J. Dairy Sci. 98:6583–6587
http://dx.doi.org/10.3168/jds.2015-9730

Short communication: Predictive ability of Fourier-transform mid-infrared spectroscopy to assess *CSN* genotypes and detailed protein composition of buffalo milk

V. Bonfatti,*[1] A. Cecchinato,† and P. Carnier*
*Department of Comparative Biomedicine and Food Science, and
†Department of Agronomy, Food, Natural resources, Animals and Environment (DAFNAE) University of Padova, 35020 Legnaro, Padova, Italy

ABSTRACT

The aim of this work was to test the applicability of Fourier-transform mid-infrared spectroscopy (FT-MIR) for the prediction of the contents of casein (CN) and whey protein fractions in buffalo milk. Buffalo milk samples spectra were collected using a MilkoScan FT2 (Foss, Hillerød, Denmark) over the spectral range from 5,000 to 900 wavenumber × cm^{-1}. Contents of protein fractions, as well as *CSN1S1* and *CSN3* genotypes, were assessed by reversed phase HPLC. The highest coefficients of determination in cross-validation (1 − VR) were obtained for the contents (g/L of milk) of total protein and CN (1 − VR = 0.92), followed by the content of β-CN, total whey protein, and α_{S2}-CN (1 − VR of 0.87, 0.77, and 0.63, respectively). Conversely, contents of α_{S1}-CN, γ-CN, glycosylated-κ-CN, total κ-CN, and whey protein fractions were predicted with poor accuracy (1 − VR <0.51). When protein fractions were expressed as percentages to total protein, 1 − VR values were never greater than 0.61 (β-CN). Only 56 and 70% of the observations were correctly classified by discriminant analysis in each of 2 groups of *CSN1S1* and *CSN3* genotypes, respectively. Results showed that FT-MIR spectroscopy is not applicable when prediction of detailed milk protein composition with high accuracy is required. Predictions may play a role as indicator traits in selective breeding, if the genetic correlation between FT-MIR predictions and measures of milk protein composition are high enough and predictions of protein fraction contents are sufficiently independent from the predicted total protein content.

Key words: buffalo milk, protein composition, casein fractions, spectroscopy

Short Communication

Detailed protein composition of buffalo (*Bubalus bubalis*) milk is influenced by nongenetic effects due to parity, DIM, and milk yield (Bonfatti et al., 2012a) and by genetic effects due to CN genotypes (Bonfatti et al., 2012c). In addition, protein composition is related to buffalo milk technological properties (Bonfatti et al., 2013a). Hence, assessment of the detailed milk protein composition might be relevant to evaluate the technological quality of buffalo milk.

Fourier-transform mid-infrared (**FT-MIR**) spectroscopy has been used in many studies to predict compositional traits of cow milk including protein composition (Bonfatti et al., 2011; Rutten et al., 2011). Conversely, composition of buffalo milk has scarcely been investigated, even though buffalo milk accounts for nearly 13% of worldwide milk yield (FAOSTAT, 2012). The aim of our study was to investigate the ability of FT-MIR spectroscopy to predict the detailed milk protein composition of individual buffalo milk.

A total of 174 buffaloes were sampled once in 5 herds located in northern Italy from January to May 2013. Individual milk samples were collected during the morning milking. Samples were stored at 4°C until acquisition of FT-MIR spectra. Individual milk sample spectra were collected within maximum 2 h of sampling using a MilkoScan FT2 (Foss, Hillerød, Denmark) over the spectral range from 5,000 to 900 wavenumber × cm^{-1}. Transmittances (T) were converted to absorbances (A) as $A = \log_{10}(1/T)$. Two spectral acquisitions were carried out for each sample and results were averaged before data analysis. An aliquot of milk was frozen and stored at −40°C until reversed phase (**RP**)-HPLC analysis was performed.

Contents of α_{S1}-CN, α_{S2}-CN, β-CN, γ-CN, glycosylated κ-CN, unglycosylated κ-CN, α-LA, and β-LG were assessed by RP-HPLC using the method developed by Bonfatti et al. (2013b). Total CN content (**TCN**; g/L) was computed as the sum of α_{S1}-CN, α_{S2}-CN, β-CN, γ-CN, and total κ-CN (the sum of glycosylated and

Received April 21, 2015.
Accepted June 2, 2015.
[1]Corresponding author: valentina.bonfatti@unipd.it

Table 1. Descriptive statistics for traits and model fitting parameters for predictions of milk protein composition

			Parameter[2]				
Item[1]	Mean	SD	Obs	Terms	Math	SE_{CV}	1 − VR
Protein, g/L	51.94	6.76	167	9	A 1,15,5,1	1.84	0.92
Casein, g/L	46.13	6.26	168	9	A 1,15,5,1	1.70	0.92
Whey protein, g/L	5.81	1.24	166	10	B 1,5,5,1	0.56	0.77
Casein number, %	88.77	2.17	165	10	B 1,5,5,1	1.16	0.51
Protein fractions, g/L							
α_{S1}-CN	14.13	4.16	173	6	B 0,0,1,1	2.91	0.51
α_{S2}-CN	7.03	1.29	167	9	B 1,5,5,1	0.72	0.63
β-CN	15.49	2.69	165	11	B 1,5,5,1	0.94	0.87
γ-CN	0.58	0.28	166	2	A 1,15,5,1	0.22	0.14
κ-CN	8.44	3.25	173	8	B 1,5,5,1	2.77	0.27
Glyco-κ-CN	4.38	2.90	172	8	B 1,5,5,1	2.50	0.26
α-LA	3.31	0.89	165	6	B 1,5,5,1	0.56	0.46
β-LG	2.50	1.10	166	8	B 1,5,5,1	0.67	0.51
Protein composition, %							
α_{S1}-CN	27.90	6.29	173	5	B 0,0,1,1	5.67	0.30
α_{S2}-CN	13.59	2.15	172	7	B 1,5,5,1	1.58	0.56
β-CN	29.82	3.38	166	10	B 1,5,5,1	1.79	0.68
γ-CN	1.12	0.55	164	3	B 1,5,5,1	0.47	0.12
κ-CN	16.33	6.25	173	5	B 0,0,1,1	5.82	0.25
Glyco-κ-CN	8.50	5.75	172	8	B 0,0,1,1	5.24	0.28
α-LA	6.42	1.68	168	6	B 1,5,5,1	1.35	0.43
β-LG	4.81	1.94	165	10	B 1,5,5,1	1.16	0.51

[1]Protein: α_{S1}-CN + α_{S2}-CN + β-CN + γ-CN + κ-CN + β-LG + α-LA; casein = α_{S1}-CN + α_{S2}-CN + βγ-CN + κ-CN; whey protein = β-LG + α-LA; casein number = (casein/protein) × 100. Glyco = glycosylated. Protein composition is expressed as contents of protein fractions to total protein content.

[2]Obs = number of observation in the calibration set after outlier elimination; terms = number of modified partial least square regression latent variables; math = mathematical treatments of the spectral data (A: only spectral regions from 5,011 to 3,673 cm^{-1} and from 3,048 to 930 cm^{-1} were used; B: only the spectral region from 3,048 to 930 cm^{-1} was used) where the first number is the order of the derivative, the second number is the segment length in data points over which the derivative was taken, the third and fourth numbers are the segment length for first and second smoothing, respectively; SE_{CV} = standard error of cross-validation; 1 − VR = coefficient of determination of cross-validation.

unglycosylated κ-CN). Total whey protein content was calculated as the sum of α-LA and β-LG content. Total protein content (**PRT**) was the sum of TCN and total whey protein content. Genotypes at *CSN1S1* and *CSN3* were derived by the same RP-HPLC method. Genotypes at *CSN1S1* corresponded to the C > T transition at nucleotide 578 of *Bubalus bubalis CSN1S1* (complete coding sequence EMBL no. AJ005430.1), resulting in the AA substitution Leu^{178}(A) → Ser^{178}(B) of the mature α_{S1}-CN polypeptide chain (from exon 3 to 17 of the reference sequence O62823). Genotypes at *CSN3* corresponded to the T > C transition at nucleotide 467 of the complete coding sequence HQ677596 results in the amino acid substitution Ile^{135}(X1) → Thr^{135}(X2) of the mature κ-CN polypeptide chain (Bonfatti et al., 2012b).

The FT-MIR calibration models were developed using modified partial least square regression (**MPLS**; Shenk and Westerhaus, 1991) procedures as implemented in the software WinISI II (InfraSoft International, State College, PA). Several mathematical treatments of raw spectra were compared before regression analysis. Samples exhibiting large spectral distance (i.e., global Mahalanobis distance >3) from the population centroid as well as samples for which the difference between the reference and the predicted value was much larger than the standard error of cross-validation were considered outliers and discarded from the calibration analysis. The number of samples used for calibration after outlier elimination is reported in Table 1.

Prediction models were validated using a 4-fold random cross-validation. The standard error of cross validation and the coefficient of determination of cross-validation (**1 − VR**) were calculated and used to evaluate the predictive ability of calibration models.

To investigate the ability of FT-MIR spectroscopy to predict *CSN1S1* and *CSN3* individual genotypes, discriminant analysis was also performed. The prediction model was developed by MPLS as implemented in the software WinISI II (InfraSoft International). Due to their low number, *CSN1S1* AA animals (n = 12) were grouped with AB animals and *CSN3* X2X2 animals (n = 15) were grouped with X1X2 animals. Then, in the MPLS discriminant analysis, spectra from 2 different groups of genotypes (BB vs. AA and AB, for *CSN1S1*; X1X1 vs. X1X2 and X2X2, for *CSN3*) were used. The software set up a calibration matrix with binary variables (0 and 1) against the different genotype groups.

The calibration was then conducted by regressing the wavelength data on the groups defined as 0 or 1. A 4-fold cross validation was used to test the accuracy of the models.

Descriptive and model-fitting statistics for predictions of contents of major milk protein fractions are reported in Table 1. High PRT can be ascribed to skimming of milk before chromatographic analysis. Average protein composition of buffalo milk was consistent with findings of Addeo (1979), D'Ambrosio et al. (2008), and Bonfatti et al. (2012a).

The most accurate prediction models for protein contents (g/L of milk) were those for PRT and TCN (1 − VR = 0.92), followed by those for β-CN, total whey protein, and α_{S2}-CN (1 − VR of 0.87, 0.77, and 0.63, respectively). Total κ-CN, glycosylated-κ-CN, α_{S1}-CN, γ-CN, and whey protein fractions were predicted with poor accuracy (1 − VR <0.51). The relatively low values of 1 − VR might be due to the fact that RP-HPLC analysis of protein fraction contents was measured on skim milk samples whereas spectra acquisition was performed on raw milk.

In general, prediction accuracy observed in our study was in agreement with results obtained by Bonfatti et al. (2011) for cow milk, except for κ-CN and glycosylated-κ-CN predictions, which exhibited poor accuracy; our results were also similar to those of other studies performed using cow and ewe milk (Ferrand et al., 2012). Variation in content of these fractions is markedly affected by *CSN3* genotypes (Bonfatti et al., 2012c), and frequency distributions of such traits are bimodal (Figure 1). This can partly explain why contents of κ-CN and glycosylated-κ-CN were poorly modeled by linear regressions. Likewise, Bonfatti et al. (2011) reported that the multimodal frequency distribution of β-CN and κ-CN content in cow milk caused by *CSN2* and *CSN3* genotype effects was responsible for the low prediction accuracy of these protein fractions.

When calibrations focused on protein composition (i.e., relative contents of protein fractions measured as the percentages of total protein), all 1 − VR estimates were lower than 0.7. Percentages of protein fractions in total protein were also calculated indirectly from predictions of PRT and contents of all protein fractions. Performances of calibration models for such predictions were consistent with those obtained when protein fraction percentages were predicted directly (data not reported in tables).

Prediction of *CSN1S1* and *CSN3* genotypes, based on infrared information, might be of interest because a relationship between milk protein genotypes and variation in coagulation properties has been previously described (Bonfatti et al., 2012b). Results of discriminant analysis for *CSN1S1* are reported in Table 2. A second derivative applied to raw spectra with segment length of 5 data points, a 5 data point- and a 2 data point-segment as first and second smoothing, respectively, in the spectral regions from 5,011 to 3,673 cm^{-1} and from 3,048 to 930 cm^{-1}, and using 7 MPLS terms yielded the best results. For *CSN1S1*, 87 samples had AA or AB genotype and 87 samples had BB genotype. The fraction of samples correctly classified for genotype at *CSN1S1* was only 56%.

Result of the discriminant analysis for genotype classification at *CSN3* is reported in Table 3. Data consisted of 82 and 92 samples having *CSN3* genotype X1X1 or X1X2 and X2X2, respectively. The best results were obtained when the first derivative, with a segment length of 5 data points, a 5 data point- and

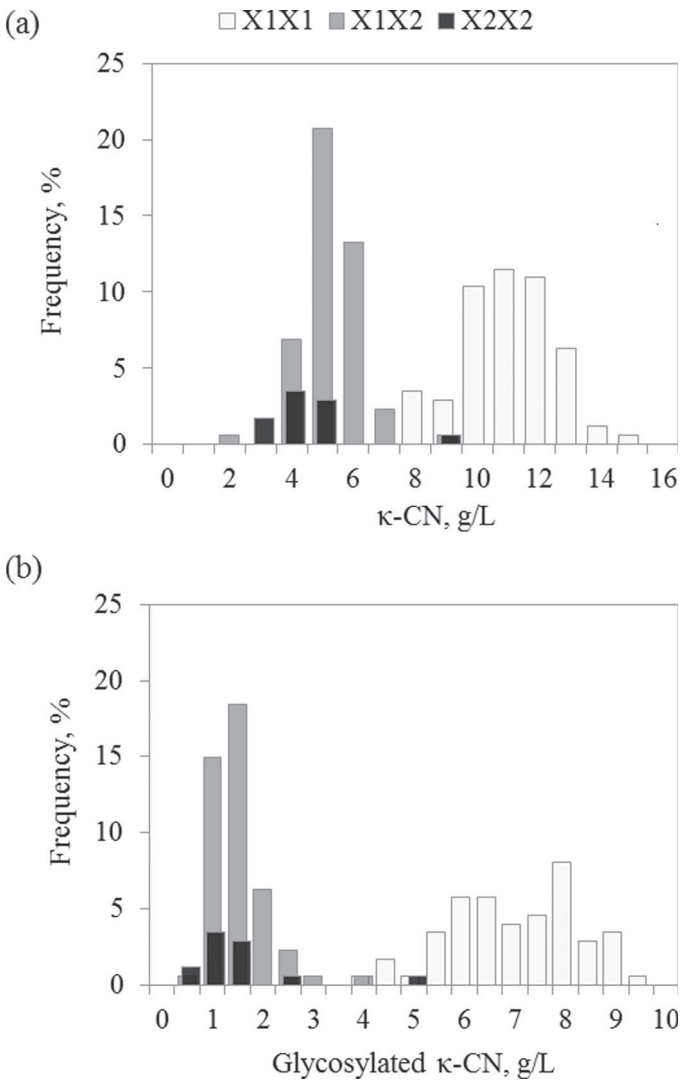

Figure 1. Frequency distribution across *CSN3* genotypes (X1X1, X1X2, and X2X2) for the content of κ-CN (a) and glycosylated κ-CN (b) in milk.

Table 2. Number of observations classified by discriminant analysis in each of two groups of *CSN1S1* genotypes of buffalo milk

	Classified genotype	
True genotype	AA or AB	BB
AA or AB	50	37
BB	39	48
Missclassification, %	44.83	37.93

Table 3. Number of observations classified by discriminant analysis in each of 2 groups of *CSN3* genotypes of buffalo milk

	Classified genotype	
True genotype	X1X1	X1X2 or X2X2
X1X1	56	26
X1X2 or X2X2	26	66
Missclassified, %	31.71	28.26

a 1 data point-segment as first and second smoothing, respectively, was applied to the spectra region from 3,048 to 930 cm^{-1} using 6 MPLS terms. The percentage of samples correctly classified for genotype at *CSN3* was 70%. The ability of the models to classify the genotypes had worsened when the less-frequent genotypes *CSN1S1* AA and *CSN3* X2X2 were excluded from the analysis.

In our study, FT-MIR spectroscopy was not able to discriminate between *CSN* genotypes with sufficient accuracy. The low predictive ability of the models might be due the low number of observations. However, Berget et al. (2010), although analyzing only 45 samples, were able to classify *CSN1S1* goats genotypes through FT-MIR procedures with good accuracy (nearly 87% of samples were correctly classified). Variation in α_{S1}-CN content of goat milk and, as a consequence, variation in total protein content, is markedly affected by genotypes at *CSN1S1* because null alleles, responsible for reduced null α_{S1}-CN content, are present in goat populations. In buffalo, animals carrying different genotypes at milk protein genes show small differences in protein composition and total protein content when compared with variation observed in goats. Likely, this is a major cause for low discriminating capacity in genotype classification exhibited by FT-MIR relative to that detected by Berget et al. (2010).

Results of our study indicate that FT-MIR is not able to provide accurate predictions of detailed milk protein composition. Such predictions may, however, be of practical interest as indicator traits for selective breeding applications. The practical utility of FT-MIR predictions of protein composition for selective breeding depends upon the heritability of the predicted traits and upon the magnitude of the genetic correlation between the predicted and the measured values (Cecchinato et al., 2009; Rutten et al., 2011). When such correlation and the genetic variance of predictions are large enough, even predictions exhibiting moderate accuracy may be of practical value. In addition, sire proofs would be obtained from predictions based on progeny spectra and on multiple spectral records per offspring over lactation, which would beneficially affect the accuracy of sire evaluations for protein composition.

For milk FA, many authors (Soyeurt et al., 2006, 2011; Eskildsen et al., 2014) reported high prediction accuracies for contents relative to percentage concentrations in fat. Eskildsen et al. (2014) stated that such accuracies arise from the correlation between FA contents and total fat, which is accurately FT-MIR predicted, rather than on specific absorption bands associated with individual FA. Predictions of protein fractions likely also rely on the correlation between contents of single fractions and total protein. Prediction of the contents of protein fractions may not be useful in milk recording systems or breeding programs if they are exclusively related to the prediction of total protein and do not provide any additional information. Moreover, if indirect correlations are used to build a calibration equation, the model will not be valid for future samples unless the indirect correlations are conserved for these samples (Eskildsen et al., 2014). Thus, the prediction of protein fraction contents may not be valid for populations having a covariance structure different from that of the calibration set. Further research is needed to elucidate these aspects.

REFERENCES

Addeo, F. 1979. The composition of whole water buffalo casein. Annali della Facoltà di Scienze Agrarie dell'Università di Napoli in Portici. Serie IV, 13:149–159.

Berget, I., H. Martens, A. Kohler, S. K. Sjurseth, N. K. Afseth, B. Narum, T. Ådnøy, and S. Lien. 2010. Caprine *CSN1S1* haplotype effect on gene expression and milk composition measured by Fourier transform infrared spectroscopy. J. Dairy Sci. 93:4340–4350.

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.

Bonfatti, V., M. Gervaso, A. Coletta, and P. Carnier. 2012a. Effect of parity, days in milk, and milk yield on detailed milk protein composition in Mediterranean water buffalo. J. Dairy Sci. 95:4223–4229.

Bonfatti, V., M. Gervaso, R. Rostellato, A. Coletta, and P. Carnier. 2013a. Protein composition affects variation in coagulation properties of buffalo milk. J. Dairy Sci. 96:4182–4190.

Bonfatti, V., M. Giantin, M. Gervaso, A. Coletta, M. Dacasto, and P. Carnier. 2012b. Effect of *CSN1S1–CSN3* (α_{S1}-κ-casein) composite genotype on milk production traits and milk coagulation properties in Mediterranean water buffalo. J. Dairy Sci. 95:3435–3443.

Bonfatti, V., M. Giantin, M. Gervaso, R. Rostellato, A. Coletta, M. Dacasto, and P. Carnier. 2012c. Short communication: *CSN1S1–CSN3* (α_{S1}-κ-casein) composite genotypes affect detailed milk pro-

tein composition of Mediterranean water buffalo. J. Dairy Sci. 95:6801–6805.

Bonfatti, V., M. Giantin, R. Rostellato, M. Dacasto, and P. Carnier. 2013b. Separation and quantification of water buffalo milk protein fractions and genetic variants by RP-HPLC. Food Chem. 136:364–367.

Cecchinato, A., M. De Marchi, L. Gallo, G. Bittante, and P. Carnier. 2009. Mid-infrared spectroscopy predictions as indicator traits in breeding programs for enhanced coagulation properties of milk. J. Dairy Sci. 92:5304–5313.

D'Ambrosio, C., S. Arena, A. Salzano, G. Tenzone, L. Ledda, and A. Scaloni. 2008. A proteomic characterization of water buffalo milk fractions describing PTM of major species and the identification of minor components involved in nutrient delivery and defense against pathogens. Proteomics 8:3657–3666.

Eskildsen, C. E., M. A. Rasmussen, S. B. Engelsen, L. B. Larsen, N. A. Poulsen, and T. Skov. 2014. Quantification of individual fatty acids in bovine milk by infrared spectroscopy and chemometrics: Understanding predictions of highly collinear reference variables. J. Dairy Sci. 97:7490–7951.

FAOSTAT (Food and Agriculture Organization of the United Nations). 2012. Agriculture statistics. Accessed Nov. 6, 2014. http://faostat.fao.org/site/569/DesktopDefault.aspx?PageID=569#ancor.

Ferrand, M., G. Miranda, H. Larroque, O. Leray, S. Guisnel, F. Lahalle, M. Brochard, and P. Martin. 2012. Determination of protein composition in milk by mid-infrared spectrometry. Pages 1–5 in Proc. 38th Int. Comm. Anim. Rec. Ann. Meet. Cork. ICAR, Rome, Italy.

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.

Shenk, J. S., and M. O. Westerhaus. 1991. Population definition, sample selection and calibration procedures for near infrared reflectance spectroscopy. Crop Sci. 31:469–474.

Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. J. Dairy Sci. 89:3690–3695.

Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. P. Berry, M. Coffey, and P. Dardenne. 2011. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. J. Dairy Sci. 94:1657–1667.