spectroscopy and protein fraction contents measured by reversed-phase HPLC

Trait ¹	Mean	SD	CV, %	Minimum	Maximum
Milk quality, %					
Fat	3.85	0.98	25.4	1.53	6.22
Protein	3.44	0.43	12.5	2.48	4.55
Casein fraction, mg/mL					
α -CN	16.15	3.00	18.6	10.10	24.13
β -CN	10.15	2.45	24.1	5.17	16.47
κ -CN	6.60	1.79	27.1	3.00	11.54
TCN, mg/mL	32.90	6.41	19.5	20.91	49.11
Whey protein fraction, mg/mL					
α -LA	0.82	0.13	15.8	0.42	1.18
β -LG	3.07	1.48	48.2	0.71	8.90
TWP, mg/mL	3.89	1.50	38.0	1.54	9.88
TP, mg/mL	36.79	6.59	18.0	23.33	53.77

¹TCN = total casein (the sum of α -CN, β -CN, and κ -CN); TWP = total whey protein (the sum of α -LA and β -LG); TP = total protein (the sum of TCN and TWP).

Table 2. Fitting statistics¹ of mid-infrared prediction models developed using partial least squares (PLS) regression before and after uninformative variable elimination (UVE) procedure applied to milk protein fraction contents and total protein (mg/mL)

Trait ²	#PC	PLS before UVE				PLS after UVE			
		N ₁	RMSE _{CV}	1 – VR	RPD	N ₂	RMSE _{CV}	1 – VR	RPD
Casein fraction									
α-CN	13	865	1.23	0.83	2.43	315	1.05	0.88	2.86
β-CN	13	865	1.95	0.36	1.25	110	1.53	0.60	1.60
κ-CN	13	865	1.05	0.66	1.71	120	0.88	0.74	2.03
TCN	10	865	3.05	0.77	2.09	390	2.51	0.88	2.55
Whey protein fraction									
α-LA	9	865	1.11	0.31	1.20	167	0.10	0.37	1.30
β-LG	15	865	1.21	0.31	1.12	190	1.10	0.47	1.34
TWP	20	865	1.06	0.53	1.41	150	0.88	0.69	1.70
TP	10	865	2.62	0.78	2.13	350	2.10	0.88	3.15

¹#PC = number of principal components; N₁ = number of wavenumbers before UVE procedure; N₂ = number of wavenumbers after UVE procedure; RMSE_{CV} = root mean square error in cross-validation; 1 – VR = coefficient of determination in cross-validation; RPD = ratio performance deviation, calculated by dividing the standard deviation of reversed-phase HPLC reference values to the RMSE_{CV}.

²TCN = total casein (the sum of α-CN, β-CN, and κ-CN); TWP = total whey protein (the sum of α-LA and β-LG); TP = total protein (the sum of TCN and TWP).

elimination of uninformative variables before PLS analysis improved the accuracy of predictions models for all traits. In particular, the percentage increment of 1 – VR ranged from 6.0 (α-CN) to 66.7% (β-CN), as depicted in Figure 1. Overall, the most visible improvements were obtained for protein fractions that exhibited the lowest 1 – VR values before UVE procedure. Gottardo et al. (2015) applied the UVE procedure to select the most informative wavenumbers to build prediction models for milk calcium content and titratable acidity, and they reported a significant enhancement of the accuracy of prediction for these traits. As depicted in Figure 2, low degrees of scatter were observed for TP, TCN, and α-CN, intermediate scatter level was observed for κ-CN, TWP, and β-CN, and high degrees of scatter were found for α-LA and β-LG. The use of UVE procedure before PLS analysis led to an improvement of RPD values for all traits (Table 2). In particular,

RPD of κ-CN increased from 1.71 to 2.03, suggesting that predictions of this casein fraction could be considered for analytical purposes (Karoui et al., 2006).

In conclusion, the results of the present study indicated that UVE combined with PLS is a valid approach to improve the accuracy of MIRS prediction models compared with standard PLS. As the procedure allowed for a large reduction of variables to be processed through PLS, an improvement of the efficiency of MIRS models to predict milk protein fraction contents on a large scale is expected.

ACKNOWLEDGMENTS

The research was funded by the “Burbacco” project (PSR 2007 – 2013, Mis. 124 – Aiuto N° 2308615). The authors thank the milk laboratory of the South Tirol Dairy Association (Bolzano, Italy) for providing spec-

Figure 1. Percentage increment (Δ%) of coefficient of determination in cross-validation (1 – VR) of prediction models for milk protein fractions and total protein after uninformative variable elimination (UVE) procedure compared with 1 – VR of prediction models before UVE.

Figure 2. Scatter plots of predicted milk protein fraction contents and total protein (mg/mL), after uninformative variable elimination procedure (y-axis) versus measured (x-axis) milk protein fraction contents and total protein.

tra data used in this study, and Giulio Visentin and Sofia Ton (University of Padova) for technical support. The generosity of the four farms who participated in the trial is also gratefully acknowledged. G. Niero collected milk samples and performed laboratory analyses. G. Niero and M. Penasa wrote the first draft of the paper. P. Gottardo performed statistical analyses. M. De Marchi and M. Cassandro designed the research. All authors contributed to the interpretation and discus-

sion of the results, and approved the final version of the manuscript.

REFERENCES

- Auldust, M. J., K. A. Johnston, N. J. White, W. P. Fitzsimons, and M. J. Boland. 2004. A comparison of the composition, coagulation characteristics and cheesemaking capacity of milk from Friesian and Jersey dairy cows. *J. Dairy Res.* 71:51–57.
- Bonfatti, V., G. Di Martino, and P. Carnier. 2011. Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein

- composition and contents of protein genetic variants of individual milk of Simmental cows. *J. Dairy Sci.* 94:5776–5785.
- Caroli, A., A. Poli, D. Ricotta, G. Banfi, and D. Cocchi. 2011. Invited review: Dairy intake and bone health: A viewpoint from the state of the art. *J. Dairy Sci.* 94:5249–5262.
- Centner, V., D.-L. Massart, O. E. de Noord, S. de Jong, B. M. Vandeginste, and C. Sterna. 1996. Elimination of uninformative variables for multivariate calibration. *Anal. Chem.* 68:3851–3858.
- Comin, A., M. Cassandro, S. Chessa, M. Ojala, R. Dal Zotto, M. De Marchi, P. Carnier, L. Gallo, G. Pagnacco, and G. Bittante. 2008. Effects of composite β - and κ -casein genotypes on milk coagulation, quality, and yield traits in Italian Holstein cows. *J. Dairy Sci.* 91:4022–4027.
- De Marchi, M., V. Bonfatti, A. Cecchinato, G. Di Martino, and P. Carnier. 2009. Prediction of protein composition of individual cow milk using mid-infrared spectroscopy. *Ital. J. Anim. Sci.* 8(Suppl. 2):399–401.
- De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa. 2014. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. *J. Dairy Sci.* 97:1171–1186.
- Gottardo, P., M. De Marchi, M. Cassandro, and M. Penasa. 2015. Technical note: Improving the accuracy of mid-infrared prediction models by selecting the most informative wavenumbers. *J. Dairy Sci.* 98:4168–4173.
- Hewavitharana, A. K., and B. van Brakel. 1997. Fourier transform infrared spectrometric method for the rapid determination of casein in raw milk. *Analyst (Lond.)* 122:701–704.
- Jōudu, I., M. Henno, T. Kaart, T. Püssa, and O. Kärt. 2008. The effect of milk protein contents on the rennet coagulation properties of milk from individual dairy cows. *Int. Dairy J.* 18:964–967.
- Karoui, R., A. M. Mouazen, É. Dufour, L. Pillonel, D. Picque, J. De Baerdemaeker, and J.-O. Bosset. 2006. Application of the MIR for the determination of some chemical parameters in European Emmental cheeses produced during summer. *Eur. Food Res. Technol.* 222:165–170.
- Maurmayr, A., A. Cecchinato, L. Grigoletto, and G. Bittante. 2013. Detection and quantification of α_{S1} -, α_{S2} -, β -, κ -casein, α -lactalbumin, β -lactoglobulin and lactoferrin in bovine milk by reverse-phase high-performance liquid chromatography. *Agric. Conspec. Sci.* 78:201–205.
- Mills, S., R. P. Ross, C. Hill, G. F. Fitzgerald, and C. Stanton. 2011. Milk intelligence: Mining milk for bioactive substances associated with human health. *Int. Dairy J.* 21:377–401.
- Niero, G., M. De Marchi, A. Masi, M. Penasa, and M. Cassandro. 2015. Short communication: Characterization of soluble thiols in bovine milk. *J. Dairy Sci.* 98:6014–6017.
- Penasa, M., F. Tiezzi, A. Sturaro, M. Cassandro, and M. De Marchi. 2014. A comparison of the predicted coagulation characteristics and composition of milk from multi-breed herds of Holstein-Friesian, Brown Swiss and Simmental cows. *Int. Dairy J.* 35:6–10.
- Pereira, P. C. 2014. Milk nutritional composition and its role in human health. *Nutrition* 30:619–627.
- Pretto, D., M. De Marchi, M. Penasa, and M. Cassandro. 2013. Effect of milk composition and coagulation traits on Grana Padano cheese yield under field conditions. *J. Dairy Res.* 80:1–5.
- R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Recio, I., L. Amigo, and R. López-Fandiño. 1997. Assessment of the quality of dairy products by capillary electrophoresis of milk proteins. *J. Chromatogr. B Biomed. Sci. Appl.* 697:231–242.
- Rutten, M. J. M., H. Bovenhuis, J. M. L. Heck, and J. A. M. van Arendonk. 2011. Predicting bovine milk protein composition based on Fourier transform infrared spectra. *J. Dairy Sci.* 94:5683–5690.
- Wedholm, A., L. B. Larsen, H. Lindmark-Månsson, A. H. Karlsson, and A. Andrén. 2006. Effect of protein composition on the cheese-making properties of milk from individual dairy cows. *J. Dairy Sci.* 89:3296–3305.
- Wehrens, R. 2011. *Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences*. Springer, Heidelberg, Germany.
- Williams, P. C. 1987. Implementation of near-infrared technology. Pages 143–167 in *Near-Infrared Technology in the Agricultural and Food Industries*. 2nd ed. Williams, P. C., Norris, K., ed. American Association of Cereal Chemists, St. Paul, MN.
- Williams, P. C. 2003. Near-infrared technology getting the best out of light. Page 109 in *A Short Course in the Practical Implementation of Near Infrared Spectroscopy for the user*. 1.1. ed. PDK Projects Inc., Nanaimo, BC, Canada.