ESTIMATION OF THE EFFECT OF MILK PROTEIN POLYMORPHISM ON PRODUCTION TRAITS IN DAIRY CATTLE BY TAIL ANALYSIS

D. Boichard, F. Grosclaude*, and P. Le Roy
INRA, Station de Génétique Quantitative et Appliquée,
*Laboratoire de Génétique Biochimique et de Cytogénétique, 78 352 Jouy-en-Josas Cedex France

SUMMARY

A Maximum-Likelihood method is presented to estimate gene effects in a design where only extreme animals are genotyped. This method was applied to estimate the effects of caseins and β-lactoglobulin on production traits in the bovine Montbéliarde breed. Out of 3,451 candidates, 505 cows selected within sire on protein content were genotyped. Protein content was found to be affected by the αs1-Cn and β-Cn loci, and fat content by β-Lg. However, this study did not confirm any effect of the κ-Cn locus on protein content or protein yield.

INTRODUCTION

Milk protein genes are natural candidate genes to explain at least a fraction of the variability of dairy traits. In cattle, milk proteins seem to have a smaller effect than in goat (Grosclaude et al., 1987). However, the polymorphism at the κ-Cn locus is known to strongly affect the cheese making properties of milk (Grosclaude, 1988). Since Grosclaude’s review (1988), Mao et al. (1991), Bovenhuis et al. (1992) and Hill (1993) published the results of new large scale studies and concluded to a favourable effect of κ-CnB and αs1-CnC on protein content, and of βLgB on fat content. However, they did not agree regarding potential effects on milk, fat, or protein yields.

Tail analysis requires only the genotyping of most extreme animals for the trait of interest (Lander & Botstein, 1989; Plotsky et al., 1993), but provides biased estimates for the gene effects, unless a Maximum Likelihood analysis is applied to the entire data set. The aim of this study was 1) to present a method to estimate the gene effects when the genotyped animals are selected, and 2) to estimate the frequencies of milk protein haplotypes and their effects on production traits, with a tail analysis on protein content in the bovine Montbéliarde breed.

MATERIAL AND METHODS

The population analysed included 3,451 Montbéliarde heifers born from 30 preselected sires, with a first calving between September 1988 and February 1989, with more than 250 days in milk in first lactation and belonging to milk recorded herds in Eastern France with at least 8 such animals. Milk, fat, and protein yields, and fat and protein contents were analysed, but emphasis was put on protein content. Before any analysis, performances were adjusted for the effects of herd-year, age at calving and month of calving effects, estimated separately in the national genetic evaluation system (Bonaiti & Boichard, 1990). This choice was motivated by the non-exhaustive within-herd sampling, which made it impossible to estimate these effects within the present analysis.

Figure 1. Within sire distribution of protein content

About 20% of the 3,451 candidates were selected within-sire according to their low or high protein content performance. The difference between the low and the high group reached 7g/kg, i.e. 3 within-sire standard deviations (Figure 1). In practice, 505 milk samples were collected by four milk recording organizations. Skimmed milk was analysed by electrophoresis on starch gel. The genotype of the females was determined at four loci: αs1-Cn (B and C alleles), βCn (A1, A2, B, and C), κCn (A and B), and βLg (A, B, and D).

The analysis of the joint effect of the three casein loci was based upon the following model:

\[ y_{ijk} = \mu + s_i + g_j + e_{ijk} \]

with \( y_{ijk} \) being the performance of female \( k \), adjusted for environmental effects, \( \mu \) being a constant, \( s_i \) being the fixed effect of the sire \( i \), \( g_j \) being the effect of genotype \( j \) of the female \( k \),
combining the three casein loci, and eijk being the residual, assumed to be normally distributed with 0 expectation and variance \( \sigma^2 \) (i.e. dams were assumed unrelated). Two analyses were carried out. In the first one, \( g_j \) was defined as the sum of the main effects of the three casein loci, whereas in the second analysis, \( g_j \) was defined as the effect of each combination of the three casein loci, including the main effects and the interactions.

The sire effect was considered as fixed. Consequently, the genotype effects were estimated from within-sire differences only. This choice led to loose some information but minimized the probability to reveal false positive effects. The haplotypes were unknown even for the genotyped females, and were inferred from within-family segregations and population parameters. The following parameters were estimated: the constant \( \mu \) and the sire effects, the genotype effects, the residual variance, the haploptic frequencies in the sire and dam populations, and the probability of false pedigree \( q \). The latter parameter was added in the analysis to account for some incompatibilities between the known genotype of a female and the putative genotype of her sire. In this case, the genotype of the female was kept and her pedigree was discarded. The likelihood could be written as

\[
M_i = \prod_{i=1}^{m_i} \prod_{G_i=1}^{n_i} \sum_{H_i=1}^{n_i} p(H_i) \left[ \prod_{j=1}^{n_j} \sum_{G_j=1}^{n_j} \sum_{G_{ij}} p(G_{ij}/H_i) p(G_j/H_i) f(y_{ij}/G_{ij}) \right] 
\]

\[
M_i = \prod_{i=1}^{m_i} \prod_{G_i=1}^{n_i} \left[ \prod_{j=1}^{n_j} p(G_{ij}/H_i) + q \prod_{j=1}^{n_j} p(G_j) f(y_j/G_j) \right]
\]

with \( m_i \) being the number of sires, \( n_{i1} \) and \( n_{i2} \) being the number of untyped and typed daughters of sire \( i \), respectively, \( G_{ij} \) being the genotype of daughter \( j \) of sire \( i \) (for instance BC A2B AA for the \( \alpha_s \), \( \beta \), and \( \kappa \) casein loci respectively), \( n_g \) being the number of genotypes (\( n_g=54 \)), \( H_i \) being the ordered genotype of sire \( i \) (for instance BBA/C A2A), \( H_{ij} \) being the ordered genotype of daughter \( j \) of sire \( i \), \( n_h \) being the number of ordered genotypes (\( n_h=78 \)), \( p(H_i) \) being the probability of the ordered genotype \( H_i \) given the haploptic frequencies \( f_s \) in the sire population and assuming Hardy-Weinberg equilibrium, \( p(H_{ij}/H_i) \) being the probability of \( H_{ij} \), given \( H_i \) and the haploptic frequencies \( f_d \) in the dam population, \( p(G_{ij}/H_{ij}) \) being equal to 1 if the ordered genotype \( H_{ij} \) corresponded to \( G_{ij} \), and to 0 otherwise, \( f(y_{ij}/G_{ij}) \) being the penetrance, i.e. the density function of the performance given the genotype and the sire, which could be written under normality as

\[
f(y_{ij}/G_{ij}) = (2\pi\sigma^2)^{-1/2} e^{-\frac{(y_{ij}-\mu_{ij}-g_{ij})^2}{2\sigma^2}} \quad g_{ij} \text{ being the effect of genotype } G_{ij} \text{ defined as above, and } f(y_j/G_j) \text{ being the penetrance given the genotype but assuming the sire unknown.}
\]

Absence of genotype effects was tested with the likelihood ratio \( 2\log(M_1/M_0) \), with \( M_0 \) being the likelihood under \( H_0 \) hypothesis of no genotype effects. This ratio asymptotically follows a \( \chi^2 \) distribution, which number of degrees of freedom is the number of parameters of \( H_1 \) fixed under \( H_0 \), i.e. the number of genotype effects minus one. Similarly, the presence of interactions was tested by the ratio of likelihoods for models with and without interactions. The likelihood was maximized by a Quasi-Newton algorithm, which evaluated numerically the Hessian matrix of the log-likelihood. The inverse of this matrix was used to estimate the asymptotic error variance of the vector of parameters. For the \( \beta \)Lg locus, unlinked to the caseins, a single locus model was used, which likelihood \( m_1 \) is

\[
m_1 = \prod_{i=1}^{m_1} \prod_{G_i=1}^{n_g} \left[ \prod_{j=1}^{n_j} p(G_{ij}/G_i) f(y_{ij}/G_{ij}) \right] \left[ \prod_{j=1}^{n_j} p(G_j) f(y_j/G_j) + q \prod_{j=1}^{n_j} p(G_j) f(y_j/G_j) \right]
\]

### RESULTS

Raw allelic frequencies observed in both groups of cows differed only slightly between groups (table 1). However, a higher frequency of \( \alpha_{s1} \)Cn\(^C\) and \( \kappa \)Cn\(^B\) and a lower frequency of \( \beta \)Cn\(^B\) was observed in the high

<table>
<thead>
<tr>
<th>Protein Allele</th>
<th>Low (PC)</th>
<th>High (PC)</th>
<th>(dam)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha_{s1} )-Cn</td>
<td>( \beta )</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>( \beta )-Cn</td>
<td>A</td>
<td>20.6</td>
<td>26.3</td>
</tr>
<tr>
<td>A2</td>
<td>54.8</td>
<td>61.8</td>
<td>62.7</td>
</tr>
<tr>
<td>B</td>
<td>24.6</td>
<td>11.6</td>
<td>17.9</td>
</tr>
<tr>
<td>( \kappa )-Cn</td>
<td>A</td>
<td>64.7</td>
<td>56.0</td>
</tr>
<tr>
<td>C</td>
<td>35.3</td>
<td>44.0</td>
<td>38.2</td>
</tr>
<tr>
<td>( \beta )-Lg</td>
<td>A</td>
<td>36.2</td>
<td>37.7</td>
</tr>
<tr>
<td>B</td>
<td>63.0</td>
<td>61.5</td>
<td>61.6</td>
</tr>
<tr>
<td>D</td>
<td>0.8</td>
<td>0.7</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Table 1. Allelic frequencies (%), observed in low and high protein content (PC) groups, and estimated in the dam population with the single locus model.
The β-CN locus seemed to affect the concentration of milk, since the alleles favourable to protein content also increased fat content and decreased milk yield, without modifying fat and protein yields. The βLg-B allele was found to strongly increase fat content without affecting protein content. This result was in agreement with previous studies (Grosclaude, 1988; Hill, 1993) which reported an effect of the βLg on the casein/protein ratio, but not on the overall protein content.

The κ-CN locus did not affect any of the five analysed traits, in contrast with most previous studies, which generally found a strongly favourable effect of the B allele on protein content and protein yield. The present methodology only accounted for within-sire family information, and information between sires was not used. Whereas the sires with the best protein content proofs were often B carriers, this association could be due to chance or to a founder effect, and could not be confirmed by a within-sire analysis. In other words, the variability of protein content was exactly the same within AA and AB sires, whereas an important genotype effect would have increased this variability in heterozygous sire families.

Excepted the effect of the β-CN locus on milk yield, no significant genotype effect was observed for milk, fat, or protein traits. However, the number of genotyped females was limited and the sampling procedure increased the design power only for the selected trait, i.e. for protein content. The detection power of the design was lowest for fat and protein contents. The κ-CN allele seemed to affect the concentration of milk, since the alleles favourable to protein content also increased fat content and decreased milk yield, without modifying fat and protein yields. The βLg-B allele was found to strongly increase fat content without affecting protein content. This result was in agreement with previous studies (Grosclaude, 1988; Hill, 1993) which reported an effect of the βLg on the casein/protein ratio, but not on the overall protein content.

Significant interactions between casein genotypes were found only for protein content. Whereas no effect of κ-CN was found on protein content in the analysis without interactions, the κ-CN^B allele seemed to have a favourable effect on protein content in association with β-CN^B, no effect in association with β-CN^A2, and an unfavourable effect in association with β-CN^A1 (figure 2). This interaction could explain the lack of effect of the κ-CN locus in the present study, and more generally that the estimate
Figure 2. Within haplotype effect of the $\kappa$-Cn locus on protein content

 could vary between studies, according to the haplotypic frequencies in the different populations. No interaction appeared for the $\alpha_{S1}$-Cn and $\beta$-Cn loci: the favourable effect of the $\alpha_{S1}$-Cn allele on protein content was found in any genotypic combination. Similarly, for the $\beta$-Cn locus, the ranking $A_1>A_2>B$ observed in the analysis without interaction, was confirmed in the genotype combinations that were most represented in the data set.

**CONCLUSION**

The tail analysis is known to increase the detection power of a design, for a given number of typings. However, it can also be used to estimate the genotypic effects, provided that an adequate statistical method accounts for the non-random sampling procedure. However, potential drawbacks are threefolds: 1) such a design is efficient only for one trait, or a set of highly correlated traits; 2) the statistical analysis is more complex and requires a non standard software; 3) it may be less robust than an analysis of complete data and requires a good prior knowledge of the distribution of the random errors.

The protein content in milk seemed to be partially determined by the $\alpha_{S1}$-Cn and $\beta$-Cn loci. This study also confirmed the strong effect of the $\beta$-Lg locus on fat content, but did not confirm the effect of the $\kappa$-Cn locus on protein content observed in many studies. As the design was oriented towards the study of protein content, its power was not large enough to draw general conclusions on yield traits.

**REFERENCES**