A comparison of interval mapping procedures including
the QTL-cluster mapping

Hirokazu Matsuda¹ and Hiroaki Iwaisaki¹,²*

¹Course of Environmental Management Science, Graduate School of Science and Technology
²Department of Agro-biology, Faculty of Agriculture
Niigata University, Niigata 950-2181, Japan

(Received 22 November 2004, accepted 17 March 2005)

Using the deterministic sampling, patterns of the log-likelihood surfaces expected in some interval mapping procedures for estimating the position of, and the effect for, QTL(s) were investigated for the situations where a single QTL or closely linked QTLs are contained in a chromosome segment bracketed with two markers. The mapping procedures compared were the conventional, likelihood-based interval mapping (IM), the regression interval mapping (RIM), and the QTL-cluster mapping (CM) in which the conditional probabilities of transmission of the whole segment marked by the flanking markers were taken into consideration. The half-sib design was used here, and several cases of the true genetic model were considered, differing in the number of QTLs contained in the marker interval, the linkage phase for the sire, and the magnitude of the QTL(s) effect. For the true genetic models where a single QTL or closely linked QTLs being in coupling phase are contained in the interval, with (R)IM, clear global maxima of the log-likelihood were observed within the range of the marker interval. It was shown that the estimates of the QTL(s) effect at the marked segment level are expected to be unbiased. On the other hand, in a setting where the linkage phase for the linked QTLs located in the interval was different from coupling and repulsion, there was found a ridge along the interval for the log-likelihood surface with (R)IM, indicating the dependency between the estimates of the position of, and the effect for, the putative QTL. For this case, it was found that the position of the putative QTL could be estimated as that of one of the flanking markers, and the estimate of the QTL effect be biased. In contrast, it was revealed that CM is expected to provide the unbiased estimate of the QTL(s)-effect at the segment level for any case of the true genetic models considered here. If the aim is for marker-assisted selection rather than mapping closely linked QTLs individually, then the CM approach is considered to be useful.

Key words: deterministic sampling, interval mapping, linked QTLs, QTL-cluster mapping

INTRODUCTION

In recent years, a variety of statistical methods to detect and locate quantitative trait loci (QTLs) using genetic marker information have been developed (e.g., Hoeschele et al., 1997). Of these methods, interval mapping using a pair of flanking markers (Lander and Botstein, 1989) provides that the effect and the location of a QTL could be estimated separately, while these estimates could be confounded with each other in the procedure using only a single marker. Interval mapping is likely to be relatively powerful for QTL mapping, when compared to the single-marker approach, and then it has widely been used in animal and plant populations.

With this method, it is allowed that the marker interval contains one QTL, and usually the remainder of QTLs in the same linkage group is not taken into account. Therefore, if there are several QTLs with moderate to large effects in the chromosome, then use of the conventional interval mapping procedure could lead to reduced power for QTL detection, biased estimates of the relevant parameters and detection of a ghost QTL (Knott and Haley, 1992; Martinez and Curnow, 1992; Wright and Kong, 1997). Jansen (1993) and Zeng (1994) addressed these problems by proposing an improved method, or the
so-called composite interval mapping, that is able to absorb the effects of QTLs located outside the marker interval by including additional markers as cofactors in the model and to increase the mapping resolution by reducing the influence of co-segregating QTLs. However, when tightly linked QTLs or a cluster of QTLs exist in the same interval, wrong inference would possibly be made by (composite) interval mapping procedures, since again these usually require the assumption that one QTL is located in the marker interval.

The objective of this study was to investigate properties of some interval mapping procedures under the true situations where a single QTL or the plural number of QTLs exist within one marker interval by using the deterministic sampling. The approach of deterministic sampling was previously applied to examine the nature of the parameter estimates and the convergence of the estimate of the QTL location to an incorrect position on the considered chromosome under model misspecification (Mackinnon and Weller, 1995; Mackinnon et al., 1996; Wright and Kong, 1997). In this study, in addition to the conventional interval mapping, we considered the regression interval mapping (Kao, 2000) and the ‘QTL cluster mapping’ using the chromosome segment model (Matsuda and Iwaisaki, 2001, 2002) that is the mapping using the information on the conditional identity-by-descent (IBD) proportion for the marked chromosome region. The true genetic models for the quantitative trait were described here by certain parameters, in which the number of QTLs contained in the marker interval, the linkage phase, and the effect due to QTL(s) were varied.

METHODS

Experimental design used In this study, a design for QTL mapping based on a single sire family with n half-sib offspring in an outbred population, or the half-sib design resulting from random mating between a single sire and n unrelated dams, was used. It was supposed that the sire is heterozygous for two marker loci, 1 and 2, flanking QTL(s), where the map distance between the two loci is l (in Morgans) and the corresponding recombination rate is $R = 0.5(1 - e^{-2l})$ (Haldane, 1919). The marker-QTL linkage phase for the sire was assumed to be known with the ordered genotype be $M_1Q_1M_2q_2/m_1q_2m_2$, where alleles at QTL and two marker loci are $(Q, q)$, $(M_1, m_1)$ and $(M_2, m_2)$. Then the half-sib offspring could receive one of the four types of marker haplotype, $h_1 = M_1M_2$, $h_2 = m_1m_2$, $h_3 = M_1m_2$ and $h_4 = m_1M_2$, from the sire and the probabilities of these events $(P(h_j); j = 1-4)$ are given as $(1-R)/2$, $R/2$, $R/2$ and $(1-R)/2$, respectively.

Mapping procedures compared Three procedures for mapping QTL(s) were compared herein. One is the conventional interval mapping (denoted as IM), in which a model with a single putative QTL fitted in the marker interval is used. The IM model for a single half-sib family is described as

$$y_i = \mu + \alpha g_i + e_i,$$  \tag{1}

where $y_i$ is the phenotypic value of offspring $i$, $\mu$ is the overall mean of the family, $\alpha$ is the substitution effect at the putative QTL of the sire with the effects of the two alleles $Q$ and $q$ being equal to $1/2\alpha$ and $-1/2\alpha$, respectively, $g_i$ is the indicator variable taking 1/2 if offspring $i$ receives the $Q$ allele and $-1/2$ if receives the $q$ allele from the sire, and $e_i$ is the residual effect ($N(0, \sigma^2)$). Then, denoting the position of the putative QTL in the marker interval by $\hat{d}$, and the resulting recombination rates between the QTL and the left and the right markers, 1 and 2, by $\hat{r}_l = 0.5(1-e^{-2\hat{d}})$ and $\hat{r}_R = 0.5(1-e^{-2(l-\hat{d})})$, respectively, and representing the probabilities of transmission of the QTL alleles from the sire to the offspring, given marker haplotype, by $P(Q|h_i) = p_1$ and $P(q|h_i) = 1-p_1$, we have the conditional probabilities:

$$p_1 = p(M_1Q_1M_2) = \frac{(1-\hat{r}_l)(1-\hat{r}_R)}{1-R},$$

$$p_2 = p(m_1Q_1m_2) = \frac{\hat{r}_l\hat{r}_R}{1-R},$$

$$p_3 = p(M_1Q_2m_1) = \frac{(1-\hat{r}_l)\hat{r}_R}{R},$$

$$p_4 = p(m_1Q_2m_2) = \frac{\hat{r}_l(1-\hat{r}_R)}{R},$$

for marker haplotypes $h_j$ ; $j = 1-4$, respectively.

The likelihood of the phenotypic value $y$ conditional on marker haplotype $h_j$ under the IM model is then written as

$$\hat{L}_i = P(y|h_j, \hat{\theta}) = P(Q|h_i) \times P(y|Q, \hat{\theta}) + P(q|h_i) \times P(y|q, \hat{\theta}) = \frac{1}{\sqrt{2\pi \hat{\sigma}^2}} \left[ p_1 \times \exp \left( \frac{(y - \hat{\mu} - 1/2\hat{\alpha}^2)^2}{2\hat{\sigma}^2} \right) \right] + (1-p_1)$$

\tag{2}

where $\hat{\theta}$ denotes the vector of the estimates of the unknown parameters $\mu$, $\sigma^2$ and $\alpha$, excluding the estimate of the QTL position $\hat{d}$.

The QTL-cluster mapping (CM) was also considered, in which the transmission of the whole chromosome segment marked, rather than the single QTL, is taken into account. The CM model can be formulated as

$$y_i = \mu + DR_i + e_i,$$ \tag{3}

with $y_i$, $\mu$, and $e_i$ as defined in the IM model (1), where $D$ represents the substitution effect for the marked segment, and $R_i$ is given as $1/2[E(\hat{L}_i|h_j) - E(\hat{L}_m|h_j)]$ with $I_p$ and $I_m$ representing the proportions of IBD for the marked segment between the paternal and maternal gametes of the sire and the gamete for the offspring that originated...
from the sire, respectively. Then, applying the procedure of Guo (1994a, b, 1995) and using Haldane’s (1919)
mapping function, the conditional expectations $E(I_p \mid h_j)$ and $E(I_m \mid h_j)$ can be expressed as

$$E(I_p \mid h_j) = \frac{1}{T} \int \left[ \frac{r_{js}(t)w_{js}(l-t)}{r_{js}(l)} \right] dt$$

with the transition probabilities at the distance $t$ (Morgans) apart,

$$r_{js}(t) = \frac{1}{2} \left( e^{-2st} + e^{-2et} \right)$$

where $k$ is the variable indicating the paternal and maternal gametes of the sire by 0 and 1, respectively, and $s$ and $e$ are the variables each taking 0 or 1 that represent the origin of the allele from the sire to the offspring at the flanking marker loci with $t = 0$ and $t = l$, respectively. For example, $s = 0$ and $e = 0$ indicate offspring’s haplotype $M_1M_2$.

For the current CM method, the likelihood of the phenotypic value $y$ conditional on marker haplotype $h_j$ is given as

$$\hat{L}_j = P(y \mid h_j, \hat{\theta}) = P(y \mid E(I_p \mid h_j), E(I_m \mid h_j), \hat{\theta})$$

$$= \frac{1}{\sqrt{2\pi \hat{\sigma}^2}} \exp \left( -\frac{(y - \hat{\mu} - 1/2[E(I_p \mid h_j) - E(I_m \mid h_j)]\hat{\sigma}^2}{2\hat{\sigma}^2} \right)$$

We here note that this CM likelihood is similar to the likelihood of the regression interval mapping (RIM; Kao, 2000), except for the assumption of the genetic model within the marker interval. Therefore, the CM method is equivalent to the RIM method, when the conditional probabilities $p_j$ and $1-p_j$ are used instead of the conditional expectations $E(I_p \mid h_j)$ and $E(I_m \mid h_j)$ in equation (6). In this study, the expected log-likelihood of the RIM model was also evaluated.

**True genetic models considered** For the true genetic model, we considered the situations where one marker interval contains a single QTL or closely linked QTLs. For the latter case, five equally-spaced QTLs, each of which has the same effect, were considered, and three types of linkage phase for the sire were included: coupling, or $(+ + + + + / - - - - -)$, repulsion, or $(+ - + - + / - + + - -)$, and a specific phase in which the effects of the left two QTLs and the right three QTLs were in opposite direction, or $(- - + + + / + + - - -)$. Average effect of the gene-substitution at the marked segment level for QTLs varied in some ways.

The likelihood according to the number of QTLs, the type of linkage phase, and the individual effects of QTLs is written as

$$L_j = \frac{1}{\sqrt{2\pi \sigma^2}} \sum_{k=1}^{q} P(QH_k \mid h_j) \times P(y \mid QH_k, \theta) \times \exp \left( -\frac{(y - \mu - A_k)^2}{2\sigma^2} \right)$$

where $\theta$ is the vector of the true parameters, $QH_k$ represents the $k$th QTL haplotype, and $A_k$ is the effect for $QH_k$.

**Deterministic sampling and simulations** Deterministic sampling (Mackinnon and Weller, 1995; Mackinnon et al., 1996; Wright and Kong, 1997; Ronin et al., 1999) was used to evaluate the expected log-likelihoods for the IM, RIM and CM procedures under the true genetic models. The data were deterministically simulated by systematically drawing values of a random variable. Then the number of observations with the value $y$ at the fixed point in the $j$th marker haplotype group is $n_p(h_j)L_j$, since this is equal to the height of the density function at the point $y$ multiplied by the size of the marker haplotype group. Hence, the log$L_j$ over all the observations with the value $y$ is $n_p(h_j)L_j \log L_j$. The expected total log-likelihood of the distribution of observations was obtained by summing over all values of $y$, as follows:

$$E(\log L) = \sum_{j=1}^{4} E \left[ \log \prod \hat{L}_j \right]$$

$$= \sum_{j=1}^{4} \sum_{i=1}^{n_j} \log \hat{L}_j$$

$$= \sum_{j=1}^{4} n_j \int L \log \hat{L}, dy,$$

where $n_j$ is the number of offspring having marker haplotype $j$, $L_j$ is the likelihood conditional on marker haplotype $j$ under the true genetic model, and $\hat{L}_j$ is the likelihood conditional on marker haplotype $j$ under the (R)IM or CM model.

In the deterministic simulation, the size of marker interval was set to 20 cM and the family size to 200. The parameters $\mu$ and $\sigma$ were supposed to be known, setting to 0 and 1, respectively. Then, for the (R)IM methods, the unknown parameters were the position of, and the substitution effect for, a putative QTL. We will show the three-dimensional plots of $\log L$ as a function of the two unknown parameters and evaluate the shapes of the likelihood surfaces for the considered cases. For CM, it was allowed that the unknown parameter was only the substitution effect for QTL(s) at the segment level, and the
log\(L\) profiles as a function of this unknown parameter were investigated. The likelihood for estimating the value(s) of the unknown parameter(s) was calculated for the range of \(\pm 0.1\) of the true (or expected) value with the increment of 0.01. The numerical integration in equation (8) was conducted for the range of \(\pm 7\sigma\), using the subroutine QTRAP (Press et al., 1986).

**RESULTS**

The log-likelihood surface expected for the IM procedure is shown in Fig. 1 for the true situation where a single QTL with the substitution effect of 0.5 is located in the middle of the marker interval of 20 cM. Note that in this case, the assumption of the IM procedure corresponds well to the true genetic model. As would be expected, the log-likelihood was maximized when the estimate of the QTL position was 10 cM (from the origin) and the estimated substitution effect of the QTL was 0.25, or when the estimates were equal to the true values. It is consequently shown that the IM model is successfully fitted to the true genetic model and is quite valid. For the true setting where the marker interval contains the five equally-spaced QTLs that are in coupling phase and each of which has the same effect with the substitution effect of 0.5 at the segment level, the plot of log\(L\) expected for IM is similar to that for the case of the single QTL as shown in Fig. 1. That is, there was observed a clear unimodal global maximum with the estimated location of 10 cM for the putative QTL. Under this true genetic model, the estimation of the QTL effect at the segment level was expected to be accurate, giving that \(\hat{\alpha}/2 = 0.25\) (data not shown).

Fig. 2 concerns the true situation where the five equally-spaced QTLs of equal effects are in repulsion phase with the resulting substitution effect of 0.1 at the segment level. For this setting, with the IM approach, there was found a ridge along the marker interval for the likelihood surface. This pattern is obviously different from those in Fig. 1 and the coupling cases. Since the ridge was found to be parallel with the axis of QTL position, it is indicated that there was the independency between the estimates of the QTL position and the QTL effect. For the QTL effect, the estimate was expected to be unbiased with \(\hat{\alpha}/2 = 0.05\). For the estimation of QTL position, although a slight peak of the likelihood emerged in the center of the marker interval, because of the presence of the ridge, all the positions within the interval could practically be an almost equally possible estimate of the location of the putative QTL. Such an undesirable phenomenon would be due to the fact that the actual substitution effect at the segment level was expected to be smaller in spite that the effect of each QTL was equal to that in the coupling case considered above. With a larger substitution effect of 0.5 at the segment level, the pattern of the likelihood surface for the IM procedure was rather similar to those of the single QTL and the coupling cases as shown in Fig. 1 (results not shown).

The case of a particular linkage phase for the five QTLs in the marker bracket, or the case of \((- - + + +)\), was also examined. With the IM procedure, as depicted in Fig. 3, a ridge was found with the substitution effect of 0.1 for the whole segment, similarly as in Fig. 2. However, the direction of the ridge was not parallel with \(\hat{\alpha}\), and there was found a tendency that the estimate of the QTL effect became higher, as the estimated QTL position became a righter-hand side of the interval, indicating a
Comparison of interval mapping procedures

Fig. 3. The profile of the log-likelihood expected with the estimates of QTL-position and -effect by IM for the case where the five equally-spaced QTLs with the substitution effect of 0.1 at the segment level within the marker interval are in a specific phase, or (−−++ +).

dependent relationship between the two estimates. For the present case, it was expected that the position of the putative QTL was estimated as 20 cM, or the position of the right-side one of the flanking markers, with the apparent effect of \( \hat{\alpha}/2 = 0.06 \), whereas the truly expected effect for the whole segment was 0.05. When the segment-substitution effect was a higher value of 0.5, a more upward bias in the estimated effect was expected, showing that \( \hat{\alpha}/2 = 0.29 \) (Fig. 4). For the RIM method, the maximum values of the log-likelihood were found to be only very slightly to slightly different from the corresponding values for IM.

Fig. 4. The profile of the log-likelihood expected with the estimates of QTL-position and -effect by IM for the case where the five equally-spaced QTLs with the substitution effect of 0.5 at the segment level within the marker interval are in a specific phase, or (−−++ +).

Fig. 5 shows the results for the CM procedure for the selected case where the five equally-spaced QTLs with the substitution effect of 0.5 at the segment level within the marker interval are in a specific phase (−−++ +), together with those for (R)IM in which the QTL position was fixed at the point with the maximum value of the log-likelihood or the right flanking marker. For the true genetic models considered, except for the cases of the particular linkage phase (−−++ +), it was expected that all the three procedures yield the unbiased estimate of the segment-effect. For the cases of (−−++ +), it was revealed that the CM method is expected to give the unbiased estimate of the segment-effect, whereas the (R)IM procedures provide biased estimates.

DISCUSSION

Usually, the (R)IM analyses require the assumption of a single QTL residing within one marker interval. However, it was suggested that multiple linked QTLs or a cluster of QTLs may exist in a particular chromosomal region (e.g., Khavkin and Coe, 1997, 1998). Then the single QTL model may not be adequate in such situations. On this point, the current results demonstrate that the parameter estimates by the (R)IM analyses may possibly be biased seriously in certain situations. It has been found that when the (R)IM analyses resulted in quite biased estimation of the QTLs-effect, the putative QTL may be mapped at the position of one of the flanking marker loci. This would be due to the fact that the search for QTL in this study was restricted within the marker interval.

In contrast, it has been suggested that the CM method could be totally robust to estimate the QTLs-effect for the true genetic models of various types. The likelihood
function for CM is quite similar to that for the RIM model shown by Kao (2000), except that the current CM assumes the genetic model within the marker interval. Kao (2000) suggested that RIM tends to give estimates with larger mean-squared errors and smaller likelihood-ratio test statistics, and therefore it is less powerful in QTL detection, when compared with the likelihood-based IM methods. Kao (2000) also found that such differences between the IM and RIM methods become larger, as the proportion of the variance explained by the QTL becomes higher, hence the marker interval becomes wider, and the QTL position moves more from the boundary to the middle of the interval. In the current study, it was found that the maximum log-likelihood is expected to be slightly larger for IM than for RIM and CM (results not shown), agreeing with the finding of Kao (2000). However, the difference in the maximum value between these methods would be not large, as long as the effect of the QTLs is not extremely large. While Kao (2000) used the operational model that exactly described the true situation where a single QTL exists within the marker interval, for the true situation where the marker interval contains multiple linked QTLs, we here found that the influences of using an inappropriate operational model on the QTL mapping are similar for IM and RIM. That is, in such cases, both the procedures would similarly give biased estimates of the parameters. It appears that these biases in the estimates are caused by use of the single QTL model against the existence of multiple linked QTLs in the marker interval rather than by the methodological difference between the maximum likelihood and the regression procedures. The patterns of the likelihood surfaces examined in this study indicate that there exists the dependency between the estimates of the position of, and the effect for, the putative QTL. Thus, it is quite likely that the occurrence of the bias mentioned above is associated with such dependency. Whittaker et al. (1996) stated that the estimated QTL position is represented by the ‘center of gravity’ of loci within the interval that affect the trait. This is an interesting statement, and our results may suggest a possibility that a region not involving the QTLs becomes the ‘center of gravity’.

The mapping studies under the genetic situation in which multiple linked QTLs contribute to the genetic variance showed that it is difficult to distinguish between the effect due to the single major QTL and the polygenic effects (Visscher and Haley, 1996; Liu and Dekkers, 1998). If the aim is for marker-assisted selection (MAS), the mapping with high resolution is obviously important. For this point, a mapping method using genetic algorithm was proposed recently by Nakamichi et al. (2001). Furthermore, fine-mapping was examined by Meuwissen and Goddard (2000). The investigation on discriminating between modes of inheritance for a single QTL or a cluster of QTLs is also important for detected genetic variation (Visscher and Haley, 1996; Liu and Dekkers, 1998).

We should finally comment that this study was for a single half-sib family and the specific region of the genome with fixed QTL(s)-effect. For whole genome studies in outbred populations with complex pedigrees, further investigation would be necessary to assess in details the validities of the QTL interval mapping (Grignola et al., 1996, 1997; van Arendonk et al., 1998; George et al., 2000) and the QTL-cluster mapping using the mixed model methodology.

REFERENCES


Comparison of interval mapping procedures


