Determination of protein composition in milk by mid-infrared spectrometry

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

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

PhénoFinlait

www.phenofinlait.fr

phenofinlait@inst-elevage.asso.fr

Outline

Context and motivations

Materials and methods

Results

Conclusions and perspectives

Context

Milk = complex product with a lot of components

- nutritional interests
- technological properties

→no cheap and large scale easy to use method to measure all milk components

PhénoFinlait: aims

- Develop and control methods to analyze fine milk composition easily
- Use the analytical development to
  - study genetic and feeding management impact on milk composition
  - build up new tools to manage milk composition (Dairy Herd Improvement (DHI) and genomic)

Outline

Context and motivations

Materials and methods

Results

Conclusions and perspectives

Major milk proteins

6 main milk proteins

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

} Caseins (*ca.* 80%)

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

} whey proteins (*ca.* 20%)

Reference method (Miranda *et al.*)

Need to establish a reference method to identify and quantify major milk proteins:

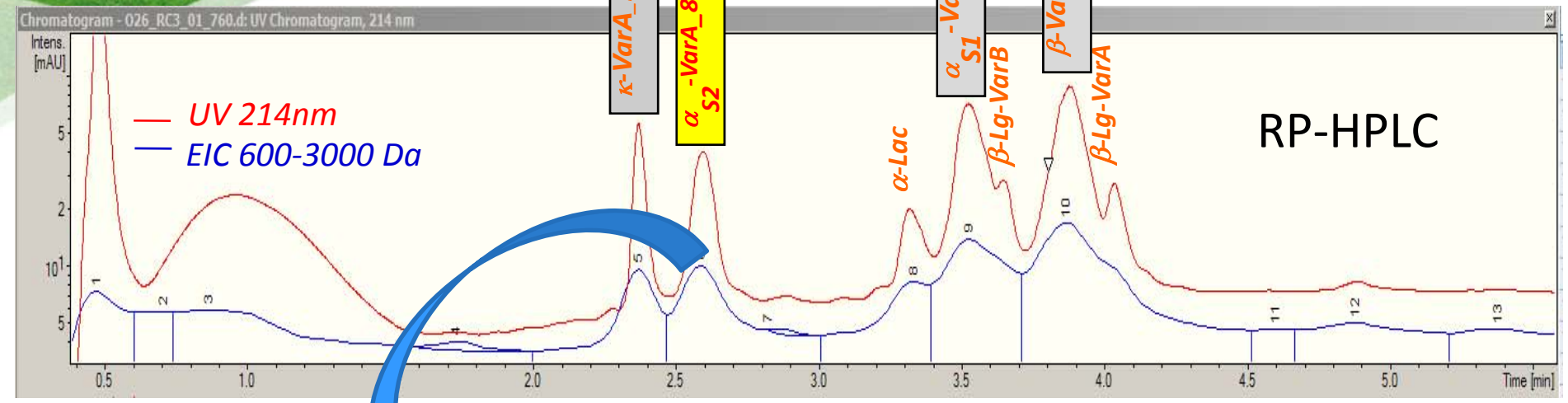
→ Liquid Chromatography + Mass Spectrometry (LC-MS)

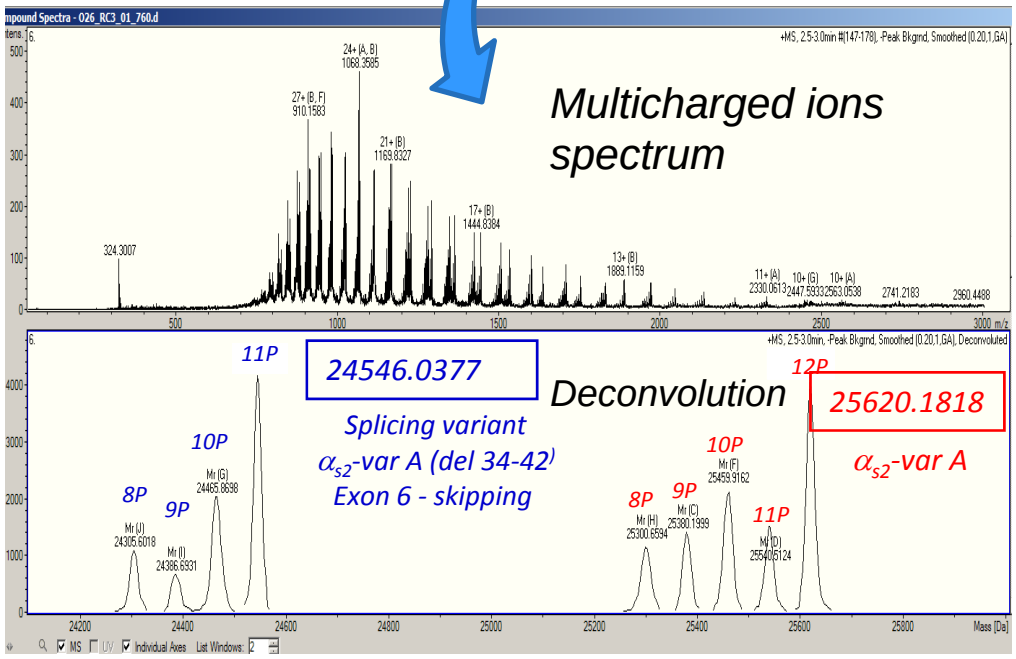
Creation of a database of masses including genetic variants, splicing variants, post-translational modifications and main proteolysis products.

Bovine: 3000 referenced masses
Ovine: 1700 referenced masses

LC-MS milk proteins profiling

Ovine milk



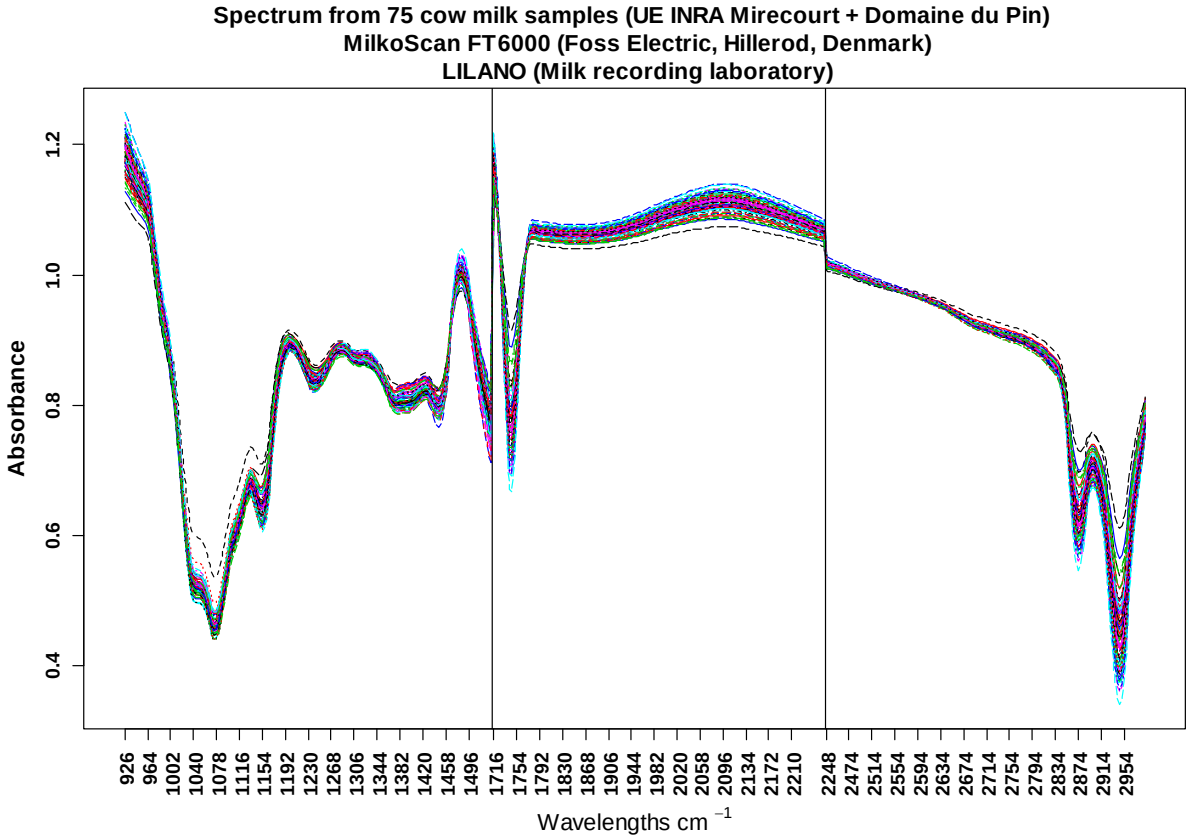


Bovine milk: 25 molecules identified

5 κ-Cn isoforms (1 to 3 glycosylation motifs)
2 κ-Cn isoforms (phosphorylation levels)
5 α_{s2}-Cn isoforms (phosphorylation levels)
2 α_{s1} and β-Cn isoforms (phosphorylation)
1 α_{s1}-Cn splicing variant (Del Q78)
2 β-Lactoglobulin genetic variants
1 α-Lactalbumin genetic variant
5 β-Cn fragments (γ-Cn and complementary fragments arising from plasmin proteolysis)

Method in routine

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



Already used to estimate FA and protein composition in cow milk (Soyeurt, 2006 – Rutten, 2011 – Bonfatti, 2011)

Development of equations

- Traditionally by **PLS regression**
- **Pretreatments** can be useful to eliminate spectral variations → derivation to eliminate uncontrolled spectral variations (Soyeurt, 2011)
- 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 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

- Optimization method based on evolutionary biology
- **Principle**: evolution of a population of solutions (=wavelength selection) using genetic operators like reproduction, mutation and selection
- **Objective**: obtain a population with the best solutions (=wavelength selection)

Penalization method

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
- Elastic Net (EN): combination of RR and LASSO (two penalization parameters) → more flexible

Samples analyzed

- **193 cow milk samples** from Holstein, Normande and Montbéliarde cows analyzed by MIR spectrometry and the reference method
- **153 ewe's milk samples** from Lacaune and Manech tête rousse
- **153 goat milk samples** from Saanen and Alpine

Outline

Context and motivations

Materials and methods

Results

Conclusions and perspectives

Cow milk: selected wavelengths

	GA	LASSO	EN
Number of retained wavelengths	**8 to 83**	**4 to 29**	**22 to 68**

- 2272-1944 cm^{-1} band rarely selected
- 2970-2278 cm^{-1} and 2272-1944 cm^{-1} selected for most proteins

Previous study

Cow milk (independant validation)	N	Mean	Sd	Sy,x /Mean (%)						
				PLS1	dérivée + PLS1	AG 1 tour + PLS1	AG 2 tours + PLS1	EN (α=0,5)	EN (α=0,5) + PLS1	LASSO + PLS1
Caseins	58	2.457	0.269	3.93	3.72	3.88	3.85			
glycosylated κ-CN	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 proteins	58	0.387	0.06	9.96	9.64	13.67	9.35			
α-LA	57	0.123	0.018	11.8	10.90	10.90	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

Ewe's milk (cross-validation)	N	Mean	Sd	Sy,x /Mean (%)			R^2_{CV}		
				PLS1	AG 1 tour + PLS1	AG tour + PLS1	PLS1	AG 1 tour + PLS1	AG tour + PLS1
Caseins	149	4,09	0,67	2,92	2,73	3,02	0,97	0,97	0,97
κ-CN	146	0,44	0,07	7,02	6,71	6,57	0,80	0,82	0,82
α_{S2}-CN	147	0,56	0,10	7,46	9,41	9,50	0,83	0,73	0,72
α_{S1}-CN	143	1,22	0,21	5,06	5,45	5,62	0,91	0,91	0,90
β-CN	148	1,84	0,32	4,67	4,14	4,33	0,93	0,94	0,94
Whey proteins	146	0,56	0,10	9,55	8,89	9,29	0,69	0,73	0,71
α-LA	144	0,15	0,03	17,42	17,12	16,45	0,20	0,21	0,26
β-LG	145	0,41	0,09	10,47	10,49	17,31	0,77	0,77	0,38

Application of ovine equations on PFL MIR-database

g/100 ml	Mean	Std
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	0,640	0,089
α-LA	0,158	0,017
β-LG	0,472	0,085

Outline

Context and motivations

Materials and methods

Results

Conclusions and perspectives

Conclusions

- In first place, to have robust equations, it seems fundamental to have a robust sample dataset with variability and accurate measurements by the reference method
- Gain of accuracy by reallocating the proteolysis
- To implement these equations at a large scale, it is also central to establish an harmonization system between laboratories (Leray *et al.*, 2011)

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

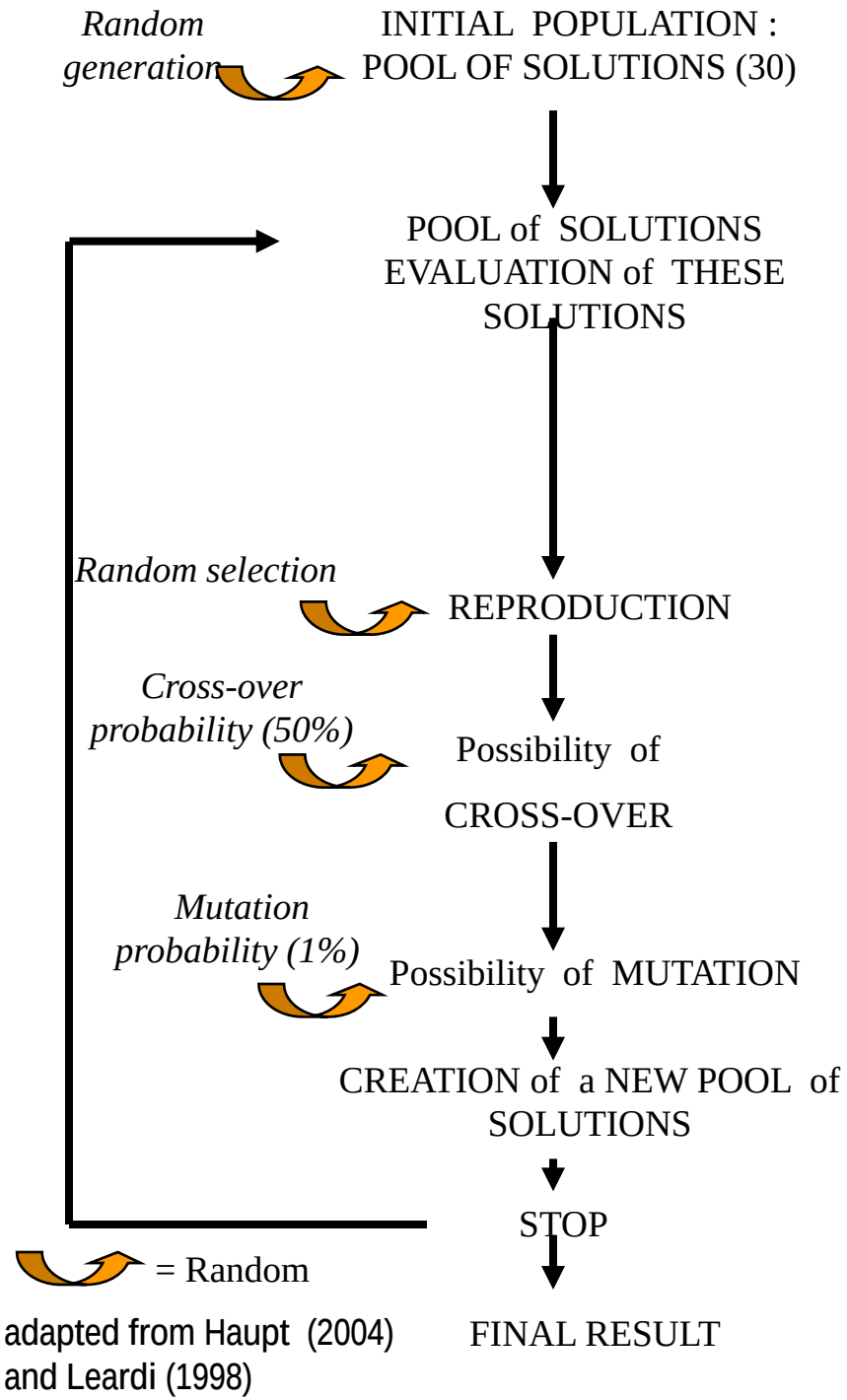
Many thanks to every partners of the project

Thank you for you attention !

www.phenofinlait.fr

phenofinlait@inst-elevage.asso.fr

PhénoFinlait



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

Genetic algorithms use

- Use of the algorithm developed by Leardi
- Check 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)