Chimiometrie 2009

Abstract/Résumé

Determination of fatty acid profile in cow milk using Mid-Infrared spectrometry: interest of applying a variable selection before a PLS regression

FERRAND M. (1), HUQUET B. (1), BARBEY S. (2), BARRILLET F. (3), BROCHARD M. (1), LARROQUE H. (3), LERAY O. (4)

(1) Institut de l'Elevage, 149 rue de Bercy, 75595 Paris cedex 12 – marion.ferrand@inst-elevage.asso.fr

(2) UE INRA du Pin-Au-Haras, (3) INRA-SAGA, (4) ACTILAIT

Keywords: mid-infrared (MIR) spectrometry, milk, fatty acid, Partial Least Squares (PLS) regression, genetic algorithms

1 Introduction

Milk contains many components such as proteins, fatty acids, lactose, minerals, and a lot of other molecules at low concentration. As food, these components can be more or less beneficial or bad in regard of human health and at the present time social demand is increasing in terms of healthy products.

Environmental (feeding and breeding techniques) and genetic factors (species, breed and animal genotype) influence the fine composition of milk. To evaluate their impact, it becomes crucial to measure elementary milk components accurately and at low cost.

The first aim of PhenoFinLait program is to develop reliable, cheap and easy-to-use methods for individual proteins and fatty acids content measurement and to apply these methods to identify the genetic and environmental factors, affecting these contents.

In 2008, the project's goal was to develop equations to predict fatty acid content with Mid Infra-Red (MIR) spectra from French experimental farm cow milk samples. Univariate and multivariate Partial Least Square (PLS1,PLS2) regression applied on MIR spectra give good predictions for main fatty acids. The small rate of some fatty acids of interest make their predictions more difficult. To improve equations and quality of prediction, we decided to apply a selection of variables before PLS regression as suggested by several authors[1,2]. We retained the use of genetic algorithms, - so-called Genetic Algorithm method or GA method -, as an efficient method that we applied to our data sets.

2 Theory

Genetic algorithms method is based on evolutionary biology. A population of candidate solutions evolves using genetic operators like reproduction, mutation and selection. A solution is a vector where each variable is coded with 0 (not-selected) or 1 (selected). Initial population has 30 candidate solutions. The evolution is controlled by a fitness function. In this study, the fitness function is the cross-validated explained variance of PLS regression applied on selected variables. To breed a new generation (two new solutions), two candidate solutions are selected. During this step of reproduction, crossing-over or mutation (probability of 1% for each variable) can happen. The solutions created integrate the population if they appears better than others solutions. The population is constant, so the worst solutions are discarded when new solutions integrate the population. This process is repeated until the fixed number of generations is reached. For more details, it is possible to consult the article from Leardi [1].

3 Material and methods

Appropriate samples from cow milk for calibration purposes, with a large variability in their composition, were collected on Pin-Au-Haras INRA experimental farm. Milk samples from 154 crossbred Holstein X Normande dairy cows were analyzed using MIR spectrometry with defined routine FT-MIR analyzers (Milkoscan FT6000, Foss and Bentley FTS) and using gas chromatography according to ISO standards [3] which is the benchmark method. Quantities of 64 fatty acids were expressed in g/100mL. Spectra have been recorded from 5012 to 926 cm-1. According Foss [4], only informative wavelength bands, i.e. bands not spoiled by water molecule, were kept (a total of 446 wavelengths). No pre-treatments were applied as suggested by Soyeurt et al. [5].

Calibration equations were developed by univariate and multivariate PLS regression according to Bertrand D. et al. [6], data being centered but not reduced. For each equation, optimal number of latent variables was chosen according to root mean square error of cross-validation (RMSEP_cv). In order to improve prediction equations, a selection of variables by genetic algorithm (GA) [1] took place before PLS regression.

We used the algorithm developed by Leardi [1] after checking its robustness by varying the parameters like population size or mutation probability. We retained the same parameters than Leardi [1] even if the algorithm seems very stable and we performed algorithm on autoscaled data. Following variables selection, PLS regression were applied as explained before.

To compare and to assess the equations, several statistical parameters were computed: mean, standard deviation (Sd), standard error of cross-validation (SECV), and cross-validation coefficient of determination (R^2CV). We consider a prediction is precise enough and robust to be applying in routine, when R^2CV is upper than 0.80. For R^2CV from 0.70 to 0.80, we suggest using these equations with caution. The accuracy is checked according to SECV criteria. To evaluate the performance of genetic algorithm, we compare the SECV of the model issued of PLS regression on variables selected by GA with the SECV of the multivariate PLS regression without variables selection.

Genetic algorithms were performed with MATLAB 7.8 and PLS regression with R 2.8.1

4 Results and discussion

Genetic algorithms select in average 46 variables out of 446 in the form of wavelength bands. The 2272-1905 cm^{-1} band is rarely selected, while the 2970-2278 cm-1 band is selected for most fatty acids. Some fatty acids groups like saturated fatty acids (FA) or C18:0 families have specific wavelengths.

Using PLS regression, we have good predictions for 16 fatty acids (R2CV>80%) and correct quality of predictions for 14 FA (70<R2CV<80%). Using genetic algorithms + PLS regression, we have good quality of predictions for 19 fatty acids (R2CV) >80%) and correct quality of predictions for 14 FA (70<R2CV<80%). These results are promising, especially regarding the non-worsening of overall predictions.

The most significant improvement is the overall accuracy gain of 8% as measured in the standard error of cross validation. For instance, accuracy increases of 8% for linoleic acid (C18:2 9c12C) and of 15% for linolenic acid (C18:3 n-3) (Table 1). These fatty acids are of a wide interest regarding nutrition and thus a gain of accuracy, even small, is very important.

The main disadvantage of genetic algorithms is the high computing time required (3 hours by fatty acid) which must be counterbalanced by a significant accuracy improvement that fits to purpose, where such an algorithm choice is made. In this study, we think that the gain is worthwhile because the developed equations are the base of big genomic selection programs, and they could be extended to other livestock problematic like payment of milk.

Table 1. Statistical parameters for each calibration equation (PLS regression only or genetic algorithm + PLS regression)

			PLS2			GA + PLS				
	Mean	**Sd**	**Latent var.**	**SECV**	**R2CV**	**Var. GA**	**Latent var.**	**SECV**	**R2CV**	**Improvement %**
Fat content	3,978	0,568	13	0,022	1,00	65	9	0,016	1,00	27,91
C4:0	0,152	0,023	22	0,009	0,87	63	7	0,009	0,87	3,09
C6:0	0,092	0,015	6	0,004	0,94	50	9	0,003	0,97	25,51
C8:0	0,055	0,011	20	0,002	0,96	42	9	0,002	0,97	11,63
C10:0	0,128	0,034	20	0,009	0,93	61	14	0,008	0,95	16,58
C12:0	0,146	0,046	20	0,013	0,93	69	9	0,011	0,95	15,99
C14:0	0,458	0,094	9	0,043	0,81	31	10	0,040	0,83	6,83
C15:0	0,047	0,009	14	0,006	0,63	39	6	0,005	0,67	8,74
C16:0	1,314	0,288	13	0,095	0,90	18	7	0,087	0,91	8,86
C17:0	0,028	0,005	19	0,002	0,75	60	11	0,002	0,74	1,17
C18:0	0,371	0,095	8	0,055	0,68	28	7	0,048	0,75	12,19
C18:111t10t	0,052	0,017	18	0,011	0,63	57	13	0,010	0,68	8,71
Total:18:1trans	0,083	0,021	18	0,013	0,64	33	12	0,011	0,71	11,36
Total:18:1cis	0,737	0,229	15	0,049	0,95	23	8	0,041	0,97	16,09
Total:18:1	0,819	0,243	14	0,047	0,96	18	8	0,037	0,97	20,29
C18:29t12c	0,006	0,002	13	0,001	0,54	87	15	0,001	0,60	6,56
C18:29c12c	0,052	0,012	19	0,006	0,72	73	15	0,006	0,76	8,67
Total 18:2 n-6	0,058	0,012	16	0,007	0,67	38	15	0,006	0,71	6,44
C18:29c11t	0,018	0,006	18	0,004	0,65	78	7	0,004	0,66	0,05
Total C18:2	0,076	0,013	16	0,008	0,59	36	11	0,008	0,64	7,88
C18:3 n-3	0,018	0,008	19	0,004	0,80	68	7	0,003	0,85	15,21
Total C18:3	0,019	0,007	19	0,004	0,80	38	7	0,003	0,82	10,82
Saturated	2,907	0,492	14	0,047	0,99	38	9	0,045	0,99	3,53
monounsaturated	0,932	0,265	14	0,046	0,97	20	7	0,044	0,97	4,11
polyunsaturated	0,106	0,017	16	0,011	0,58	31	9	0,010	0,62	7,81
Trans	0,108	0,029	18	0,016	0,71	47	14	0,015	0,76	9,41
Omega 3	0,026	0,009	19	0,004	0,76	39	7	0,004	0,79	10,40
Omega 6	0,083	0,016	20	0,009	0,71	36	6	0,008	0,74	8,85

Latent var. : number of latent variables introduced in the PLS regression. SECV : standard error of croos-validation.R2CV cross-validation coefficient of determination. Improvement % : improvement brought by genetic improvement

5 Conclusion

The mid-Infrared Spectrometry is of a strong interest for the prediction of milk fat detailed composition. It is possible to obtain prediction equations rather quickly, but it seems wise to spend time to improve these equations by applying genetics algorithms before. This allows to improve the quality of the predictions and to stabilize the equations over the time. Since null coefficients are applied on discarded wavelengths, we avoid to spoil predictions in case of temperature or pressure change influencing these bands.

Similar task is undertaken on goat and ewe milk. First results concerning goat milk are very interesting. With genetic algorithm we can achieve results comparable with those presented in this paper, whereas we didn't have good predictions by PLS regression for goat milk unlikely ewe and cow milk.

Acknowledgements

The authors thank INRA experimental farms for the technical support and steering committee of Phenofinlait for the constructive discussions.

This study received financial support from Apis-Gène, French Ministry of Agriculture and France Génétique Elevage.

6 References

[1] Leardi R., Lupiañez G. (1998). Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemometrics and Intelligent Laboratory Systems*, 41:195-208.

[2] Hoskuldsson A (2001). Variable and subset selection in PLS regression. *Chemometrics and intelligent laboratory systems*, 55, 23-38.

[3] Kramer J. K. G. et al., 1997. Evaluating acid and base catalysts in the methylation of milk and rumen fatty acids with special emphasis on conjugated dienes and total trans fatty acids. *Lipids*, Vol.32, N°11: 1219-1228.

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

[5] Soyeurt H, Dardenne P, Dehareng F, Lognay G, Veselko G, Marlier M, Bertozzi C, Mayeres P, Gengler N (2006). Estimating Fatty Acid Content in Cow Milk UsingMid-Infrared Spectrometry. *Journal of Diary Science*, 89:3690-3695.

[6] Bertrand.D, Dufour.E, coord (2006). La spectrometrie infrarouge et ses applications analytiques. Tec&Doc Lavoisier, Paris.