Use of genetic algorithm on mid-infrared spectrometric data: Application to estimate fatty acid profile of goat milk

M. FERRAND and B. HUQUET
(1 – Institut de l’Elevage).

F. Bouvier (2 - UE Bourges), H. Caillat (3 – INRA-SAGA),
F. Barillet (3), M. Brochard (1), F. Faucon (1,4 - CNIEL),
H. Larroque (3), O. Leray (5 - Actilait)., I. Palhière (3)

www.phenofinlait.fr

phenofinlait@inst-elevage.asso.fr

PhénoFinlait

Context

- Consumers are aware of the food impact on their health, especially FA
- In France, more and more farmers are paid on the FA composition of their milk

But…

⇒ No reference method to routinely analyze milk FA composition

⇒ No tools (animal genetic and feeding strategy) to adapt fine milk composition to consumers demand

Goat milk characteristics

- Higher concentration of short and medium chain fatty acids and lower level of palmitic acid (Tomotake, 2006)
- Fatty acid composition depends on diet but also on genotype at the αs1 casein gene (Mahé, 1994)

→ No knowledge at a large scale of factors affecting fine goat milk composition and of QTL responsible for the composition variation

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

Method choice

- MIR spectra routinely obtained by milk recording laboratories for fat and protein percentage measurements
- Can also be used to predict FA composition in cow milk (Soyeurt et al. 2006)

Prediction of FA composition

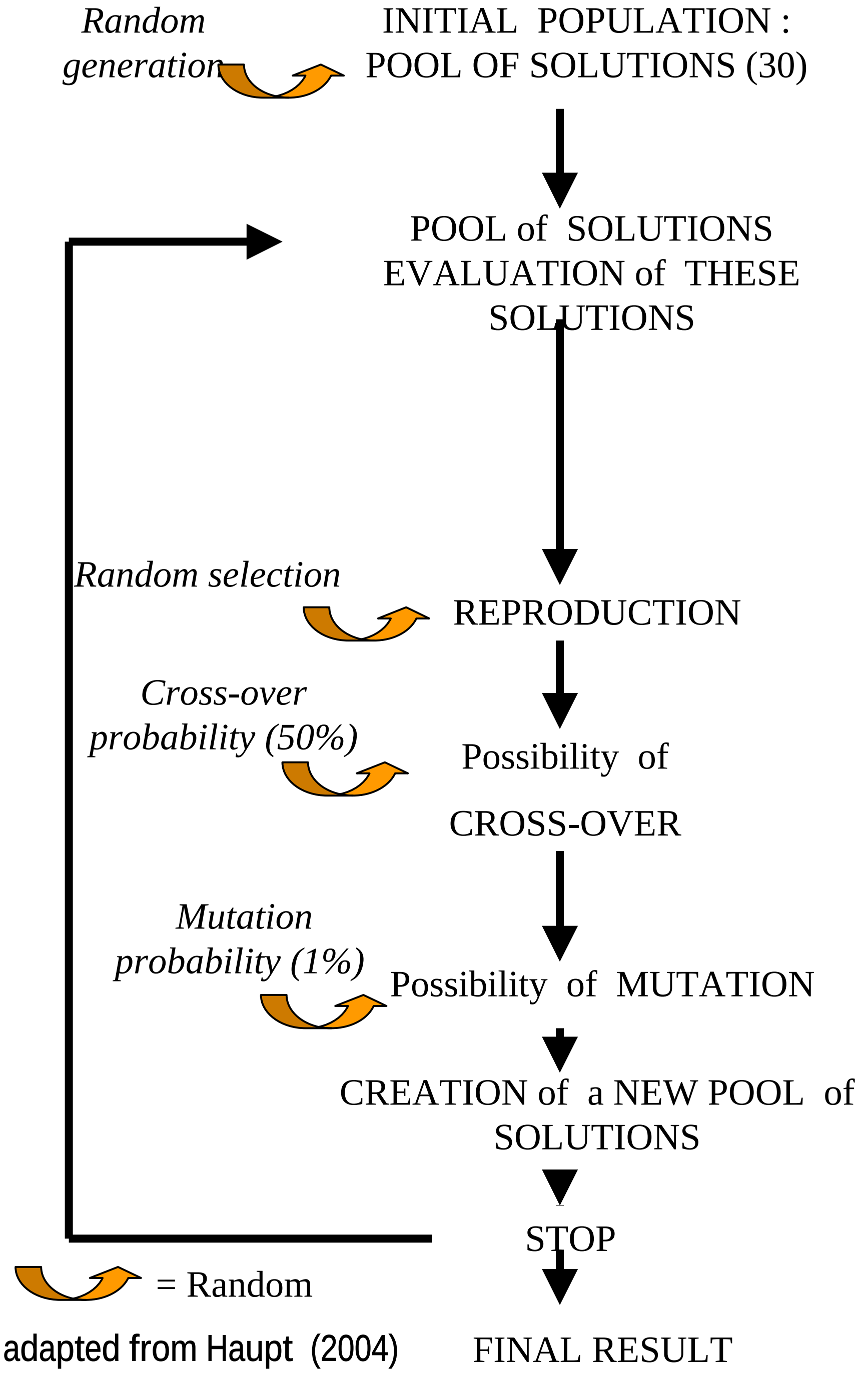
- **149 milk samples** from Alpine dairy goat analyzed by MIR spectrometry and gas chromatography
- Spectra recording from 5012 to 926 cm^{-1}
- **446 wavelengths** are kept (Foss, 1998)
- **No pre-treatments**
- In a first time development of **predictive equations by PLS regression** for 64 FA and some ratios
- Good prediction for 9 FA and correct prediction for 8 FA: estimations not as good as in cow milk (16+14 FA)

How to improve equations accuracy ?

- 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 in cow milk with good results (Ferrand, 2009)

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


Random generation
INITIAL POPULATION : POOL OF SOLUTIONS (30)
POOL of SOLUTIONS EVALUATION of THESE SOLUTIONS
Random selection
REPRODUCTION
Cross-over probability (50%)
Possibility of CROSS-OVER
Mutation probability (1%)
Possibility of MUTATION
CREATION of a NEW POOL of SOLUTIONS
STOP
FINAL RESULT
= Random

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 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)

Results: selected wavelengths

- Selection in average of **72 variables** out of 446 in the form of wavelengths bands (46 in cow milk)
- 2272-1944 cm^{-1} band rarely selected
- 2970-2278 cm^{-1} and 2272-1944 cm^{-1} selected for most fatty acids

Results: improvement

- Good prediction for 9 FA and correct prediction for 10 FA
- Accuracy gain of 7% on average
- Notable improvement for FA of a crucial interest regarding nutrition (C14:0, C16:0…)
- Stabilization of the equations over the time

			PLS2		GA+PLS1 or PLS2		
	Mean	Sd	SECV	R2CV	SECV	R2CV	Improvement
C12:0	0,134	0,041	0,023	0,69	0,019	0,81	18%
C14:0	0,307	0,077	0,034	0,82	0,029	0,87	13%
C16:0	0,996	0,197	0,059	0,92	0,053	0,93	10%
C18:29c12c	0,086	0,020	0,012	0,67	0,012	0,69	4%
C18:29c11t	0,017	0,005	0,004	0,45	0,003	0,55	8%
C18:3n-3	0,013	0,004	0,003	0,41	0,003	0,44	2%
Saturated	2,351	0,485	0,087	0,97	0,086	0,97	2%
Monounsat.	0,798	0,184	0,074	0,85	0,073	0,85	1%
Polyunsat.	0,128	0,028	0,018	0,63	0,016	0,67	6%
Trans	0,100	0,031	0,021	0,53	0,020	0,60	6%

Limits

- High computing time required (3 hours per fatty acid)
- Several manual stages: important error risk, variable results between individuals

Conclusions

- Ambitious multispecies program with a lot of stakes
- Importance to produce robust and accurate equations
- Genetic algorithms before PLS regression is of a strong interest to predict individual milk fatty acid profile: improvement of the quality of the predictions and stabilization of the equations over the time
- Validation with new data is planned in the future

Perspectives

- Beyond PLS: alternative methods like wavelets
  - Accuracy improvement?
  - Time efficient methods ?
  - Ease-of-use in routine ?
- Multispecies

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

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

APIS-GENE

INSTITUT DE L'ÉLEVAGE

INRA

Actilait

Thanks to every partners of this project

Thank you for you attention !

www.phenofinlait.fr

phenofinlait@inst-elevage.asso.fr

PhénoFinlait

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.