Determination of protein composition in milk by mid-infrared spectrometry

M. Ferrand[1], G. Miranda[2], H. Larroque[3], O. Leray[4], S. Guisnel[1], F. Lahalle[1,5], M. Brochard[1], P. Martin[2]

[1]Institut de l'Elevage, 149 rue de Bercy, 75595 Paris cedex 12, France
[2]INRA, UMR1313, Animal Genetics and Integrative Biology, Domaine de Vilvert 78352 Jouy-en-Josas cedex, France
[3]INRA, UR0631, Station d'Amélioration Génétique des Animaux, F-31326 Castanet-Tolosan cedex, France
[4]ACTILAIT 39, BP 70129, 39802 Poligny Cedex, France
[5]CNIEL, 42 rue de Châteaudun, 75314 Paris cedex 09, France

Abstract

The fine milk composition is currently studied with attention, some components being of a particular interest. To get a better description of these components, it is necessary to have rapid and reliable methods to analyze several hundreds of milk samples in order to estimate genetic and environmental effects on milk quality.

An innovative method based on liquid chromatography coupled with mass spectrometry was developed to identify and quantify the main milk proteins and some works are now on-going to determine milk protein composition from MIR (Mid Infra-Red) spectra usually obtained by milk analysing laboratories.

The first results show that it is possible to estimate each of the 4 casein contents with a good accuracy, especially in ewe's and cow's milk. By contrast, it seems more difficult to estimate the whey proteins contents. Some improvements are under evaluation to enhance the performance of these equations.

This study which is part of the PhenoFinlait programme was funded by Apis-Gène, CASDAR, CNIEL, FranceAgriMer, France Génétique Elevage, French Ministry of Agriculture and French Ministry of Research (National Agency for Research).

Keywords: milk proteins, liquid chromatography, mass spectrometry, mid infrared spectroscopy

Introduction

Milk is a complex product which contains a lot of components with different properties. The main milk proteins are divided into two important families: caseins (α_{s1}, α_{s2}, β, κ) accounting for 80% of milk proteins, and whey proteins (including α-lactalbumin and β-lactoglobulin) which represent ca. 20%, in cattle. Some of these proteins are particularly important for cheese production, for example a milk with a high content in κ-casein will clot rapidly and provide a firm curd (Grosclaude, 1988).

A part of the PhénoFinLait programme consists in developing two methods to analyse the protein composition of milk: the first one is a qualitative and quantitative determination using liquid chromatography coupled with mass spectrometry (LC/MS), the second one aims to estimate some major milk proteins contents from MIR spectral data. The LC/MS reference method allows quantification and characterization of the six main proteins, but this method is

time consuming and more expensive than the MIR spectrometry that is used in routine in control laboratories. Previous studies (De Marchi, 2009; Rutten, 2011; Bonfatti, 2011) show that it was possible to estimate milk protein composition using this technology, but the errors remained high, in particular as far as κ-casein and the two main whey proteins are concerned.

Materials and methods

MIR spectra database

A MIR spectra database has been created between 2009 and 2011 with 73 149 cow milks from 1 072 herds, 33 623 ewe milks from 167 herds and 54 932 goat milks from 191 herds distributed in different areas of France. Spectra were recorded on Foss FT6000 or FT+ and Bentley spectrometers.

First, studies on proteins were focused on Foss data. Some spectra areas were removed following the constructor recommendations (Foss, 1998) since these areas are sensitive to water molecule. Finally, we kept wavelengths from 965 to 1544 cm^{-1}, from 1 716 to 2 272 cm^{-1} and from 2 434 to 2 970 cm^{-1}

Quantitative and qualitative analysis using LC/MS

Concomitantly, milks from 271 cows of different breeds (Montbéliarde, Holstein and Normande) 157 ewes (Lacaune and Manech-Tête-Rousse), 151 goats (Saanen and Alpine) were selected and analyzed with an innovative methodology based on Liquid Chromatography (Ultimate 3000 HPLC from Dionex) coupled with Mass Spectrometry (MicroToF focus from Bruker). This method (Miranda, Bianchi and Martin, manuscript in preparation) allows the identification and a relative quantification of the 6 main milk proteins:κ-casein (glycosylated or not),α_{s1}, α_{s2} and β-caseins, as well as β-lactoglobulin and α-lactalbumin. The quantification is based on the integration of the UV signal recorded at 214 nm of each peak of the chromatogram. Surfaces are expressed as percent of the total of peaks of the chromatogram. The identification is achieved by comparing the observed masses of each protein to the predicted ones referenced in a milk protein masses database designed and implemented by the authors:

(APP: IDDN.FR.001.460019.000.R.C.2011.000.10300).

This database contains ca. 3 000 mass values corresponding to these 6 main milk proteins, including genetic variants, splicing isoforms, post-translational modifications (phosphorylation, glycosylation) and the main degradation products (due to the action of plasmin, i.e. essentiallyγ-caseins and related proteose-peptones).

In order to establish reliable equations, only samples with a proteolysis rate lower than 20% were retained, i.e. 193 samples in cow milks, 152 samples in ewe milks, and 147 samples for goat milks. For each protein, outliers were removed by Grubb's test as indicated in the norm ISO 5725-2.

Calibration equations

For cow milk, the samples were divided into calibration and validation sets (n calibration=135 and n validation=58). For ewe and goat milk, a cross-validation was used, the sample number being a little low.

The equations were developed by univariate PLS regression (Tenenhaus, 2002), data being centered but not reduced according to Bertrand *et al*. (2006). For each equation, optimal number of latent variables was chosen according to root mean square error of cross-validation (RMSEPcv). To improve equations and quality of estimation, a selection of wavelengths by genetic algorithm was performed before PLS regression in cow and ewe milk (Ferrand, 2010). The genetic algorithm (GA) used is the algorithm developed by Leardi (1998) which is specific to wavelength selection. Mutation rate, initial population, and number of variables selected in the solution of initial population were fixed to 1%, 30 and 5 respectively.

GAs were performed with MATLAB 7.8 and PLS regressions were performed with the package PLS in R 2.8.1.

To compare and assess the equations, several statistical parameters were computed: mean, standard deviation (SD), standard error of validation ($SE_{validation}$), validation coefficient of determination ($R^2_{validation}$) and the relative error[$SE_{validation}$/Mean (%)].

$SE_{validation}$ is defined as

$$\sqrt{\frac{\left(\sum_{i=1}^{N}(\hat{y}_i - y_i)^2\right)}{N-k-1}}$$

with N being the number of samples and k the number of latent variables introduced in PLS regression.

We considered that estimation was accurate enough and robust to be applied in routine, when the relative error was under 8%. For relative error in the range of 8 to 12%, we advised using these equations with caution. We chose to use this parameter rather than the R^2validation because this latter depends on the standard deviation of our population.

Results

The equation performances for cow milk and ewe milk are presented in table 1. The best results are obtained for ewe milk, where the relative error is lower than 5% for total casein and β-casein, and inferior to 10% for the other three individual caseins (κ, α_{s1}, α_{s2}). For whey proteins, we can estimate the total, but not in the detail, equation for ?-lactalbumin presenting a very low R^2.

In cow milk, we have higher relative errors and lower R^2. Total casein, β-casein, α_{s1}-casein are correctly estimated.

These results are comparable to Rutten *et al*. (2011), except for ?-lactoglobulin for which a higher error was obtained. Differences could be due to the reference method used that was different between the two studies (LC-MS versus capillary zone electrophoresis). Depending on the method used, quantitative values could be more or less accurate.

Table 1. Fitting statistics of prediction models (g/100ml), independent validation dataset for cow milk and cross-validation for ewe milk (PLS regression only or genetic algorithm (GA) + PLS regression).

	Cow milk					Ewe milk				
	N[1]	Mean	Std	Relative error (%)	R^2	N	Mean	Std	Relative error (%)	R^2
TP	58	3.341	0.307	0.98	0.99	147	5.080	0.772	0.38	1.00
Caseins	57	2.458	0.271	3.93	0.88	149	4.089	0.666	2.73	0.97
κ-CN	57	0.317	0.054	11.61	0.54	145	0.441	0.069	6.57	0.82
α_{S2}-CN	58	0.237	0.041	10.29	0.65	147	0.557	0.102	7.46	0.83
α_{S1}-CN	57	0.860	0.100	6.32	0.71	143	1.216	0.210	5.06	0.91
β-CN	57	1.037	0.130	5.91	0.78	147	1.834	0.315	4.14	0.94
Whey proteins	57	0.385	0.058	9.96	0.58	145	0.564	0.097	8.89	0.73
α-LA	57	0.123	0.018	10.91	0.48	143	0.152	0.029	16.45	0.26
β-LG	58	0.263	0.054	15.29	0.45	145	0.410	0.090	10.47	0.77

[1]*Number of samples used in the validation after removing outliers.*

Table 2. Estimation of proteins composition of ewe milks from the PhenofinLait database, Descriptive statistics (PLS regression only or genetic algorithm (GA) + PLS regression).

	Mean	SD
TP	5.456	0.768
Caseins	4.399	0.68
κ-CN	0.465	0.07
α_{S2}-CN	0.592	0.131
α_{S1}-CN	1.375	0.217
β-CN	1.908	0.307
Whey proteins	0.64	0.089
α-LA	0.158	0.017
β-LG	0.472	0.085

By contrast, the accuracy is lower with goat milk. It may be due to the lower amount of total protein in goat milk and to its polymorphism at the α_{s1}-casein locus which is responsible for large quantitative variations in milk protein content as it is for fat content and its fatty acid composition (Mahé *et al.*, 1994).

To validate these first equations, we have applied the ewe equations on the 127 040 ewe milk spectra of the PhenoFinLait database. Distributions are normal and means are coherent with the bibliography (Table 2).These results confirm the equation abilities to estimate the protein contents on average, even if some verifications are still necessary.

Conclusions

These first results show it makes it possible to obtain accurate estimations for each casein in individual milk samples of ewe and cow.

In goat milk, equations are clearly insufficiently accurate, probably due to the strong polymorphism existing at the α_{s1}-casein locus in that species, which may impact the accuracy of the results obtained with the reference method. Some works are still necessary to improve predictive equations for goats.

Further researches will focus on the improvement of the reference method, while increasing simultaneously the initial sampling size to get more accurate estimation of milk protein profile.

Advancements in the PhenoFinLait program are available on *www.phenofinlait.fr*.

List of References

Bertrand, D., Dufour, E., 2006. La spectrométrie infrarouge et ses applications analytiques. Second ed. Tec&Doc Lavoisier, Paris, 660pp.

Bonfatti, V., 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 Science. 94: 5776-5785.

De Marchi M. *et al.*, 2009. Prediction of protein composition of individual cow milk using mid-infrared spectroscopy. Ital.J.Anim.Sci 8(2) : 399-401.

Ferrand, M., Huquet, B., Barbey, S., Barillet, F., Faucon, F., Larroque, H., Leray, O., Trommenschlager, J.M. and Brochard, M., 2011. Determination of fatty acid profile in cow's milk using mid-infrared spectrometry: Interest of applying a variable selection by genetic algorithms before a PLS regression. Chemometr. Intell. Lab. Syst. 106: 183-189.

Foss, 1998. Reference Manual of Milkoscan FT120 (Type 71200). Denmark.

Grosclaude, F., 1988. Le polymorphisme génétique des principales lactoprotéines bovines. Relations avec la quantité, la composition et les aptitudes fromagères du lait. INRA Prod. Anim. 1(1): 5-17.

Grubbs, F., 1969. Procedures for Detecting Outlying Observations in Samples. Technometrics. 11: 1-21.

Mahé, M.F., Manfredi, E., Ricordeau, G., Piacère, A. and Grosclaude, F., 1994. Effets du polymorphisme de la caséine α_{s1} caprine sur les performances laitières: analyse intradescendance de boucs de race Alpin. Genet. Sel. Evol. 26: 151-157.

Rutten, M.J.M. *et al.*, 2011. Prediction of β-lactoglobulin genotypes based on milk Fourier transform infrared spectra. Journal of Dairy Science, 94: 4183-88.

Rutten, M.J.M. *et al.*, 2011. Predicting bovine milk protein composition based on Fourier transform infrared spectra. Journal of Dairy Science, 94: 5683-5690.

Tenenhaus, M., 2002. La regression PLS, Technip, Lassay-les-Chateaux.