Mapping Multiple Interacting Quantitative Trait Loci with Multidimensional Genome Searches

by

Malgorzata Bogdan
Purdue University

R.W. Doerge
Purdue University

Technical Report #04-03

Department of Statistics
Purdue University

2004

Mapping multiple interacting quantitative trait loci by multidimensional genome searches

Małgorzata Bogdan [1] & R. W. Doerge[2,3,4]

[1] Institute of Mathematics, Wrocław University of Technology, Wrocław, Poland

[2] Department of Statistics, Purdue University, West Lafayette, IN 47907, USA

[3]Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA

[4]Computational Genomics, Purdue University, West Lafayette, IN 47907, USA

Correspondence should be addressed to:

R.W. Doerge
Department of Statistics
1399 Mathematical Sciences Building
Purdue University
West Lafayette, IN 47907-1399

Phone: (765) 494-6030
Fax: (765) 494-0558
Email: doerge@purdue.edu

ABSTRACT

The popular methods for mapping quantitative trait loci (QTL), like interval mapping or composite interval mapping, do not allow to locate QTL which do not have additive effects and influence the trait only by interacting with other genes. The direct solution to the problem of locating epistatic QTL is to use multiple regression model with interactions and search for several QTL simultaneously using multidimensional version of interval mapping. The utility of the multiple interval mapping is however limited by two interconnected issues. First of them is the difficult task of deciding how many terms should be included in the corresponding regression model. The second is the computational complexity of the search over the great space of possible multidimensional models. In this paper we report the results of the simulation study in which we demonstrate that multiple interval mapping can appropriately locate linked and epistatic QTL and address important statistical and computational issues related to its application. In particular we test the popular model selection criteria designed for multiple regression and show that in the context of QTL mapping they have a tendency to overestimate the number of QTL and their interactions. We explain this phenomenon and argue that an easy modification of Schwarz Bayesian Information Criterion can handle the problem of the overfitting. We also give a brief review of different random search procedures which can be used to speed up the search over the entire model space and suggest that these developments coupled with improved computing power have the potential to solve main problems related to the application of multiple interval mapping.

keywords: Multiple Interval Mapping, Epistasis, Model Selection Criteria

INTRODUCTION

Locating genes which influence quantitative traits (QTL) has a long history and still remains one of the most important tasks of statistical genetics. The earliest methods for QTL location (see e.g. Soller et al. (1976)) used standard statistical procedures like t-test or ANOVA to look for association between marker genotypes and trait values. These methods have a relatively low power of detection if the QTL is located in the middle of the interval between two distant markers and do not allow to separate the QTL effect and its location. To make it possible to locate QTL which lies inside the interval between two markers Lander and Botstein (1989) introduced Interval Mapping (IM). This procedure employs EM algorithm (Dempster et al.,1977) to compute likelihood ratio tests for the presence of QTL on a dense grid of possible QTL locations. Though IM is based on a single QTL model it is usually used in genome scans aiming to locate all relevant QTL. Presently interval mapping is a widely accepted procedure for QTL location and has become the basis for further developments. The extensions of IM went in many directions. To avoid using EM algorithm an approximation to IM by a simple regression scheme was proposed by Hayley and Knott (1992) and Martinez and Curnow (1992). Another extensions, composite interval mapping (CIM, Zeng 1993, 1994) and multiple QTL mapping (MQM, Jansen 1993) improve the accuracy of locating multiple QTL by including some markers' genotypes as additional predictors in regression. A review of these classical methods for detecting and locating QTL can be found e.g. in Doerge et al. (1997) or Churchill and Doerge (1997).

While the classical methods of QTL mapping cope very well with the problem of locating QTL with additive effects they are usually not able to detect those QTL which influence the trait only by interacting with other genes. The problem of detecting such epistatic QTL was recently addressed by Jannink and Jansen (2001) and Boer at al. (2002), who proposed two new methods based on one dimensional genome searches. In particular in Jannink and Jansen (2001) epistatic QTL are located by identifying loci of high QTL by genetic background interaction. This method requires large populations derived from multiple related inbred - line crosses and inherits characteristic to MQM problems related to the choice of marker cofactors. In turn Boer et al. (2002) address the problem of the choice of cofactors. They propose to include all available markers in the regression equation and use a Bayesian approach to penalize large values of the corresponding regression coefficients.

One of shortcomings of this method is that it requires the initial choice of "the effective dimension" for epistatic interactions, which has a strong influence on the power of detection.

In this paper we point at another direction in the search for an efficient method to locate multiple interacting QTL.

The direct solution to the problem of mapping epistatic QTL is to use multiple regression model with interactions and search for several QTL simultaneously. The corresponding multidimensional version of interval mapping (MIM) was proposed by Kao et al. (1999). The utility of MIM is however limited by two interconnected issues. First of them is the need to develop a statistical tool for deciding how many terms (additive effects and epistasis) should be included in the fitted model. The second is the computational complexity of the search over the great space of possible multidimensional models.

In this paper we report results of the simulation study in which we demonstrate that, contrary to standard one dimensional interval mapping, its multidimensional version can appropriately locate linked and epistatic QTL. We also address important statistical and computational issues related to the application of MIM. In particular we consider the problem of the estimation of QTL number. We test the popular model selection criteria designed for multiple regression and show that they have a tendency to overestimate the number of QTL and their interactions. We explain this phenomenon and argue that an easy modification of Schwarz Bayesian Information Criterion can handle the problem of the overfitting. We also briefly mention some of recent developments in Bayesian variable selection and discrete optimization methods and suggest that these procedures, coupled with improved computing power, have the potential to solve the main problems reducing the applicability of MIM.

METHODS

Advantages of MIM: To illustrate the difficulties which standard one dimensional interval mapping has with detecting linked and epistatic QTL and show the superiority of using appropriate multiple regression model we performed some computer simulations. All the programs used in this paper were written in *Matlab* and were run on the machine IBM R5/600 F50 in the Department of Statistics of Purdue University.

We simulated samples of 200 individuals from backcross population for three different genetic models. For each individual we used Haldane's model of no interference to simulate genotypes of 11 markers spaced every 10 cM on one 100 cM chromosome. We also simulated genotypes of two QTL located at 24 cM and 56 cM from one end of the chromosome.

To show problems with separating linked QTL we simulated the trait data Y_i according to the linear model

$$Y_i = \mu + a_1 X_{1i} + a_2 X_{2i} + \xi_i \ \ , \ i = 1, \ \ldots \ 200 \ \ , \qquad (1)$$

where X_{li} is the genotype of l-th QTL for i-th individual ($X_{li} = \frac{1}{2}$ for homozygote and $-\frac{1}{2}$ for heterozygote) and $\xi_i \sim$N$(0, \sigma^2)$ is an environmental noise.

For the first example we used parameters from Example 13 in Piepho and Gauch (2001): $\mu = 0$, $a_1 = 1.5$, $a_2 = 1.25$ and $\sigma = 1$, which yield heritability (percent of the variation in trait explained by QTL) $h^2 = 59\%$.

At first we analyzed our simulated data with a standard interval mapping based on a folowing single QTL model:

$$Y_i = \mu + a X_i + \xi_i \ \ . \qquad (2)$$

Here Y_i denotes the trait value for i-th individual, X_i is the genotype of the putative QTL and $\xi_i \sim$N$(0, \sigma^2)$ is the random error. For each of the potential locations of QTL, which we spaced every 1 cM, we used the genotypes of flanking markers to assign conditional probabilities to possible QTL genotypes and fitted the model (2) by invoking EM algorithm of Jansen and Stam (1994). The resulting maximum likelihood estimators of the parameters in the model (2) are denoted by $\hat{\mu}$, $\hat{a}$, and $\hat{\sigma}$. Then at each location we computed the lod score statistic;

$$Lod = \log_{10} \frac{L(\hat{\mu}, \hat{a}, \hat{\sigma})}{L(\tilde{\mu}, 0, \tilde{\sigma})} \ \ ,$$

where $L(\hat{\mu}, \hat{a}, \hat{\sigma})$ is the likelihood of the data under the model (2) with parameters $\hat{\mu}$, $\hat{a}$, $\hat{\sigma}$ and $L(\tilde{\mu}, 0, \tilde{\sigma})$ is the likelihood of the data under the null hypothesis of no QTL ($a = 0$ and $\tilde{\mu}$ and $\tilde{\sigma}$ are maximum likelihood estimators of the mean and standard deviation computed upon the assumption that the trait data are normally distributed). Large values of the lod score statistic suggest that a QTL is in the neighborhood of the considered position. Though there exist many theoretical and empirical methods to choose the suitable threshold (for references see e.g. Doerge and Rebaï, 1996) in practical applications the presence of QTL is usually acknowledged if the lod score statistic exceeds value 3 somewhere on the genome.

HERE FIGURE 1

In the first graph in Figure 1 we present the plot of the lod score statistic resulting from standard interval mapping as a function of the location of the putative QTL. We can observe that the maximum of the lod score function exceeds 37 (which strongly suggests QTL presence) and is obtained at the position of the stronger QTL (24 cM). Then the lod score function slowly decreases and the second QTL (56 cM) is not distinguishable. Thus in this example the standard one dimensional interval mapping appropriately locates the stronger QTL and does not detect the weaker one.

In the second graph in Figure 1 we present the results of the analysis of the same data by the two dimensional interval mapping based on the appropriate model (1). At each point on the net of locations of two putative QTL, which we spaced every 1 cM in both directions, we computed the lod score statistic according to the formula

$$Lod = \log_{10} \frac{L(\hat{\mu}, \hat{a}_1, \hat{a}_2, \hat{\sigma})}{L(\tilde{\mu}, 0, 0, \tilde{\sigma})} \ ,$$

where $\hat{\mu}$, $\hat{a}_1$, $\hat{a}_2$, $\hat{\sigma}$ are maximum likelihood estimators of the parameters of the model (1) computed by EM algorithm. In Figure 1 we can observe that the lod score function resulting from the two dimensional interval mapping obtains its maximum at the point (24 cM, 61 cM), which is very close to the true location of simulated QTL. Thus the two dimensional interval mapping performs better than its one dimensional version and allows for appropriate location of both QTL.

For the second example we simulated the trait data according to the model (1) with parameters $\mu = 0$, $a_1 = 1.25$, $a_2 = 1.25$ and $\sigma = 1$, which yield heritability $h^2 = 54.4\%$. Note that in this model both QTL have the same genetic effects.

HERE FIGURE 2

In Figure 2 we present the plots of lod score functions resulting from analysis of our data with one and two dimensional interval mapping. Results of the one dimensional interval mapping demonstrate a typical example of the so called "ghost" effect. The maximum of the lod score statistic is obtained at the point 35 cM, which is in between the locations of true QTL (24 cM, 56 cM). There is a weak local maximum at 27 cM (close to the location of the first QTL) and no local maximum close to 56 cM. Thus the one dimensional interval wrongly suggests the presence of QTL at 35 cM and does not detect the second QTL.

The second graph in Figure 2 shows that the two dimensional interval mapping allows for precise localization of both QTL. The corresponding lod score statistic obtains a maximum at the point (25 cM, 54 cM), which is very close to (24 cM, 56 cM) - the true location of QTL.

To illustrate the problems with detecting epistatic QTL for the third example we simulated the trait data according to the model

$$Y_i = \mu + a_1 X_{1i} + a_2 X_{2i} + a_3 X_{1i} X_{2i} + \xi_i \ \ , i = 1, \ \ldots \ 200 \ . \tag{3}$$

The term $a_3 X_{1i} X_{2i}$ corresponds to the epistasis. We used parameters $\mu = 0$, $a_1 = 0$, $a_2 = 0$, $a_3 = 4$, $\sigma = 1$. Thus our simulated QTL have no additive effects and influence the trait only by the epistatic (interaction) term. The percent of the variation of the trait explained by QTL is 41.9 %.

HERE FIGURE 3

In Figure 3 we present results of the analysis of our simulated data with one and two dimensional interval mapping. The two dimensional interval mapping is based on the correct model (3) and the lod score statistic is computed according to the formula

$$Lod = \log_{10} \frac{L(\hat{\mu}, \hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{\sigma})}{L(\tilde{\mu}, 0, 0, 0, \tilde{\sigma})} \ .$$

In Figure 3 we observe that the lod score statistic resulting from one dimensional interval mapping doesn't exceed the value 0.5, which is definitely below the threshold to detect the QTL presence. The presence of epistatic QTL is however well detected by the two dimensional interval mapping. The maximum of the corresponding lod score statistic is equal to 27.57 and is obtained at (26cM, 55cM), which is again very close to the true location - (24 cM, 56 cM).

Search over markers: While, as shown above, MIM can help in detecting linked and epistatic QTL its utility is restricted by its computational complexity.

For example we consider the problem of locating two QTL on a 100 cM chromosome. In this case the two dimesional interval mapping with the step of 1 cM requires fitting the model (1) or (3) $\binom{101}{2}$=5050 times. Moreover each time we fit the model we have to invoke EM algorithm, which on its own requires some iterations. The complexity of the problem drastically increases with the size of the part of the genome we want to search. For example two dimensional interval mapping with the step of 1cM over the genome of the length 2000 cM requires fitting the regression model approximately $\binom{2000}{2}$=1 999 000 times.

To speed up the search process some authors proposed to use random search methods, like e.g. genetic algorithms (see e.g. Carlborg, Andersson and Kinghorn (2000) and Nakamichi, Ukai and Kishino (2001)). These methods are usually much quicker than the extensive multidimensional search but they do not guarantee to find the optimal solution.

Another way to reduce the complexity of MIM relies on identifying interesting genome regions by using the initial search based on a coarser grid of locations. In the Bayesian setting this approach was suggested e.g. by Sen and Churchill (2001), who use the initial scan based on a 10 cM pseudomarker grid. In the situation when there exists an accurate genetic map (distance between markers $\leq$ 15 cM) it is natural to base this initial search on the net of marker positions, which in many cases helps to avoid using EM algorithm. The search over markers was succesfully used to locate multiple QTL e.g by Ball (2001), Broman and Speed (2002) and Bogdan et al. (2003).

Model selection criteria: Restricting the search for QTL to marker positions reduces the problem of QTL mapping to the problem of the choice of the best multiple regression model, with marker genotypes being the regressor variables. An important and difficult part in fitting such a regression model is the estimation of the number of additive and interaction terms which should be included. As shown above, if we make a mistake at this point we may obtain wrong QTL locations or skip some important QTL.

The standard way of deciding how many additive and interacting (QTL) terms should appear in the model relies on using many consecutive statistical tests (see e.g. Kao et al. (1999)). A disadvantage of this approach is that it allows the comparison of only nested models. It is also unclear how to adjust the significance threshold for each test to provide the full control over the type I error.

An alternative way to decide on the number QTL and their interactions is to use one of many statistical criteria for the choice of the optimal regression model. This approach was succesfully used by Broman and Speed (2002), Piepho and Gauch (2001), Nakamichi et al. (2001) and Bogdan et al. (2003).

In particular Broman and Speed (2002) and Piepho and Gauch (2001) recommend using Schwarz Bayesian Information Criterion (BIC, Schwarz (1978)) to estimate the number of QTL with additive effects. BIC, which is used in many different applications in statistics, recommends choosing the model M, which maximizes the expression

$$S_M = \log L_M - \frac{1}{2} k \log n \quad , \tag{4}$$

where L_M is the likelihood of the data for the model M, k is the number of parameters in the model and n is the sample size. BIC belongs to the wide class of the so called penalized maximum likelihood methods and the second term in this criterion, $\frac{1}{2}k \log n$, is called the penalty for the complexity of the model. In the context of linear regression maximizing S is equivalent to minimizing

$$BIC = n \log\left(\frac{RSS}{n}\right) + k \log n \quad , \tag{5}$$

where RSS is the residual sum of squares from regression.

Broman and Speed (2002) use BIC to choose markers which are strongly associated with the trait. They report that in this context the original BIC works poorly and has a tendency to overestimate QTL number. Therefore they propose to increase the penalty for the number of parameters and consider the criterion

$$BIC_\delta = n \log\left(\frac{RSS}{n}\right) + \delta k \log n \quad ,$$

where δ depends on the length of the genome and the sample size via the formula

$$\delta = \frac{2L_{95}}{\log_{10} n} \quad .$$

Here L_{95} is the 95th percentile of the maximum LOD-score of interval mapping or the ANOVA at marker loci, genome wide, under the hypothesis that there are no QTL. In the examples investigated by Broman and Speed (2002) the values of δ are in the interval between 1.5 and 3.

Let us observe that since

$$\delta k \log n = \frac{2L_{95}}{\log_{10} n} k \log n = 2kL_{95} \log 10 \quad ,$$

the penalty in Broman and Speed criterion does not increase with n. (Actually the values of L_{95} reported in Broman and Speed (2002) decrease with n). Therefore this criterion differs a lot from the original BIC and lacks the property of being consistent (i.e. even for huge sample sizes the probability of chosing the proper model is essentially different from one). Simulations reported in Broman and Speed (2002) show however that their method of locating multiple QTL performs well and can yield better results than the popular composite interval mapping.

Piepho and Gauch (2001) also recommend using BIC to choose the number of QTL. Instead of including individual markers they propose to include pairs of adjacent markers. Thus, at each step they try to add a pair of markers to the model and increase the penalty by the quantity corresponding to two regressors. It turns out that this algorithm works pretty well and does not require changing penalty in BIC. However it is important to notice that the penalty in BIC is inflated in other way. Two adjacent markers genotypes are strongly correlated. For example in the backcross population the correlation between genotypes of markers separated by 10 cM is equal to 0.82. Thus the decrease in RSS we obtain by including these two markers is similar to the decrease in RSS resulting from including only one of them. Despite of that the penalty in Piepho and Gauch algorithm is increased by the quantity corresponding to two independent regressors and thus has good performance.

Nakamichi et al. (2001) propose to use Akaike Information Criterion (AIC, Akaike (1974)) for the choice of the number of QTL. Similarly to BIC, AIC is also used in many applications in statistics. It suggests to choose the model maximizing

$$A = \log L - k \quad , \tag{6}$$

where L is the likelihood of the data under the given model and k is the number of parameters in the model. In the context of the linear regression it is equivalent to choosing the model for which the quantity

$$AIC = n \log\left(\frac{RSS}{n}\right) + 2k \quad , \tag{7}$$

obtains a minimum. Contrary to BIC AIC is not consistent. It is however asymptotically efficient in terms of the Kullback-Leibler distance (i.e. for large sample sizes it chooses the model whose expected predictive power is close to the best possible).

Another criterion for the choice of the best multiple regression model was proposed by Bhansali and Downham (1977). This criterion is a modification of Akaike's (1973) final prediction error criterion and choses the model minimizing the quantity

$$FPE4 = \frac{RSS(n+3k)}{n-k} \quad .$$

Simulations reported in Piepho and Gauch (2001) show that Bhansali and Downham (1977) criterion works relatively well when used for the choice of the pairs of markers flanking QTL.

Below we present results of the simulation study in which we used original versions of AIC, BIC and FPE4 to choose single markers associated with the trait. Contrary to Broman and Speed (2002), Piepho and Gauch (2001) and Nakamichi et al. (2001), who used model selection criteria to estimate the number of QTL with additive effects, we consider also models with interactions.

To illustrate some basic problems related to the application of the model selection criteria we restrict the attention to the models including no more than two QTL. Thus our task is to choose the best model of the form

$$Y_i = \mu + \gamma_1\beta_1 X_{ji} + \gamma_2\beta_2 X_{ki} + \gamma_3\beta_3 X_{ji}X_{ki} + \xi_i \quad , \tag{8}$$

where X_{li} is the genotype of l-th marker for i-th individual ($X_{li} = \frac{1}{2}$ for homozygote and $-\frac{1}{2}$ for heterozygote) and $\xi_i \sim$N$(0, \sigma^2)$ is an environmental noise. γ_1, γ_2, γ_3 can take on values 0 or 1 and are indicators specifying which terms appear in the model.

In the general practise of statistics one usually doesn't consider regression models in which interaction terms appear without the related additive terms. However it is known that in some situations genes without additive effects may influence the trait via the epistasis (see e.g. Fijneman et al. 1996, 1998). Therefore we decided to consider also the models in which epistatic terms appeared without the related additive terms. The description of the six possible models, which need to be checked for each pair of markers, is given in Table 1.

HERE TABLE 1

RESULTS

In Tables 2,3 and 4 we report the results of the analysis of some simulated data sets with the model selection criteria. These results are based on 100 replicates consisting of 200 individuals from backcross population.

For the first experiment we simulated the distribution of marker and QTL genotypes on a short chromosome of length 40 cM, with 5 markers spaced every 10 cM. Trait values were simulated according to three models specified in Table 1: model 1 (no QTL), model 2 (one QTL) and model 4 (only epistatic effect). For model 1 we used parameters $\mu = 0$ and $\sigma = \sqrt{0.563}$. For model 2 we simulated the QTL halfway between the first and the second marker and used parameters $\mu = 0$, $\beta_1 = 0.7705$, $\sigma = 1.1557$, which yield heritability $h^2 = 0.1$. For model 4 we simulated QTL halfway between the first and the second marker and between the fourth and the fifth marker. We used parameters $\mu = 0$, $\beta_3 = 1.8125$, and $\sigma = 1.1557$ for which the broad sense heritability is equal to 0.097.

For each replicate we searched over all possible $\binom{5}{2}$=10 pairs of markers and used model selection criteria to choose the best model of the form (8). In Table 2 we give the percentage of replicates for which specific models were chosen. In the column labeled Correct we give the percentage of replicates for which both the model chosen as well as locations of detected QTL were correct. In this table we classify the location to be correct if one of the markers closest to the QTL was chosen (marker 1 or 2 for the first QTL and 4 or 5 for the second QTL). Model 4 is classified to be correct if both QTL are properly located.

HERE TABLE 2

In Table 2 we observe that when the search is done over a short chromosome with only 5 markers the BIC criterion performs very well. In all our examples it choses the true model in at least 83% of cases. Bhansali and Downham (1977) criterion performs decisively worse but still it can find the appropriate model in over 64% of cases. AIC criterion performs much worse. In all considered examples it has a strong tendency to overestimate the number of terms in the regression model. Some insight into this behavior of AIC can be obtained by comparing the penalties for the number of parameters in BIC and AIC. The constants by the number of parameters in AIC and BIC penalties are equal to 2 and $\log 200 \approx 5.3$ respectively (see (7) and (5)). Thus the penalty for the complexity of the model in AIC is much smaller than in BIC and AIC more easily chooses models with many regressors.

To get results reported in Tables 3 and 4 we simulated the distribution of marker and QTL genotypes on 10 chromosomes of the length 100 cM. On each chromosome we simulated 11 markers, located every 10 cM. The trait data were generated according to models 1,2 and 4, as specified in Table 1. For model 2 the QTL was located on the first chromosome, halfway between the first and the second marker. For model 4 both QTL were located on the first chromosome. The first QTL was located halfway between the first and second marker and the second QTL was at the position of the 10th marker. For models 1 and 2 we used the same parameters as we used to generate data for Table 2. For model 4 we used the parameters: $\mu = 0$, $\beta_3 = 1.541$ and $\sigma = 1.1557$, which yield heritability 0.097.

In Tables 3 and 4 we give the percentage of replicates for which specific models were chosen when the search was done over the first chromosome and all ten chromosomes respectively. In these tables we classify the location to be correct if the marker chosen is in the neigborhood ± 15cM of the true QTL (markers 1,2 and 3 for the first QTL and markers 9,10 and 11 for the second QTL).

HERE TABLES 3 AND 4

The results reported in Table 3 confirm that in the context of QTL mapping BIC performs better than other model selection criteria. However even this criterion has a strong tendency to overestimate the number of QTL when the search is done over entire genome.

Results reported in Tables 2, 3 and 4 show that the tendency to overestimate QTL number by all model selection criteria rapidly increases when we increase the part of the genome we are searching. For example the model 1 (no QTL) is appropriately detected by BIC in 85 % of cases when we consider only 5 markers, in 53% of cases when we search over 1 chromosome and in 0% of cases when we search over entire genome. Also the probability of appropriately detecting the model 2 (one QTL) by BIC falls from 83% when we consider five markers, to 73% when we search over one chromosome and to 0% when we search over entire genome. Instead of models 1 and 2, more complicated models 3, 4 and 5 are usually chosen.

To explain this property of model selection criteria we observe that the number of available models involving two markers increases with the size of the part of the genome we are searching much quicker than the number of models involving one marker. For example when we search over one chromosome we need to consider 11 models with one marker involved and $4\binom{11}{2}$=220 models with two markers involved. Thus the number of possible models involving two markers is twenty times larger than the number of models involving one marker. When we search over ten chromosomes the numbers of models involving one and two markers grow to 110 and $4\binom{110}{2}$=23980 respectively. Thus in this case the number of available models involving two markers is 218 times larger than the number of models involving one marker. The reported difference in proportions between the number of models involving one and two markers explains why the models with two markers involved are more easily chosen just by a random chance when the search is done over the entire genome.

For a better understanding of the phenomenon of the overestimation we will briefly analyze results reported in Table 4. In the first example we observe that when there are no QTL (model 1) BIC and FPE4 usually wrongly choose models involving one epistatic term (model 4). To explain this fact let us observe that models with only one epistatic term comprise almost one fourth of the set of models under consideration. Let us also observe that the penalty we pay for including one epistatic term is the smallest possible and is equal to the penalty we pay for including one additive term. In the result the models with one epistatic term are very often chosen just by a chance even when in reality there are no QTL.

When there is only one QTL (model 2) all model selection criteria usually appropriately detect the corresponding additive term. But they also have a tendency to add to the model some additional terms. Instead of the correct model 2 they usually choose models 3 (two additve terms) and 5 (one additive term and epistasis), with approximately equal frequencies. To explain this phenomenon we observe that the numbers of available models with two additive terms (model 3) and with one additive term and epistasis (model 5) are equal and they together comprise almost one half of the set of models under consideration. Also models of these two forms differ from the true model 2 by just one term in the regression equation. Therefore the chances that the criterion will be minimized by the model 3 or 5 are large and approximately equal.

The only model which is well detected by considered model selection criteria is the model with one epistatic term. We explain this fact by reminding that the models with one epistatic term comprise almost one fourth of the set of possible models with up to two markers involved. Therefore the probability that an alternative model, with different number of components, will minimize one of the considered criteria just by a chance is relatively small. We however expect that the high probability of appropriately classifying a model with one epistatic term will rapidly decrease if we enlarge the class of competing models by including models with three markers involved (there are $\binom{110}{3} = 215820$ possible models with three additive terms).

To solve the reported overestimation problem some authors modified standard model selection criteria by increasing the penalty for the growing dimension. This solution was chosen e.g. by Broman and Speed (2002). Also Nakamichi et al. (2001) increase the penalty in Akaike criterion by treating a location of a QTL as an extra parameter in the model. Methods of Broman and Speed (2002) and Nakamichi et al. (2001) were tested in the situation when the class of competing models consisted of models with only additive terms. They however do not generalize directly to the models with interactions, which require a special treatment. (Our simulations suggest that to balance the big difference in numbers between epistatic and additive terms the penalty for the interaction should be larger than the penalty for the additive term.)

In Bogdan et al. (2003) the phenomenon of the overestimation of the number of QTL by BIC is explained by recalling Bayesian origin of this criterion. It is observed that while approximating the Bayesian rule for the choice of the optimal model BIC neglects the prior probabilities of different models, which corresponds to assigning the same prior probability to all considered models. While in many applications this approach is well justified, in the context of QTL mapping it lends itself to assigning unrealistically high prior probabilities to the events where many regressors are involved (e.g., when 200 markers are available, the number of different models involving 100 additive terms is equal to $\binom{200}{100} \approx 9.05 * 10^{58}$ and the prior probability of the event that there are 100 regressors involved is over 10^{56} times larger than the prior probability of the event that there is just one regressor). This explains the phenomenon of the overestimation and suggests that the properties of BIC can be improved by supplementing it with a more realistic prior distribution π on the set of possible models and maximizing the expression

$$\tilde{S}_M = S_M + \log \pi(M) \ , \tag{9}$$

where S_M is given by (4).

The idea of incorporating the prior information into BIC was already suggested by Ball (2001) who uses $e^{\tilde{S}_M}$ to approximate the posterior probability of the model M. Ball (2001) uses his method to search through the class of models with additive terms and estimate locations of QTL with additive effects. Bogdan et al. (2003) extend the approach and propose the modification of BIC which allows to search through the class of models with interactions. The choice of the prior proposed in Bogdan et al. (2003) implies that the penalty for the model dimension in the resulting criterion increases with the number of available markers and that interactions are penalized more than the additive terms. Extensive computer simulations reported in Bogdan et al. (2003) show that the modified version of BIC deals very well with the problem of the overfitting the model and in many situations allows the location of both additive and epistatic QTL.

DISCUSSION

Examples reported in this paper demonstrate that multiple interval mapping (MIM), based on the appropriate regression model, allows for precise localization of multiple interacting QTL. The main difficulty in using MIM relies however on the choice of the appropriate model. We showed that the popular statistical model selection criteria have a tendency to choose models with too many QTL and argued that this problem can be solved by using Schwarz Bayesian Information Criterion supplemented with a realistic prior distribution on the number of terms in the fitted regression model.

The idea of using Bayesian methodology for QTL mapping is not new. In a series of papers Satagopan et al. (1996), Satagopan and Yandell (1996), Heath (1997), Uimari and Hoeschele (1997), Stephens and Fisch (1998), Silanpää and Arjas (1998) and Yi and Xu (2000) use the full Bayesian approach and Markov Chain Monte Carlo simulations to estimate posterior distributions of QTL locations and other parameters in the regression model. This approach requires multiple generations of all parameters and is very much computationally demanding. Moreover, as noted by Ball (2001), " a major challenge remains to obtain a rapidly converging sampler for the full Bayesian model". A simplified version of Bayesian approach to the problem of QTL mapping was proposed by Berry (1998) and Sen and Churchill (2001). In particular Berry (1998) obtains the approximation to the posterior distribution of QTL locations by representing it as a step function, using a first order approximation to the likelihood and integrating out parameters of the regression model. This makes it possible to construct a Gibbs sampler over the net of possible locations, obviating generations of other parameters. Sen and Churchill (2001) avoid using MCMC and employ independent sample Monte Carlo approach to generate multiple versions of pseudomarker genotypes on the dense grid of genomic locations. In the next step they compute weights for each pseudomarker realization by integrating out parameters of the related regression models and use them to approximate the posterior distribution of the QTL locations. The methods of Ball (2001), Broman and Speed (2002) and Bogdan et al. (2003), who restrict the search to marker positions and use modified versions of BIC, can be seen as further simplifications of standard Bayesian methodology, which are particularly useful when we search over a large space of possible models with interactions.

However, even when we use the simplified approach and restrict the search to markers positions, the exhaustive search over a huge space of possible multiple regression models is unfeasible. To solve this problem Broman (1997) and Bogdan et al. (2003) use the forward selection procedure to construct the nested sequence of “interesting” models. The results of simulations reported in

Broman (1997), Broman and Speed (2002) and Bogdan et al. (2003) show that though this simple procedure works very well, in some cases it has a tendency to include extraneous variables into the model. Broman and Speed (2002) report that when the search is done over models with additive terms this can be improved by using a random search procedure based on an efficient MCMC sampler proposed in Smith (1996). The method used by Broman and Speed (2002) is one of many available random search methods which can be applied to traverse the space of possible multiple regression models. In the Bayesian setting such procedures were proposed e.g. by Raftery et al. (1997, Sec. 4.2), who also construct a suitable MCMC sampler or Berger and Molina (2002). On the contrary to the full Bayesian approach these methods directly search through the model space, obviating the generation of regression parameters. There exist also many discrete optimization procedures which can be used to search for the best regression model. Among the most promising are genetic algorithms (see e.g. Goldberg, 1989), simulated annealing (Kirkpatrick et al., 1983 , for an application for multiple regression see Brown et al., 1999), tabu search (Glover, 1989a,b) or ant colony optimization (see e.g. Dorigo and Di Caro, 1999). We believe that further research on the application of these methods in the context of QTL mapping will result in numerically feasible and accurate procedures for the estimation of QTL number and simultaneous location of multiple interacting QTL.

Acknowledgment We would like to thank Professor Jayanta K. Ghosh for helpful remarks on BIC.

References

[1] Akaike H. (1973) Information theory and an extension of the maximum likelihood principle pp.267–281 in 2nd International Symposium on Information Theory, edited by B. N. Petriv and F. Csaki, Akademia Kiado, Budapest.

[2] Akaike H. (1974) A new look at the statistical model identification. IEEE Trans. Automat. Control AC-19:716–723.

[3] Ball R. (2001) Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian Information Criterion. Genetics 159:1351–1364.

[4] Berger J. O. and Molina G. (2002) Discussion on *A case study in model selection* by K. Viele, R. Kass, M. Tarr, M. Behrmann and I. Gauthier. *Case Studies in Bayesian Statistics* (A. Carriquiri, et al., Eds.) Springer - Verlag, New York.

[5] Berry C. (1998) Computationally efficient Bayesian QTL mapping in experimental crosses. Joint Statistical Meetings Proceedings of the Biometrics Section 164–169.

[6] Bhansali R. J. and Downham D. Y. (1977) Some properties of the order of an autoregressive model selected by a generalized Akaike's EPF criterion. Biometrika 64:547–551.

[7] Boer M. P., ter Braak C. J. F., and Jansen R .C. (2002) A penalized likelihood method for mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics 162:951–960.

[8] Bogdan M., Ghosh J. K. and Doerge R. W. (2003) Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci. Submitted for publication.

[9] Broman K. W. (1997) Identifying quantitative trait loci in experimental crosses. PhD Dissertation. Department of Statistics, University of California, Berkeley.

[10] Broman K. W. and Speed T. P. (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. J Roy Stat Soc B 64:641 - 656.

[11] Brown P. J., Fearn T. and Vanucci M. (1999) The choice of variables in multivariate regression: A non-conjugate Bayesian decision theory approach. Biometrika 86: 635–648.

[12] Carlborg Ö., Andersson L. and Kinghorn B. (2000) The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. Genetics 155: 2003–2010.

[13] Churchill G. A. and Doerge R. W. (1997) Mapping quantitative trait loci in experimental populations. In : Molecular Dissection of Complex Traits. A. H. Paterson (ed), CRC Press: New York.

[14] Dempster A. P., Laird N.M. and Rubin D. B. (1977) Maximum likelihood from incomplete data via EM algorithm. J. Roy. Statist. Soc. B 39:1–38.

[15] Doerge R. W., Rebaï A. (1996), Significance thresholds for QTL interval mapping tests, *Heredity* **76** 459–464.

[16] Doerge, R. W., Zeng Z-B. and Weir B. S. (1997) Statistical issues in the search for genes affecting quantitative traits in experimental populations . Statistical Science 12:195–219.

[17] Dorigo M. and Di Caro G. (1999) The ant colony optimization meta-heuristic. In : New Ideas in Optimization. Corne D., Dorigo M. and Glover F. (ed). McGraw-Hill.

[18] Fijneman R. J. A., de Vries S. S., Jansen R. C. and Demant P. (1996) Complex interactions of new quantitative trait loci, *Sluc1, Sluc2, Sluc3,* and *Sluc4*, that influence the susceptibility to lung cancer in the mouse. Nat. Gen. 14:465–467.

[19] Fijneman R. J. A., Jansen R. C., Van der Valk M. A. and Demant P. (1998) High frequency of interactions between lung cancer susceptibility genes in the mouse: mapping of Sluc5 to Sluc14. Cancer Res. 58:4794–4798.

[20] Gelfand A. E. and Smith A.F.M. (1990) Sampling-based approaches to calculating marginal densities. J. Am. Statist. Assoc. 85:398–409.

[21] Glover F. (1989a) Tabu search - Part1. ORSA Journal on Computing 1:190–206.

[22] Glover F. (1989b) Tabu search - Part2. ORSA Journal on Computing 2:4–32.

[23] Goldberg D. E. (1989) Genetic algorithms in search, optimization and machine learning. Addison Wesley.

[24] Hayley, C.S. and Knott S.A. (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324.

[25] Heath S. C. (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. 61:748–760.

[26] Jannink J.-L. and Jansen R. (2001) Mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics 157: 445–454.

[27] Jansen R. C. (1993) Interval mapping of multiple quantitative trait loci. Genetics 135: 205–211.

[28] Jansen R. C. and Stam P. (1994) High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136:1447–1455.

[29] Kao C-H., Zeng Z-B., Teasdale R. D. (1999) Multiple interval mapping for quantitative trait loci. Genetics 152:1203–1216.

[30] Kirkpatrick S., Gelatt C. D. and Vecchi M. P. (1983) Optimization by simulated annealing. Science 220:671–680.

[31] Lander, E. S. and Botstein, D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199.

[32] Martinez, O. and Curnow, R. N. (1992) Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. 85: 480–488.

[33] Nakamichi R., Ukai Y. and Kishino H. (2001) Detection of closely linked multiple quantitative trait loci using genetic algorithm. Genetics 158:463–475.

[34] Piepho H.-P. and Gauch H. G. Jr. (2001) Marker pair selection for mapping quantitative trait loci. Genetics 157:433–444.

[35] Raftery A. E., Madigan D. and Hoeting J. A (1997) Bayesian model averaging for linear regression models. J. Am. Statist. Assoc. 92:179–191.

[36] Satagopan J. M. and Yandell B. S. (1996) Estimating the number of quantitative trait loci via Bayesian model determination. Special Contributed Paper Session on Genetic Analysis of Quantitative Traits and Complex Diseases, Biometric Section, Joint Statistical Meetings, Chicago IL.

[37] Satagopan J. M., Yandell B. S., Newton M. A. and Osborn T. C. (1996) Bayesian model determination for quantitative trait loci. Genetics 144: 805–816.

[38] Schwarz, G. (1978) Estimating the dimension of a model. Ann. Stat. 6: 461–464.

[39] Sen S. and Churchill G. A. (2001) A statistical framework for quantitative trait mapping. Genetics 159:371–387.

[40] Silanpää M. J. and Arjas E. (1998) Bayessian mapping of multiple quantitative trait locus from incomplete inbred line cross data. Genetics 148:1373–1388.

[41] Smith M. S. (1996) Nonparametric regression: a Markov chain Monte Carlo approach. PhD dissertation. University of New South Wales, Sydney.

[42] Soller M., Brody T. and Genizi A. (1976) On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor Appl Genet 47:35–39.

[43] Stephens D. A. and Fisch R. D. (1998) Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. Biometrics 54:1334–1347.

[44] Uimari P. and Hoeschele I. (1997) Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. Genetics 146:735–743.

[45] Yi N. and Xu S. (2000) Bayesian mapping of quantitative trait loci for complex binary traits. Genetics 155:1391–1403.

[46] Zeng, Z.-B. (1993) Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA 90:10972–10976.

[47] Zeng, Z.-B. (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468.

Table 1: Specification of models with up to two QTL involved. Indicators γ_i are as in the definition of the model (8).

id	model	γ_1	γ_2	γ_3
1	no QTL	0	0	0
2	1 QTL	1	0	0
3	2 QTL with only additive effects	1	1	0
4	only epistatic term	0	0	1
5	1 additive effect + epistatic term	1	0	1
6	2 additive effects + epistatic term	1	1	1

Table 2: Percentage of replicates for which specific models were chosen when the search was performed over one 40 cM chromosome. In the column labeled 'Correct' we give the percentage of replicates for which both the model and QTL locations were correctly identified. 'id' is the model identificator as specified in Table 1.

		Chosen model id						
True model id	Criteria	1	2	3	4	5	6	Correct
1	BIC	85	2	1	12	0	0	85
	AIC	25	6	14	44	9	2	25
	FPE4	67	8	6	19	0	0	67
2	BIC	1	89	6	0	4	0	83
	AIC	1	31	25	0	36	7	31
	FPE4	1	68	15	0	16	0	64
4	BIC	2	0	0	95	3	0	90
	AIC	0	0	0	71	23	6	67
	FPE4	1	0	0	91	7	1	85

3

Table 3: Percentage of replicates for which specific models were chosen when the search was performed over one 100 cM chromosome. In the column labeled 'Correct' we give the percentage of replicates for which both the model and QTL locations were correctly identified. 'id' is the model identificator as specified in Table 1.

		Chosen model id						
True model id	Criteria	1	2	3	4	5	6	Correct
1	BIC	53	3	3	40	1	0	53
	AIC	1	2	18	53	22	4	1
	FPE4	26	5	7	57	5	0	26
2	BIC	0	73	15	0	12	0	73
	AIC	0	10	40	0	37	13	10
	FPE4	0	58	20	0	20	2	58
4	BIC	0	0	0	98	2	0	91
	AIC	0	0	0	71	25	4	67
	FPE4	0	0	0	90	9	1	85

Table 4: Percentage of replicates for which specific models were chosen when the search was performed over ten 100 cM chromosomes. In the column labeled 'Correct' we give the percentage of replicates for which both the model and QTL locations were correctly identified. 'id' is the model identificator as specified in Table 1.

		Chosen model id						
True model id	Criteria	1	2	3	4	5	6	Correct
1	BIC	0	0	3	80	17	0	0
	AIC	0	0	8	40	42	10	0
	FPE4	0	0	5	75	17	3	0
2	BIC	0	0	48	10	42	0	0
	AIC	0	0	36	0	34	30	0
	FPE4	0	0	46	4	48	2	0
4	BIC	0	0	1	94	5	0	68
	AIC	0	0	3	54	34	9	42
	FPE4	0	0	3	79	17	1	60

Figure 1: Plots of lod score functions resulting from one and two dimensional interval mapping for the first set of simulated data. Two QTL are located at 24 cM and 56 cM. The trait data were generated according to the model (1) with parameters $\mu = 0$, $a_1 = 1.5$, $a_2 = 1.25$ and $\sigma = 1$. The maxima of the lod score functions are obtained at 24 cM and (24 cM, 61 cM) correspondingly.

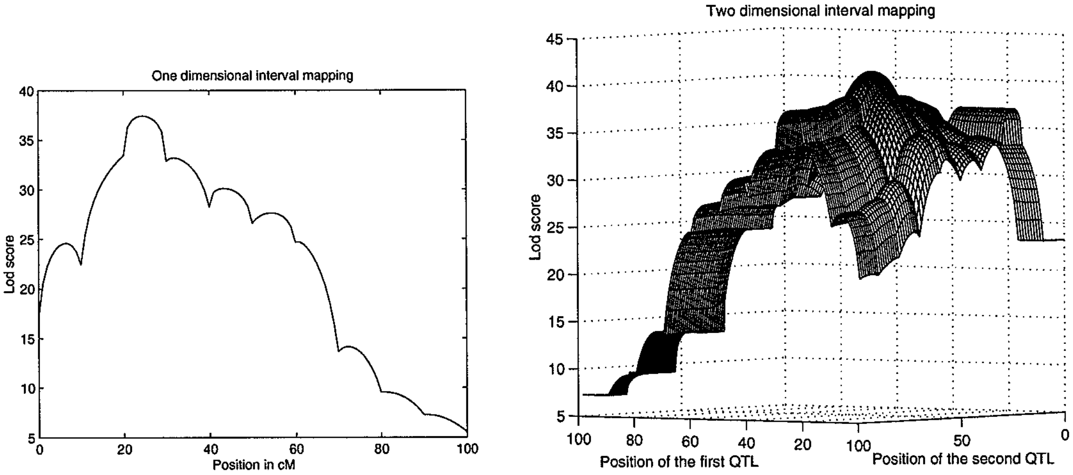


Figure 2: Plots of lod score functions resulting from one and two dimensional interval mapping for the second set of simulated data. Two QTL are located at 24 cM and 56 cM. The trait data were generated according to the model (1) with parameters $\mu = 0$, $a_1 = 1.25$, $a_2 = 1.25$ and $\sigma = 1$. The maxima of the lod score functions are obtained at 35 cM and (25 cM, 54 cM) correspondingly.

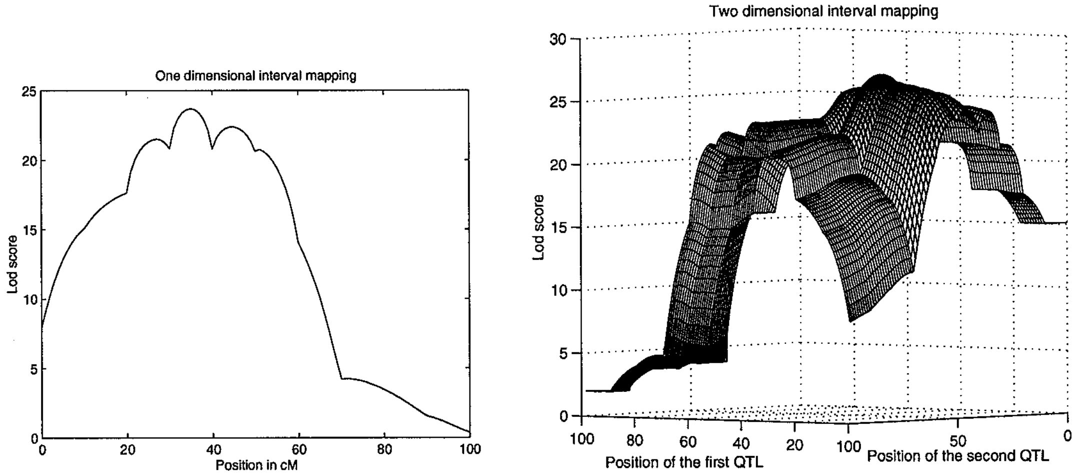
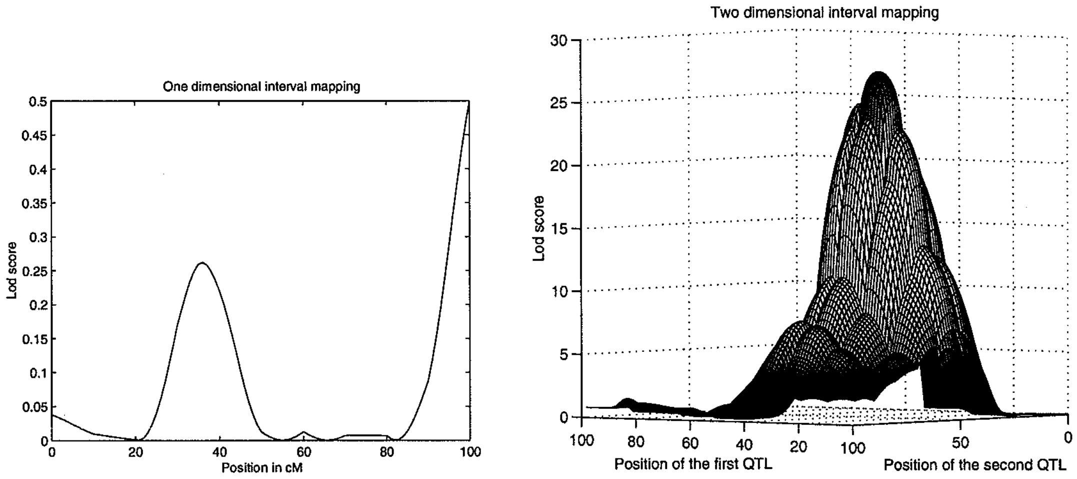


Figure 3: Plots of lod score functions resulting from one and two dimensional interval mapping for the third set of simulated data. Two QTL are located at 24 cM and 56 cM. The trait data were generated according to the model (3) with parameters $\mu = 0$, $a_1 = 0$, $a_2 = 0$, $a_3 = 4$ and $\sigma = 1$. The maximum of the lod score statistic resulting from one dimensional interval mapping is below the threshold to detect the QTL presence. The maximum of the lod score statistic resulting from two dimensional interval mapping is obtained at (26 cM, 55 cM).


One dimensional interval mapping
Lod score
Position in cM
Two dimensional interval mapping
Lod score
Position of the first QTL
Position of the second QTL