Determination of protein composition in milk by mid-infrared spectrometry : comparison of methods

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

(1) Institut de l'Elevage
(2) INRA GABI
(3) CNIEL
(4) INRA SAGA
(5) Actilait

www.phenofinlait.fr

phenofinlait@inst-elevage.asso.fr

PhénoFinlait

Context

Consumers are aware of the food impact on their health and on their environment

→ But there is no cheap and large scale easy to use methods to determine nutritional quality and environmental impact of milk production

⇒"Fine" milk composition analysis could be an answer

PhénoFinLait: aims

- **Develop and control methods to analyze fine milk composition**
- High scale analysis of milk composition and implementation of a huge data base
- Understand how genetic and feeding strategies impact fine milk composition
- Create tools (genetics + feeding strategies) to face evolving consumer demands including health requirements

Protein milk characteristics

6 main lactoproteins

- *α_{s1}-Casein*
- *α_{s2}-Casein*
- *β-Casein*
- *κ-Casein*

} Caseins (among 80%)

- *β-Lactoglobulin (β-LG)*
- *α-Lactalbumin (α-LA)*

} Soluble whey proteins (among 20%)

Reference method

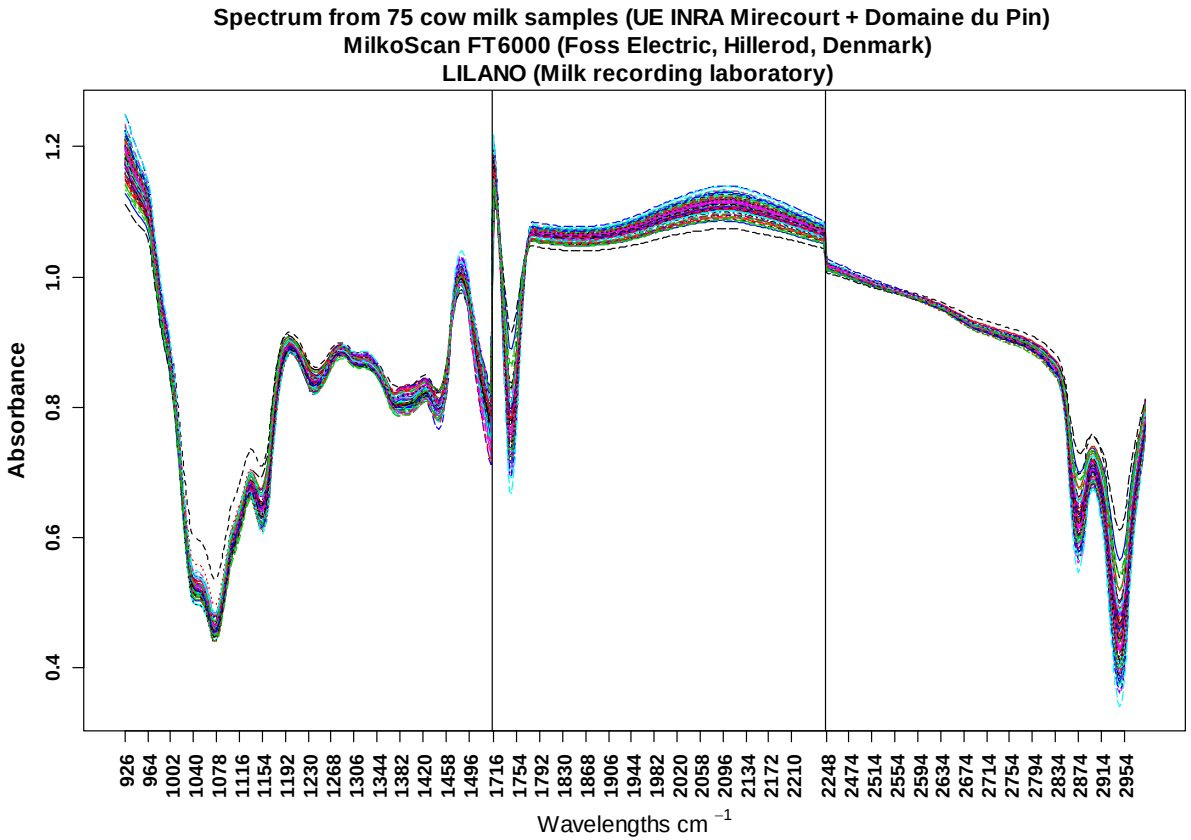
Need to establish a reference method to identify and quantify major lactoproteins:

→ **Liquid chromatography + Mass spectrometry**

Creation of a masses database taking into account genetic variants, post-translational modifications and main proteolysis products.
(Miranda et al.)

Alternativ method

MIR spectra routinely obtained by milk recording laboratories for fat and protein percentage measurements



Can also be used to predict FA and protein composition in cow milk (Soyeurt, 2006 - De Marchi, 2009)

Prediction of protein composition

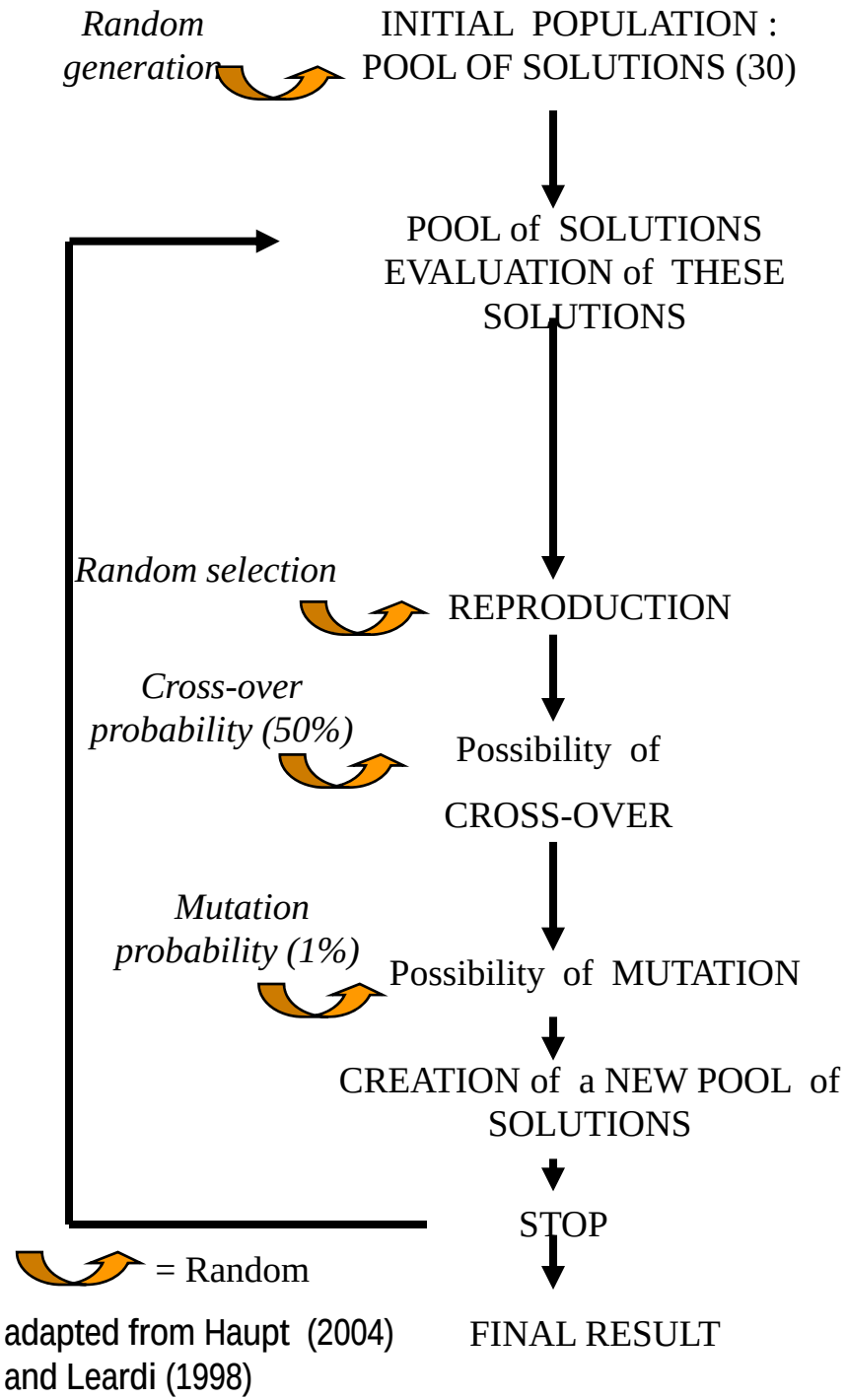
- **193 milk samples** from holstein, normande and montbéliarde cows analyzed by MIR spectrometry and reference method
- Spectra recording from 5012 to 926 cm^{-1}
- **436 wavelengths** are kept (Foss, 1998)
- **No pre-treatments**
- In a first time development of **predictive equations by PLS regression** for 7 proteins and 2 sums

How to improve equations accuracy ?

- **Pretreatments** can be useful to eliminate spectral variations → derivation to eliminate uncontrolled spectral variations
- Several authors have suggested **to apply a selection of variables before PLS regression** to improve results (Leardi 1998, Hoskuldsson 2001)
- Genetic algorithms already successfully used on IR data (Leardi R. 1998, Gomez-Carracedo 2007)
  → Previous study on fatty acids with good results (Ferrand, 2009)
- In genomic selection penalization method like LASSO, Ridge Regression or Elastic Net are used (Croiseau, 2011)

Genetic algorithms method

- Based on evolutionary biology
- **Principle**: evolution of a population of solutions using genetic operators like reproduction. mutation and selection
- **Objective**: obtain a population with the best solutions



adapted from Haupt (2004) and Leardi (1998)

N solutions generated at random

Evaluation

	Var1	Var2…	Var446	$R2_{CV}$
Solution 1	1	1	… 1	
Solution 2	1	0	… 1	
…				
Solution N	0	1	… 0	

Variable i takes value of 1 if selected , else 0. $R2_{CV}$ is obtained by PLS regression on selected variables.

Selection of 2 solutions
The better a solution is, the highest the probability of being chosen is

Combination of 2 solutions
Objective : to obtain 2 better solutions
Limit : variability of solutions decreases

Each variable has a mutation probability of x% (1 no selected variable become selected and conversely)
Objective : avoid having a pool of uniform solutions

Substitution of the 2 worst solutions by new solutions

When quality of solutions is constant, algorithm is stopped.

Getting N solutions among the bests

Penalization method (1/2)

Aim : to reduce the variance of estimators to guarantee the stability of the estimations

- Ridge Regression (RR) : all the predictors are kept
- LASSO : some coefficients are set to zero and in presence of collinearity, only one predictor of the group is retained
- Elatic Net (EN) : combination of RR and Lasso (two penalization parameters) → more flexible

Penalization method (2/2)

avec y_i une teneur en protéines pour le lait i et x_i le vecteur des absorbances (j) de ce lait

- RR : $$\hat{\beta} = \arg\min \left\{ \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \sum_j \beta_j{}^2 \right\}$$
- LASSO : $$\hat{\beta} = \arg\min \left\{ \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \sum_j \left|\beta_j\right| \right\}$$
- Elatic Net (EN) :

$$\hat{\beta} = \arg\min \left\{ \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \left((1-\alpha) \sum_j \beta_j{}^2 + \alpha \sum_j \left|\beta_j\right| \right) \right\}$$

Results: selected wavelengths

- GA : Selection of **8 to 83 variables** out of 446 in the form of wavelengths bands
- LASSO : Selection of **4 to 29 variables** out of 446 in the form of wavelengths bands
- EN : Selection of **22 to 68 variables** out of 446 in the form of wavelengths bands
- 2272-1944 cm^{-1} band rarely selected
- 2970-2278 cm^{-1} and 2272-1944 cm^{-1} selected for most proteins

				Sy,x /Mean (%)						
	N	Mean	Sd	PLS1	dérivée + PLS1	AG 1 tour + PLS1	AG 2 tours + PLS1	EN (α=0,5)	EN (α=0,5) + PLS1	LASSO + PLS1
Casein	58	2,457	0,269	3,93	3,72	3,88	3,85			
κ-CN glycosylée	57	0,11	0,032	26,4	24,12	26,87	25,99	28,47	28,49	26,97
κ-CN	57	0,316	0,052	11,61	10,89	13,42	13,15	14,05	14,56	14,59
α_{S2}-CN	58	0,237	0,041	11,25	10,43	10,29	10,59	11,43	11,76	11,64
α_{S1}-CN	58	0,861	0,099	6,32	6,86	5,47	6,31	6,37	6,97	6,65
β-CN	58	1,041	0,132	7,22	6,04	5,91	6,7	7,09	6,38	6,99
Whey protein	58	0,387	0,06	9,96	9,64	13,67	9,35			
α-LA	57	0,123	0,018	11,8	10,9	10,9	10,91	12,92	13,6	13,5
β-LG	58	0,263	0,054	15,79	15,86	15,29	15,39	16,63	16,28	15,89

Results: improvement

- Good prediction for 3 proteins and correct prediction for 7 proteins
- Similarity between the different methods
- Light advantage to derivate and genetic algorithm

Conclusions

- Ambitious multispecies program with a lot of stakes
- Importance to produce robust and accurate equations
- Wavelength selection before PLS regression could be of a strong interest to predict individual milk protein profile by improving the quality of the predictions and stabilizing the equations over the time, but not real improvement with penalization methods or genetic algorithms
- In a first time, necessity to increase the size of dataset

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE
DE L'ALIMENTATION
DE L'AGRICULTURE
ET DE LA PÊCHE

APIS-GENE

INSTITUT DE L'ELEVAGE

Cniel
Centre national interprofessionnel de l'économie laitière

INRA

Actilait

Thanks to every partners of this project

Thank you for you attention !

www.phenofinlait.fr

phenofinlait@inst-elevage.asso.fr

PhénoFinlait

Genetic algorithms use

- Use of the algorithm developed by Leardi
- Check of the robustness by varying parameters (previous study)
- Fitness function: cross-validated explained variance
- Population size: 30 solutions
- Mutation probability: 1%
- Number of GA runs: 5 (to ensure an optimal convergence)

Répétabillité (sr relatif)

TP	Casein	k-CN glyco	k-CN	α_{S2}-CN	α_{S1}-CN	β-CN	Whey	α_LA	β-LG
0,2	0,8	5,9	3,7	0,6	0,3	0,2	1,7	2,3	2,1

Protéines Individuelles Majeures
Fidélité FTMIR

Protocole d'essai interlaboratoire oct.2008

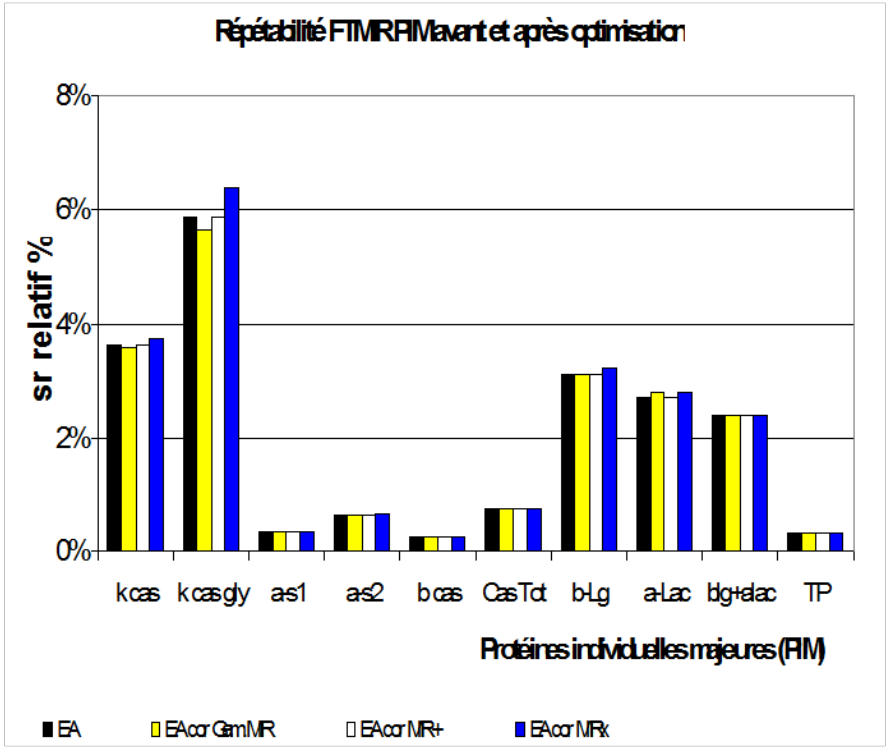
- 8 laboratoires de contrôle laitier x 12 Milkoscan FT6000
- 15 échantillons laits individuels de vache (EA 1 à 15) : gamme test
- 13 laits de calibrage Cecalait
- **Analyses** en triples consécutifs dans standardisation spectrale Foss
- **Prédiction** FTMIR des taux des protéines individuelles majeures
- **Equations** PFL bovin **avec** & **sans** dérivation 1ère de spectre
- **Correction des prédictions**:
  - 1- Régression linéaire sur 13 points,
  - 2- additive sur 1 point,
  - 3- multiplicative sur 1 point

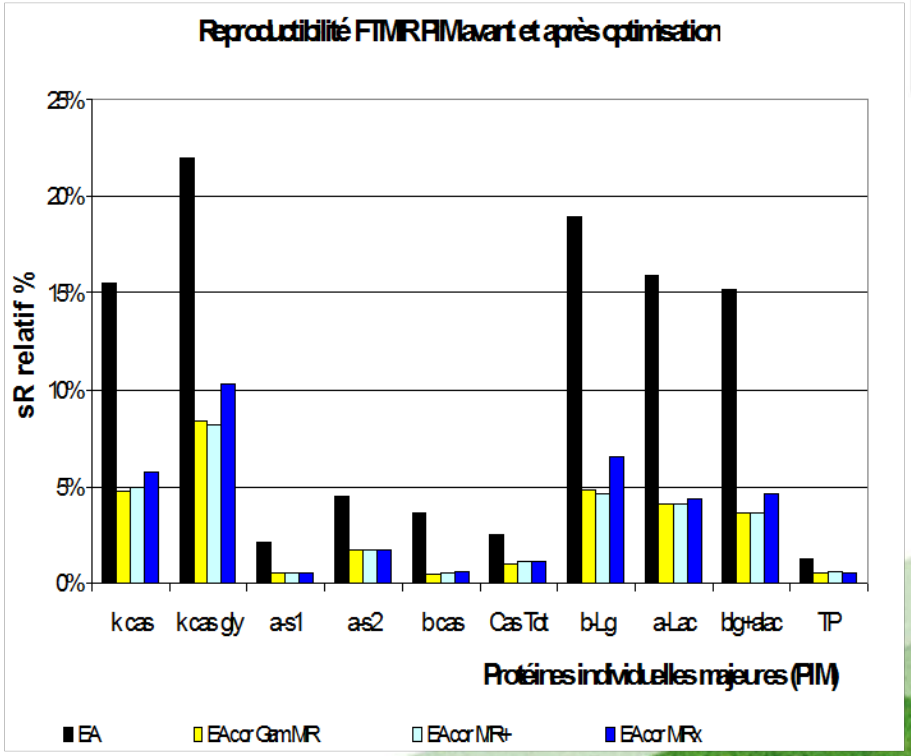
> **Analyse statistique** : Fidélité ISO 5725 / Comparaison des r et R

Résultats

1- Biais entre appareils : significatifs dans une standardisation appareil constructeur, avec et sans dérivation spectrale

2- Correction : régression linéaire 13 points, correction additive et multiplicative sur 1 point, améliore la reproductibilité.





References

Haug A. Bovine milk in human nutrition – a review. *Lipids in Health and Disease* 2007. **6**:25. 2007.

Hoskuldsson A. Variable and subset selection in PLS regression. *Chemometrics and Intelligent Laboratory Systems*. 55 :23-38. 2001.

Leardi R. and Lupiañez G. Genetic algorithms applied to feature selection in pls regression : how and when to use them. *Chemometrics and Intelligent Laboratory Systems*. 41:195-208.1998.

Legrand P. Interêt nutritionnel des principaux acides gras des lipides du lait. *Cholé-doc*.105:1-4. 2008.

Schennink and al.. Genome-wide scan for bovine milk-fat composition**.** *J. Dairy Sci*. 92 :4676–4682. 2009.