Determination of protein composition in milk by mid-infrared spectrometry

M. Ferrand\textsuperscript{1}, G. Miranda\textsuperscript{2}, H. Larroque\textsuperscript{4}, S. Guisnel\textsuperscript{1}, O. Leray\textsuperscript{5}, F. Lahalle\textsuperscript{1,3}, M. Brochard\textsuperscript{1}, P. Martin\textsuperscript{2}

\textsuperscript{(1) Institut de l’Elevage} \textsuperscript{(2) INRA GABI} \textsuperscript{(3) CNIEL} \textsuperscript{(4) INRA SAGA} \textsuperscript{(5) Actilit}

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

\( \alpha_{s1} \)-Casein
\( \alpha_{s2} \)-Casein
\( \beta \)-Casein
\( \kappa \)-Casein

\( \beta \)-Lactoglobulin (\( \beta \)-LG)
\( \alpha \)-Lactalbumin (\( \alpha \)-LA)

Caseins (ca. 80%)

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

ICAR - 28/05/2012
Ovine milk

LC-MS milk proteins profiling

UV 214nm
EIC 600-3000 Da

Multicharged ions spectrum

Deconvolution

24546.0377
Splicing variant
αs2-var A (del 34-42)
Exon 6 - skipping

25620.1818

Bovine milk: 25 molecules identified

5 κ-Cn isoforms (1 to 3 glycosylation motifs)
2 κ-Cn isoforms (phosphorylation levels)
5 αs2-Cn isoforms (phosphorylation levels)
2 αs1 and β-Cn isoforms (phosphorylation)
1 αs1-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

Spectrum from 75 cow milk samples (UE INRA Mirecourt + Domaine du Pin)
MilkoScan FT6000 (Foss Electric, Hillerod, Denmark)
LILANO (Milk recording laboratory)

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

<table>
<thead>
<tr>
<th></th>
<th>GA</th>
<th>LASSO</th>
<th>EN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of retained wavelengths</td>
<td>8 to 83</td>
<td>4 to 29</td>
<td>22 to 68</td>
</tr>
</tbody>
</table>

- 2272-1944 cm\(^{-1}\) band rarely selected
- 2970-2278 cm\(^{-1}\) and 2272-1944 cm\(^{-1}\) selected for most proteins
## Previous study

<table>
<thead>
<tr>
<th>Cow milk (independant validation)</th>
<th>N</th>
<th>Mean</th>
<th>Sd</th>
<th>PLS1</th>
<th>dérivée + PLS1</th>
<th>AG 1 tour + PLS1</th>
<th>AG 2 tours + PLS1</th>
<th>EN (α=0,5)</th>
<th>EN (α=0,5) + PLS1</th>
<th>LASSO + PLS1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Caseins</strong></td>
<td>58</td>
<td>2.457</td>
<td>0.269</td>
<td><strong>3.93</strong></td>
<td><strong>3.72</strong></td>
<td><strong>3.88</strong></td>
<td><strong>3.85</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>glycosylated κ-CN</td>
<td>57</td>
<td>0.11</td>
<td>0.032</td>
<td><strong>26.4</strong></td>
<td><strong>24.12</strong></td>
<td><strong>26.87</strong></td>
<td><strong>25.99</strong></td>
<td><strong>28.47</strong></td>
<td><strong>28.49</strong></td>
<td><strong>26.97</strong></td>
</tr>
<tr>
<td>κ-CN</td>
<td>57</td>
<td>0.316</td>
<td>0.052</td>
<td><strong>11.61</strong></td>
<td><strong>10.89</strong></td>
<td><strong>13.42</strong></td>
<td><strong>13.15</strong></td>
<td><strong>14.05</strong></td>
<td><strong>14.56</strong></td>
<td><strong>14.59</strong></td>
</tr>
<tr>
<td>α_{S_2}-CN</td>
<td>58</td>
<td>0.237</td>
<td>0.041</td>
<td><strong>11.25</strong></td>
<td><strong>10.43</strong></td>
<td><strong>10.29</strong></td>
<td><strong>10.59</strong></td>
<td><strong>11.43</strong></td>
<td><strong>11.76</strong></td>
<td><strong>11.64</strong></td>
</tr>
<tr>
<td>α_{S_1}-CN</td>
<td>58</td>
<td>0.861</td>
<td>0.099</td>
<td><strong>6.32</strong></td>
<td><strong>6.86</strong></td>
<td><strong>5.47</strong></td>
<td><strong>6.31</strong></td>
<td><strong>6.37</strong></td>
<td><strong>6.97</strong></td>
<td><strong>6.65</strong></td>
</tr>
<tr>
<td>β-CN</td>
<td>58</td>
<td>1.041</td>
<td>0.132</td>
<td><strong>7.22</strong></td>
<td><strong>6.04</strong></td>
<td><strong>5.91</strong></td>
<td><strong>6.7</strong></td>
<td><strong>7.09</strong></td>
<td><strong>6.38</strong></td>
<td><strong>6.99</strong></td>
</tr>
<tr>
<td>Whey proteins</td>
<td>58</td>
<td>0.387</td>
<td>0.06</td>
<td><strong>9.96</strong></td>
<td><strong>9.64</strong></td>
<td><strong>13.67</strong></td>
<td><strong>9.35</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>α-LA</td>
<td>57</td>
<td>0.123</td>
<td>0.018</td>
<td><strong>11.8</strong></td>
<td><strong>10.90</strong></td>
<td><strong>10.90</strong></td>
<td><strong>10.91</strong></td>
<td><strong>12.92</strong></td>
<td><strong>13.6</strong></td>
<td><strong>13.5</strong></td>
</tr>
<tr>
<td>β-LG</td>
<td>58</td>
<td>0.263</td>
<td>0.054</td>
<td><strong>15.79</strong></td>
<td><strong>15.86</strong></td>
<td><strong>15.29</strong></td>
<td><strong>15.39</strong></td>
<td><strong>16.63</strong></td>
<td><strong>16.28</strong></td>
<td><strong>15.89</strong></td>
</tr>
<tr>
<td>Ewe’s milk (cross-validation)</td>
<td>N</td>
<td>Mean</td>
<td>Sd</td>
<td>PLS1</td>
<td>AG 1 tour + PLS1</td>
<td>AG tour + PLS1</td>
<td>PLS1</td>
<td>AG 1 tour + PLS1</td>
<td>AG tour + PLS1</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------------------</td>
<td>----------------</td>
<td>------</td>
<td>------------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>Caseins</td>
<td>149</td>
<td>4,09</td>
<td>0,67</td>
<td>2,92</td>
<td>2,73</td>
<td>3,02</td>
<td>0,97</td>
<td>0,97</td>
<td>0,97</td>
<td></td>
</tr>
<tr>
<td>κ-CN</td>
<td>146</td>
<td>0,44</td>
<td>0,07</td>
<td>7,02</td>
<td>6,71</td>
<td>6,57</td>
<td>0,80</td>
<td>0,82</td>
<td>0,82</td>
<td></td>
</tr>
<tr>
<td>αs2-CN</td>
<td>147</td>
<td>0,56</td>
<td>0,10</td>
<td>7,46</td>
<td>9,41</td>
<td>9,50</td>
<td>0,83</td>
<td>0,73</td>
<td>0,72</td>
<td></td>
</tr>
<tr>
<td>αs1-CN</td>
<td>143</td>
<td>1,22</td>
<td>0,21</td>
<td>5,06</td>
<td>5,45</td>
<td>5,62</td>
<td>0,91</td>
<td>0,91</td>
<td>0,90</td>
<td></td>
</tr>
<tr>
<td>β-CN</td>
<td>148</td>
<td>1,84</td>
<td>0,32</td>
<td>4,67</td>
<td>4,14</td>
<td>4,33</td>
<td>0,93</td>
<td>0,94</td>
<td>0,94</td>
<td></td>
</tr>
<tr>
<td>Whey proteins</td>
<td>146</td>
<td>0,56</td>
<td>0,10</td>
<td>9,55</td>
<td>8,89</td>
<td>9,29</td>
<td>0,69</td>
<td>0,73</td>
<td>0,71</td>
<td></td>
</tr>
<tr>
<td>α-LA</td>
<td>144</td>
<td>0,15</td>
<td>0,03</td>
<td>17,42</td>
<td>17,12</td>
<td>16,45</td>
<td>0,20</td>
<td>0,21</td>
<td>0,26</td>
<td></td>
</tr>
<tr>
<td>β-LG</td>
<td>145</td>
<td>0,41</td>
<td>0,09</td>
<td>10,47</td>
<td>10,49</td>
<td>17,31</td>
<td>0,77</td>
<td>0,77</td>
<td>0,38</td>
<td></td>
</tr>
</tbody>
</table>
Application of ovine equations on PFL MIR-database

<table>
<thead>
<tr>
<th>Protein</th>
<th>g/100 ml</th>
<th>Mean</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>5,456</td>
<td>0,768</td>
<td></td>
</tr>
<tr>
<td>Caseins</td>
<td>4,399</td>
<td>0,68</td>
<td></td>
</tr>
<tr>
<td>$\kappa$-CN</td>
<td>0,465</td>
<td>0,07</td>
<td></td>
</tr>
<tr>
<td>$\alpha_{s_2}$-CN</td>
<td>0,592</td>
<td>0,131</td>
<td></td>
</tr>
<tr>
<td>$\alpha_{s_1}$-CN</td>
<td>1,375</td>
<td>0,217</td>
<td></td>
</tr>
<tr>
<td>$\beta$-CN</td>
<td>1,908</td>
<td>0,307</td>
<td></td>
</tr>
<tr>
<td>Whey proteins</td>
<td>0,640</td>
<td>0,089</td>
<td></td>
</tr>
<tr>
<td>$\alpha$-LA</td>
<td>0,158</td>
<td>0,017</td>
<td></td>
</tr>
<tr>
<td>$\beta$-LG</td>
<td>0,472</td>
<td>0,085</td>
<td></td>
</tr>
</tbody>
</table>
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).
Many thanks to every partners of the project

Thank you for you attention!
INITIAL POPULATION: POOL OF SOLUTIONS (30)

N solutions generated at random

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

Evaluation

<table>
<thead>
<tr>
<th>Solution 1</th>
<th>Solution 2</th>
<th>Solution N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Variable i takes value of 1 if selected, else 0. \( R_{2CV} \) 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

Getting N solutions among the bests

Random generation

Random selection

Cross-over probability (50%)

Mutation probability (1%)

= Random

adapted from Haupt (2004) and Leardi (1998)
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)