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 βlactoglobulin 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 α_{s1}-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-Cn^B and α_{s1}-Cn^C on protein content, and of βLg^B on fat content. However, they did not agree regarding potential effects on milk, fat, or protein yields.

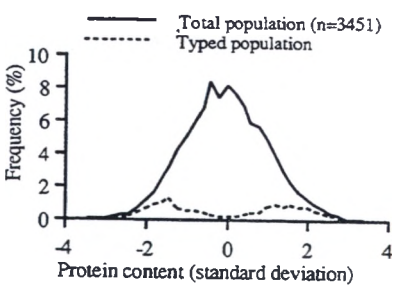
Due to the complete or very close linkage between the three casein loci, a correct analysis should explicitly account for this haplotype transmission of the casein genes.

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 (Bonaïti & 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 : α_{s1}Cn (B and C alleles), βCn (A_1, A_2, 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, μ 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 e_{ijk} being the residual, assumed to be normally distributed with 0 expectation and variance σ^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 infered from within-family segregations and population parameters. The following parameters were estimated : the constant μ and the sire effects, the genotype effects, the residual variance, the haplotypic 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

$$M_1 = \prod_{i=1}^{ns} \sum_{H_i=1}^{nh} p(H_i) \left[\prod_{j=1}^{n_{i1}} \sum_{H_{ij}=1}^{nh} \sum_{G_{ij}=1}^{ng} p(H_{ij} / H_i)\, p(G_{ij} / H_{ij})\, f(y_{ij} / G_{ij}) \right]$$

$$\left[\prod_{j=1}^{n_{i2}} \left\{ (1-q)\, f(y_{ij} / G_{ij}) \sum_{H_{ij}=1}^{nh} p(H_{ij} / H_i)\, p(G_{ij} / H_{ij}) \;+ q\; f'(y_j / G_j) \sum_{H_j=1}^{nh} p(H_j)\, p(G_j / H_j) \right\} \right]$$

with ns 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 A_2B AA for the α_{s1}, β, and κ casein loci respectively), ng being the number of genotypes (ng=54), H_i being the ordered genotype of sire i (for instance BBA / C A_2A), H_{ij} being the ordered genotype of daughter j of sire i, nh being the number of ordered genotypes (nh=78), $p(H_i)$ being the probability of the ordered genotype H_i, given the haplotypic 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 haplotypic 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 $f(y_{ij} / G_{ij}) = (2\pi\sigma^2)^{-1/2}\; e^{-.5(y_{ij}-\mu-s_i-g_j)^2/\sigma^2}$, g_j being the effect of genotype G_{ij} defined as above, and $f'(y_j / G_j)$ being the penetrance given the genotype but assuming the sire unknown.

Table 1. Allelic frequencies (%) observed in low and high protein content (PC) groups, and estimated in the dam population with the single locus model

Protein	Allele	Low PC	High PC	Freq (dam)
α_{s1}-Cn	B	91.7	83.9	89.0
	C	8.3	16.1	11.0
β-Cn	A_1	20.6	26.3	19.4
	A_2	54.8	61.8	62.7
	B	24.6	11.6	17.9
κ-Cn	A	64.7	56.0	61.8
	B	35.5	44.0	38.2
β-Lg	A	36.2	37.7	38.4
	B	63.0	61.5	61.6
	D	0.8	0.7	ne

Absence of genotype effects was tested with the likelihood ratio $2\log(M_1/M_0)$, with M_0 being the likelihood under H0 hypothesis of no genotype effects. This ratio asymptotically follows a χ^2 distribution, which number of degrees of freedom is the number of parameters of H1 fixed under H0, *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 βLg locus, unlinked to the caseins, a single locus model was used, which likelihood m_1 is

$$m_1 = \prod_{i=1}^{ns} \sum_{G_i=1}^{ng} p(G_i) \left[\prod_{j=1}^{n_{i1}} \sum_{G_{ij}=1}^{ng} p(G_{ij} / G_i)\, f(y_{ij} / G_{ij}) \right] \left[\prod_{j=1}^{n_{i2}} \left\{ (1-q)\, f(y_{ij} / G_{ij})\, p(G_{ij} / G_i) + q\, p(G_j)\, f'(y_j / G_j) \right\} \right]$$

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 κCn^B and a lower frequency of βCn^B was observed in the high

Table 2. Haplotypic frequencies (%) for the caseins estimated in the dam population and expected under linkage equilibrium

α_{s1}-β-κ Haplotype	Estimated	Expected
B A_1 A	12.4	10.6
B A_1 B	7.2	6.6
B A_2 A	23.6	34.5
B A_2 B	26.8	21.3
B B A	15.1	9.8
B B B	3.3	6.1
C A_1 A	0.0	1.3
C A_1 B	0.0	0.8
C A_2 A	9.4	4.2
C A_2 B	2.1	2.6
C B A	0.0	1.2
C B B	0.0	0.7

Table 3. Effects of the different loci on fat and protein content

	Fat content		Protein content	
	Est	SE	Est	SE
LR	**22.1**	**	**31.4**	***
α_{s1}Cn				
BB	0.0	0.0	0.0	0.0
BC	0.4	0.3	0.7	0.2
CC	0.2	0.8	0.6	0.5
β -Cn				
A_1A_1	0.0	0.0	0.0	0.0
A_1A_2	-1.8	0.6	-0.7	0.3
A_1B	-1.9	0.6	-0.9	0.3
A_2A_2	-1.8	0.6	-1.0	0.3
A_2B	-2.1	0.6	-1.2	0.3
BB	-3.0	0.7	-1.8	0.4
κ -Cn				
AA	0.0	0.0	0.0	0.0
AB	-0.3	0.3	0.1	0.1
BB	-0.5	0.4	0.0	0.2
β Lg	**6.0**	*	**1.0**	NS
AA	-0.9	0.4	0.0	0.2
AB	-0.3	0.2	-0.1	0.1
BB	0.0	0.0	0.0	0.0

In bold, **LR** = $2\text{Log}(M_1/M_0)$
For βLg, LR = $2\text{Log}(m_1/m_0)$
NS : not significant, * : $p<.05$,
** : $p<0.01$; *** : $p<0.001$

protein content group. Some alleles ($\alpha_{s1}Cn^C$, κCn^B), rather rare in Holstein breed, appeared to be more frequent in this breed originating from the Mid-European Red-and-White populations group, making easier the assessment of their effects.

The dam sample could be considered as representative of the Montbéliarde population. The frequencies of haplotypes B A_2B, BBA, and C A_2A were higher, and those of B A_2A and BBB were lower than expected under linkage equilibrium (table 2). Only 8 haplotypes, among the twelve possible ones, were observed with a non zero frequency. Haplotypes frequency estimates were very similar to those reported by Grosclaude (1988), in spite of a clear decrease in the frequency of the B A_2 A haplotype. Only six haplotypes were found in the sire sample.

The estimated probability of false pedigree reached 7.1 %. It was higher than the observed incompatibility frequency (3%), because of the probability of a false pedigree to remain undetected. Our estimates were of the same magnitude as, or slightly lower than estimates of Bovenhuis & Van Arendonk (1991).

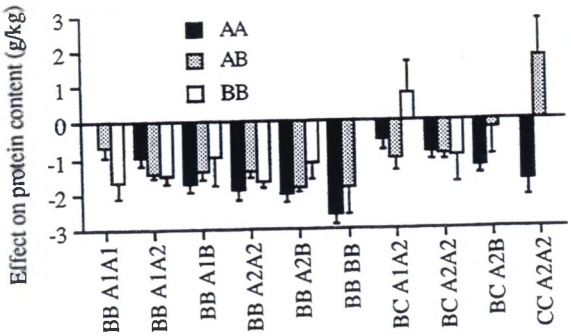
The model without interactions showed a strong and significant effect of the α_{s1}-Cn and β-Cn loci on protein content (table 3). The α_{s1}-Cn^C allele, rare in Montbéliarde, seemed to be favourable to protein content, and slightly unfavourable to milk yield, without affecting the other traits. On the other hand, 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 yields, which are almost uncorrelated with protein content, and intermediate for fat content and milk yield, which correlations with protein content are 0.6 and -0.4, respectively.

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^{A_2}, and a unfavourable effect in association with β-Cn^{A_1} (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 κ–Cn locus on protein content



Genotype at α_{S1}-Cn and β-Cn loci

could vary between studies, according to the haplotypic frequencies in the different populations. No interaction appeared for the α_{S1}-Cn and β-Cn loci : the favourable effect of the α_{S1}-Cn^C allele on protein content was found in any genotypic combination. Similarly, for the β-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-ramdom 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 α_{S1}-Cn and β-Cn loci. This study also confirmed the strong effect of the β-Lg locus on fat content, but did not confirm the effect of the κ-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

Bonaïti, B. and Boichard, D. (1990) Proc. 4th World Congr. Genet. Appl. Livestock Prod. 13 : 364-373.
Bovenhuis, H and Van Arendonk, J.A.M. (1991) J. Dairy Sci., 74 : 2728-2736.
Bovenhuis, H, Van Arendonk, J.A.M., and Korver, S. (1992) J. Dairy Sci., 75 : 2549-2559.
Grosclaude, F. (1988) INRA Prod. Anim., 1 : 5-17.
Grosclaude, F., Mahé, M.F., Brignon, G., Di Stasio, L., and Jeunet, R. (1987) Genet. Sel. Evol., 19: 399-412.
Hill, J.P. (1993) J. Dairy Sci., 76 : 281-286.
Lander, E.S., and Botstein, D. (1989) Genetics, 121 : 185-199.
Mao, I.L., Buttazzoni, L.G., and Aleandri, R. (1992) Acta Agric. Scand. (Sect. A), 42 : 1-7.
Plotsky, Y., Cahaner, A., Haberfeld, A., Lavi, U., Lamont, S.J., and Hillel, J. (1993) Anim. Genet., 24 : 105-110.