SEVENTH FRAMEWORK PROGRAMME

Quantomics

Genome Information Assisted Prediction

Theo Meuwissen

Norw.Univ.Life Sci.

Ås, Norway

Introduction

- GS adopted in many countries
- Increasingly dense SNPchips
  - Moving to sequence data
  - Expectations are high
- However, in cattle : 50k SNPs => 700k
  - Little improvement
  - $\mathbf{G_{50k} \approx G_{700k} \approx G_{sequence}}$

Sequence Data

- A wealth of data, but:
  - what about the signal to noise ratio?
  - 1000 QTN are hidden amongst ~20 million SNPs
    - As needles in a hay stack
- Yet, we need to find them in some way
  - GBLUP and BayesB/C/R seem to use too little info
- Use genomic info beyond SNP genotypes?

SEVENTH FRAMEWORK PROGRAMME

Quantomics

genomic info on SNP beyond genotype?

- In/near exon
- Conserved region (GERP score)
- In/near differetially expressed gene
- In/near eQTL (eQTL data base)
- In/near methylationQTL (mQTL)
- In/near genes of network known to affect trait
- gain/loss of stop codon
- Non-synonymous SNPs…

AIM:

- Develop a GEBV prediction model that utilises extra genomic info about the SNP
- Genome Information Assisted Prediction
- Not the aim:
  - Determine what this extra genomic info actually is

Hierarchical model: 1st layer

- Like BayesC:

$$y_i = \mu + \Sigma_j I_j X_{ij} b_j + e_i$$

X_{ij} = standardised genotype

b_j = SNP effect, with prior N(0,σ_b^2)

I_j = indicator variable whether SNP has effect (0/1)

2nd layer:

- Threshold model for I_j:

$E(I_j)=\pi_j$

$\text{Probit}(\pi_j) = S\beta$

S = a matrix containing extra genomic information

(eg: distance to gene; belong to pathway X (0/1))

β = regression coefficients

(unknown; are estimated)

Extension to random β

- Prior $\beta \sim N(0, \mathbf{A_R}\sigma_\beta^2)$

  $\mathbf{A_R}$ = Autoregressive correlation matrix

  Auto correlation ρ is assumed known

- Thus: π_j depends on π_{j+1} , π_{j-1} ,..

Implementation

- By Gibbs-sampling
  - EM-type of implementation seems possible
- Still fixed parameters:
  - ρ
  - σ_β^2
    - Determines how much π_j can deviate from its mean
- No results yet…

Conclusions

- Model was build to include extra genomic info
  - Distance to gene; pathway info; conserved region…
- Hierarchical model:
  - 1st layer: models *y* as a function of *b*
  - 2nd : models π_j as a function of extra genomic info
- 2nd model may be autoregressive
  - π_j depends on neighboring SNPs