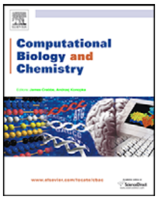
Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem



Research article

Assessing a multiple QTL search using the variance component model

Kateryna Mishchenko[a], Lars Rönnegård[b,*], Sverker Holmgren[c], Volodymyr Mishchenko[c]

[a] School of Education, Culture and Communication, Mälardalen University, Box 883, SE-721 23 Västerås, Sweden
[b] Department of Economics and Social Sciences, Statistic Unit, Dalarna University, Rodavagen 3, 78188 Borlange, Sweden
[c] Division of Scientific Computing, Department of Information Technology, Uppsala University, Sweden

ARTICLE INFO

Article history:
Received 31 March 2009
Received in revised form 7 December 2009
Accepted 8 December 2009

Keywords:
QTL mapping
REML
Variance component estimation
Average information matrix
Forward selection
Hessian approximation
Active set
Primal-dual method

ABSTRACT

Development of variance component algorithms in genetics has previously mainly focused on animal breeding models or problems in human genetics with a simple data structure. We study alternative methods for constrained likelihood maximization in quantitative trait loci (QTL) analysis for large complex pedigrees. We apply a forward selection scheme to include several QTL and interaction effects, as well as polygenic effects, with up to five variance components in the model. We show that the implemented active set and primal-dual schemes result in accurate solutions and that they are robust. In terms of computational speed, a comparison of two approaches for approximating the Hessian of the log-likelihood shows that the method using an average information matrix is the method of choice for the five-dimensional problem. The active set method, with the average information method for Hessian computation, exhibits the fastest convergence with an average of 20 iterations per tested position, where the change in variance components <0.0001 was used as convergence criterion.

1. Introduction

Quantitative trait loci (QTL) are regions on the genome that affect traits measured on a continuous scale. These traits are affected both by several genetic regions and by environmental factors. QTL detection has been a major field of research for several decades (Lynch and Walsh, 1998), where experimental data has shown to be of great importance and has given unique insights to the genetic architecture of quantitative traits (Carlborg and Haley, 2004).

Experimental data, resulting in high power for QTL detection, may be derived by crossing two breeds that are expected to differ genetically. The relationship between trait values and genotypes can be analyzed after two generations of controlled breeding. These experiments are referred to as F_2 intercrosses. A standard statistical tool for analyzing F_2 intercrosses is the simple regression model, which assumes no genetic variation between individuals of the same breed (Haley and Knott, 1992; Broman, 1997; Ljungberg et al., 2002). However, there is often some genetic variation within the two breeds, and this variation may be modeled as a random effect in a more advanced variance component model (Rönnegård and Carlborg, 2007; Perez-Enciso and Varona, 2000).

In a variance component QTL analysis, all the founders of the F_2 intercross are assumed to be unrelated with genes randomly sampled from an outbred population. QTL mapping based on a variance component model is computationally demanding. The computational procedure consists of an *inner problem* and an *outer problem*. In the *inner problem* a variance component model is fitted at a given position in the genome. The value of the likelihood ratio statistic is calculated for this model and is subsequently used in the outer problem. The *outer problem* consists of finding the position, among all tested positions, with highest likelihood ratio value. Hence, the dimensionality of the inner problem is equal to the number of variance components to be estimated, whereas the number of dimensions in the outer problem is given by the number of QTL that we wish to fit simultaneously.

Calculation of the likelihood ratio statistic requires variance component estimation, where restricted maximum likelihood (REML) estimation is used to ensure unbiased estimates of variance components. Variance component estimation consists of a non-linear optimization problem where the computation of the objective function and its derivative is rather costly. Fast variance component estimation programs developed for animal breeding problems (e.g. ASReml (Gilmour et al., 2002) and DMU (Madsen and Jensen., 2008)) are often used in QTL analysis (e.g. Rowe et al., 2009). These variance component estimation programs have been developed to analyze large data sets ($\approx 10^6$ observations) and to compare a moderate number of models (usually < 10). In QTL analysis, however, the size of the data sets are moderate ($\approx 10^3$ observations), whereas the number of models compared are large

[*] Corresponding author. Tel.: +46 241 30034.
E-mail addresses: kateryna.mishchenko@mdh.se (K. Mishchenko), lrn@du.se (L. Rönnegård), sverker.holmgren@it.uu.se (S. Holmgren).

1476-9271/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2009.12.001

(usually > 1000). Consequently, the variance component estimation program developed for QTL analysis needs to be robust so that the algorithm converges for all fitted models. Once the robustness has been verified, further efforts can be made to reduce the computational cost of the calculations.

Variance component estimation algorithms have also been developed for QTL analysis in human pedigrees consisting of independent families (for instance in the SOLAR software (Almasy and Blangero, 1998)), where the size of each family is small. This gives a block-diagonal structure in the variance component model which results in significant simplifications in the computational algorithms and convergence problems does not seem to be an issue. In the current paper, we focus on large complex pedigrees that do not have this simple structure.

A major problem in variance component estimation is that the parameter space is constrained (since variances are > 0). This fact needs to be accounted for by employing established techniques for constrained optimization (e.g. primal-dual and active set methods (Forsgren and Gill, 1998)). Convergence for variance components on, or close to, the parameter boundary may otherwise not be guaranteed. A commonly used algorithm for variance component estimation in animal breeding is the *average information REML* (Johnson and Thompson, 1995), which has been implemented in the ASReml and DMU software (Gilmour et al., 2002; Madsen and Jensen., 2008). The focus has been on speed rather than estimating parameters close to, or on, the boundary in this algorithm, since it is primarily developed for animal breeding applications. For parameter estimates on, or outside, the parameter boundary DMU combines average information REML with an expectation-maximization (EM) algorithm to enable convergence within the parameter space. ASReml does not allow zero variances and sets a lower limit to the variance components equal to a small positive value. To our knowledge, these methods do not guarantee convergence within the parameter space.

Previously we have investigated the possibilities of using active set and primal-dual methods for the simplest possible model with two variance components (Mishchenko et al., 2008), a QTL variance and a residual variance, where the given correlation structure for the QTL variance is low rank or can be approximated by a low rank correlation structure (Rönnegård et al., 2007). Fast computation of projection matrices and matrix inversions has also been derived for the two variance component problem (Mishchenko and Neytcheva., 2009).

In QTL analysis, it is also common to include random polygenic effects as well as QTL effects (Lynch and Walsh, 1998). The correlation structure for polygenic effects (i.e. the additive relationship matrix) is full rank and adds an additional complexity to the variance component estimation problem. Furthermore, possible interaction effects between QTL (i.e. *epistasis*) at several positions on the genome is important to include in the analysis (Carlborg and Haley, 2004). Hence, problems with more than two dimensions for the *inner problem* needs to be studied and will put higher requirements on the computational robustness for the variance component estimation algorithm.

The aim of the current paper is to investigate optimization techniques for the inner problem based on the active set and primal-dual algorithms for constraint optimization, and we apply these schemes for QTL mapping models with 3–5 variance component problems. We wish to find a scheme which is numerically robust and efficient. Moreover, the performance of the schemes using different methods for approximating the Hessian of the log-likelihood are compared. The methods are tested on published data (Carlborg et al., 2006), where the previous analysis was based on a regression model (Haley and Knott, 1992) assuming no within-breed variation. We briefly discuss differences and similarities between our results and these earlier analyses.

2. The restricted maximum likelihood approach

In this section, we consider models where a one-dimensional genome scan is performed for estimating 3–5 variance components. We start by considering a model of a single QTL and additional polygenic effects. Polygenic effects are the combined effects of many genes at different loci each having a small effect (Lynch and Walsh, 1998), whereas a QTL effect is the effect of a restricted part of the genome. The correlation structure for polygenic effects is given by the *additive relationship matrix* and is calculated from pedigree information, whereas the correlation structure for the QTL effect is given by the *identity-by-descent* (IBD) matrix. Elements of the IBD matrix are estimated from pedigree and marker information (Lynch and Walsh, 1998).

2.1. A single QTL and polygenic effects (3D-SCAN)

Variance component analysis for single QTL and polygenic effects is based on a general linear mixed model,

$$y = Xb + Z_1u_1 + Z_aa + e, \tag{1}$$

where y is a vector of n individual phenotypes of a normally distributed trait, X is an $n \times n_f$ design matrix for fixed effects, Z_1 is an $n \times n_r$ design matrix for random effects, b is a vector of n_f unknown fixed effects, u_1 is a vector of n_r unknown random effects for an individual QTL, Z_a is a $n \times n_a$ design matrix for additional polygenic effects, a is a vector of n_a random polygenic effects, and e is a vector of n residuals of random effects. All random effects are assumed to be normally distributed.

For the QTL analysis setting we also assume that the entries in e are identically and independently distributed and that there is a single observation for each individual. Let Π_1 be the IBD matrix and A the additive relationship matrix, then the variance–covariance matrix for (1) is

$$V = \Pi_1\sigma_1^2 + A\sigma_a^2 + I\sigma_e^2, \tag{2}$$

where σ_1^2 is the variance of the random QTL effect, σ_a^2 is the variance of polygenic effects and σ_e^2 is the residual variance.

In REML estimation, the parameters σ_1^2, σ_a^2, σ_e^2 are obtained as maximizers of the restricted likelihood function l of the observed data y. This is done by minimizing the restricted log-likelihood function $L(\Sigma)$ associated with (1),

$$L = -2\ln(l) = C + \ln(\det(V)) + \ln(\det(X^TV^{-1}X)) + y^TPy. \tag{3}$$

Here, C is normalizing constant, Σ is the vector of variance components and the projection matrix P is defined by

$$P = V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}. \tag{4}$$

In summary, we solve the inner problem, i.e. determine the estimates of σ_1^2, σ_a^2, σ_e^2, by solving the optimization problem:

$$\min L(\Sigma) \tag{5}$$

s.t.

$$\sigma_1 \geq 0, \quad \sigma_2 \geq 0, \quad \sigma_3 > 0. \tag{6}$$

Below, we use the notation $\Sigma = (\sigma_1^2, \sigma_a^2, \sigma_e^2) = (\sigma_1, \sigma_2, \sigma_3)$. To determine the main QTL and its effect we need to solve the outer problem and search for the best model fit over the genome. The position τ_0 with the best likelihood value is the most likely position of the main QTL.

2.2. Forward selection for an additional QTL (4D-SCAN)

To solve the problem of finding several QTL, a simultaneous search for them should in principal be performed. For