Determination of fatty acid profile in milk using mid-infrared spectrometry: interest of applying a variable selection before a PLS regression

M. FERRAND and B. HUQUET

(1 – Institut de l'Elevage).

S. BARBEY (2 - UE INRA). F. BARILLET (3 – INRA-SAGA).

M. BROCHARD (1). F. FAUCON (1,4 - CNIEL).

H. LARROQUE (3). O. LERAY (5 - Actilait).

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

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 new consumer demands including health requirements

Method choice

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

Prediction of FA composition

- **154 milk samples** from crossbred cows analyzed by MIR spectrometry and gas chromatography
- Spectra recording from 5012 to 926 cm^{-1}
- **446 wavelengths** are kept
- **No pre-treatments**
- In a first time development of **predictive equations by PLS regression** for 64 FA and some ratios
- Good prediction for 16 FA and correct prediction for 14 FA

How to improve equations accuracy ?

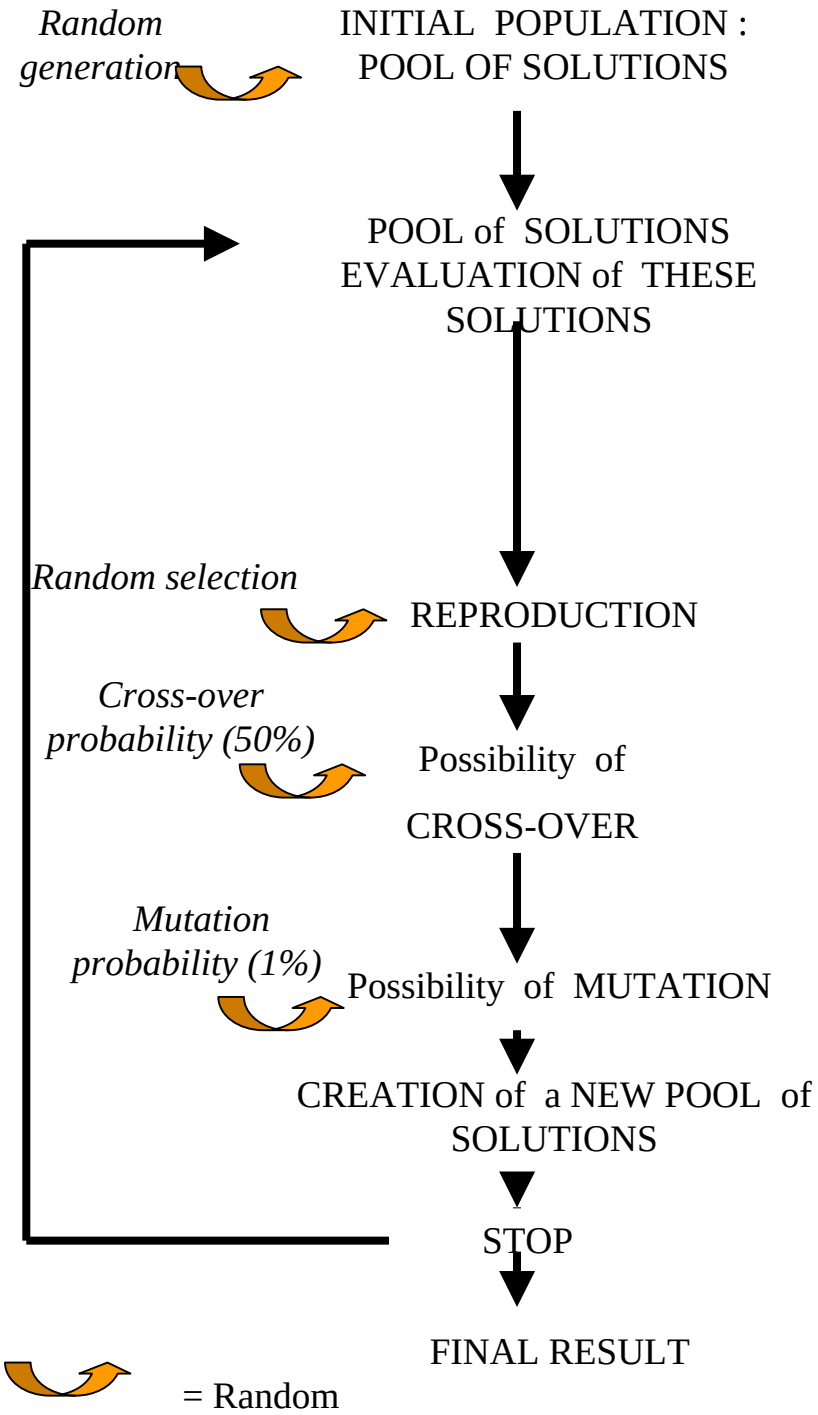
- Several authors have suggested **to apply a selection of variables before PLS regression** to improve results

  (Hoskuldsson A. 2001 – Leardi R. 1998)

- Genetic algorithms already successfully used on IR data (Leardi R. 1998)

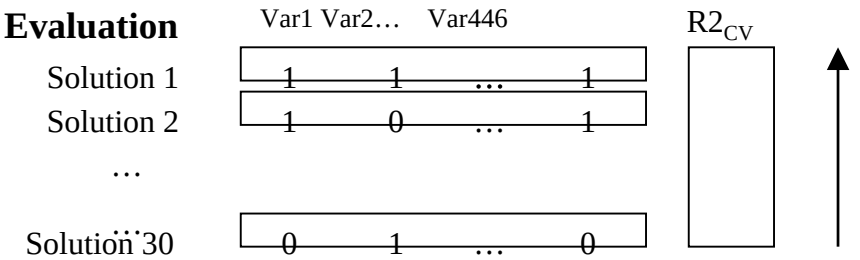
Genetic algorithms method

- 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



30 solutions generated at random

Evaluation



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 1% (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 30 solutions among the bests

Genetic algorithms use

- Use of the algorithm developed by Leardi
- Check of the robustness by varying parameters
- 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 **46 variables** out of 446 in the form of wavelengths bands
- 2272-1905 cm^{-1} band rarely selected
- 2970-2278 cm^{-1} selected for most fatty acids
- **Specific wavelengths** for saturated FA (1234-1080 cm^{-1}) or C 18:0 family (1003-926 cm^{-1} and 1061-1176 cm^{-1})

Results: improvement

- Good prediction for 19 FA and correct prediction for 14 FA
- Improvement of overall predictions
- Accuracy gain of 9% on average
- Notable improvement for FA of a crucial interest regarding nutrition
- Stabilization of the equations over the time

			PLS2		GA+PLS1 or PLS2		
	Mean	Sd	SECV	R2CV	SECV	R2CV	Improvement
C14:0	0.458	0.094	0.0428	0.81	0.0399	0.83	7%
C16:0	1.314	0.288	0.0952	0.9	0.0868	0.91	9%
C18:0	0.371	0.095	0.0547	0.68	0.0480	0.75	12%
Total 18:1trans	0.083	0.021	**0.0129**	**0.64**	**0.0114**	**0.71**	11%
C18:1 9c	0.709	0.224	0.0479	0.95	0.0385	0.97	20%
C18:3 n-3	0.018	0.008	**0.0036**	**0.80**	**0.0031**	**0.85**	15%
Polyunsat.	0.106	0.017	0.0109	0.58	0.0101	0.62	8%
Trans	0.108	0.029	0.0162	0.71	0.0147	0.76	9%
Omega 3	0.026	0.009	0.0044	0.76	0.0040	0.79	10%
Omega 6	0.083	0.016	0.0088	0.71	0.0080	0.74	9%

Limits

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

Conclusions

- Ambitious 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 planned in the future

Perspectives

- Beyond PLS : alternative methods like wavelets or random forest
  - Accuracy improvement?
  - Time efficient methods ?
  - Ease-of-use in routine ?
- Others species

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.