World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:4, No:3, 2010

Markov Chain Monte Carlo Model Composition Search Strategy for Quantitative Trait Loci in a Bayesian Hierarchical Model

Susan J. Simmons, Fang Fang, Qijun Fang, and Karl Ricanek

***Abstract*—**Quantitative trait loci (QTL) experiments have yielded important biological and biochemical information necessary for understanding the relationship between genetic markers and quantitative traits. For many years, most QTL algorithms only allowed one observation per genotype. Recently, there has been an increasing demand for QTL algorithms that can accommodate more than one observation per genotypic distribution. The Bayesian hierarchical model is very flexible and can easily incorporate this information into the model. Herein a methodology is presented that uses a Bayesian hierarchical model to capture the complexity of the data. Furthermore, the Markov chain Monte Carlo model composition (MC^3) algorithm is used to search and identify important markers. An extensive simulation study illustrates that the method captures the true QTL, even under nonnormal noise and up to 6 QTL.

***Keywords*—**Bayesian hierarchical model, Markov chain Monte Carlo model composition, quantitative trait loci.

I. Introduction

QUANTITATIVE trait loci (QTL) experiments have yielded important biological and biochemical information necessary for understanding the relationship between genetic markers and quantitative traits [1]. There are many examples where the identification of QTL have made substantial impact in industry such as the genetic markers responsible for weight gain in pigs [2], pecking-related traits in chickens [3], and starch content and composition in maize [4].

Over the last 20 years, there has been an abundance of algorithms proposed for QTL mapping such as interval mapping strategies [5]-[8], composite interval mapping strategies [9]-[12], multiple interval mapping [13]-[20], Bayesian interval mapping [21]-[24], and model selection strategies [25]-[27]. However, most algorithms allow only one observation per genotypic distribution. In situations such as plant QTL experiments where there can be cloned plants, observations within lines are summarized into a single observation. By summarizing the cloned plants into one observation, important information about variation within each line is lost.

In this paper, an MC^3 model selection strategy is proposed in a flexible Bayesian hierarchical model that incorporates replicate information and thereby taking full advantage of variability within a (plant) line. The proposed method is validated through an extensive simulation study involving a variety of QLT scenarios.

S.J. Simmons is with the Department of Mathematics and Statistics at the University of North Carolina Wilmington, Wilmington, NC 28403 USA (phone: 910-962-3296; fax: 910-962-7107; e-mail: simmonssj@ uncw.edu).

F. Fang and Q. Fang were with the Department of Mathematics and Statistics at University of North Carolina Wilmington, Wilmington, NC 28403 USA. They are now with the Department of Mathematics at University of Arizona, Tuscon, AZ 85721 USA (e-mail: fangfang1101@gmail.com and fangqijun33@hotmail.com).

K. Ricanek is with the Department of Computer Science at University of North Carolina Wilmington, Wilmington, NC 28403 USA (e-mail: ricanekk@uncw.edu).

II. Methodology

A. Bayesian Hierarchical Model

The measured trait in the QTL experiment is represented by y_{ij} where $i = 1,\ldots,L$ (L = number of lines) and $j = 1,\ldots, n_i$ (n_i = the number of replicates within line i). We assume that the observed trait follows a normal distribution with mean θ_i and variance σ_i^2 or in other words

$$y_{ij} \,|\theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2)\; i = 1, \ldots, L. \tag{1}$$

Furthermore, the means are assumed to be influenced by the marker information matrix, which is denoted as $\boldsymbol{X}$ where $\boldsymbol{X}$ is M x L and M is the number of markers. The mean of each line is assumed to follow a normal distribution with mean $\boldsymbol{X}_i'\boldsymbol{\beta}$ and variance τ^2 where $\boldsymbol{X}_i'$ is the transpose of i^{th} column of $\boldsymbol{X}$ or

$$\theta_i \,|X, \beta, \tau^2 \sim N(X_i'\beta, \tau^2). \tag{2}$$

Since no prior information is assumed to be known about which markers might be the QTL, we assign a normal distribution with mean 0 and a large variance of 100 on the coefficient for each marker ($\boldsymbol{\beta} \sim N(0, 100)$).

The prior distribution for the variance parameters σ_i^2 and τ^2 are assigned Inverse-$\chi^2(1)$ which has an infinite mean and variance.

These assumptions yield an implicit full posterior distribution; however, the full conditional distributions have a nice parametric form. The full conditional posterior distributions are represented below as $\varphi|\cdot$ for random variable φ given all other quantities.

$$\theta_i|\cdot \sim N\left(\left(\frac{1}{\tau^2}I+\gamma^{-1}\right)^{-1}\left(\frac{1}{\tau^2}X'\beta+\gamma^{-1}Y\right),\left(\frac{1}{\tau^2}I+\gamma^{-1}\right)^{-1}\right)$$

$$\sigma_i^2|\cdot \sim \text{Inv-gamma}\left(\frac{n_i+1}{2},\frac{1}{2}\left(1+\sum_j(y_{ij}-\theta_i)^2\right)\right)$$

$$\beta_j|\cdot \sim N\left(\left(\frac{1}{\tau^2}X'X+\frac{1}{100}\right)^{-1}X'\theta,\left(\frac{1}{100}+\frac{1}{\tau^2}X'X\right)^{-1}\right)$$

$$\tau^2|\cdot \sim \text{Inv-gamma}\left(\frac{L+1}{2},\frac{1}{2}\left(1+\sum_i(\theta_i-X'\beta)^2\right)\right)$$

where,

$$\gamma=\begin{pmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \sigma_2^2 & & 0\\ \vdots & & \ddots & 0\\ 0 & 0 & \cdots & \sigma_L^2\end{pmatrix}.$$

The full conditional distributions are used in the Gibbs sampler to estimate the full posterior distribution, which will be denoted as $p(D|M)$. Due to the highly complex nature of the model, the first 2000 samples are considered as the burn-in period and are discarded. An additional 50,000 samples are generated to estimate the posterior distribution.

B. Stochastic Search

The Markov chain Monte Carlo model composition strategy (MC^3) [28] first randomly generates a model selection vector. The model selection vector is a binary vector of length M. Along the model selection vector, locations with values of 1 indicate that the marker is included in the current model and locations with a value of 0 indicate that the marker is not included in the model. The likelihood of the data, $p(D|M)$, under the initial model selection vector is calculated via the Gibbs sample as discussed in the previous section. A position along the model selection vector is randomly selected, and its value is switched such that if the original value is 1 then it becomes 0 and vice versa. Next the likelihood of the data under this new model is calculated. Denoting the likelihood of the models as $p(D|M^{(i)})$ and $p(D|M^{(j)})$ where i represents the original model and j represents the new model, the transitional probability from model i to model j is defined as:

$$\alpha_{ij}=\min\left(1,\frac{p(D|M^{(j)})}{p(D|M^{(i)})}\right).$$

The transitional probability α_{ij} defines the likelihood of the chain moving to the new model, j. A Bernoulli random variable is generated with probability of success α_{ij}. If the generated Bernoulli random variable is 1, then the new model, j, becomes the current model. If, however, the generated Bernoulli random variable is 0, the current model, i, is maintained as the current model. The algorithm continues by randomly identifying a new position along the current model vector and switching its value. The likelihood of this new model is calculated and a new transitional probability is computed. The Markov chain progresses in this manner until 2,000 models have been visited. A multiplicity of chains is simultaneously generated in the fashion described above to protect against premature convergence caused by a local minimum. This work uses ten parallel chains with randomly generated initial model selection vectors.

Upon completion of all chains, 20,000 models would have been explored and their likelihoods computed. The posterior model probability for the 20,000 models are calculated using Bayes theorem defined as:

$$p(M^{(k)}|D)=\frac{p(D|M^{(k)})p(M^{(k)})}{\sum_k p(D|M^{(k)})p(M^{(k)})}. \quad (3)$$

No prior model information is assumed, so equation (3) simplifies to:

$$p(M^{(k)}|D)=\frac{p(D|M^{(k)})}{\sum_k p(D|M^{(k)})}. \quad (4)$$

Using this quantity, the activation probability for each marker is then calculated as:

$$p(\beta_j|D)=\sum_k p(\beta_j|M^{(k)},D)p(M^{(k)}|D)$$

where $p(\beta_j|M^{(k)},D)$ =1 if marker j is in the model and 0 if marker j is not in the model.

III. Simulations

In order to validate the current methodology, 60 simulations were performed on responses generated from a marker matrix from a *Bay-O x Shahdara* population with 38 markers and 165 lines [29]. The use of this $\boldsymbol{X}$ matrix incorporates the complexity and correlation structure evident in observed QTL experiments. The genetic map of the *Bay-O x Shahdara* is shown in Fig. 1.

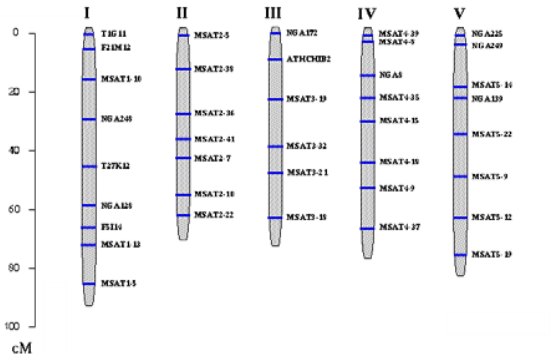

Fig. 1 Genetic map for the Bay-O x Shahdara recombinant inbred line [29]

This $\boldsymbol{X}$ matrix was used to generate quantitative traits by the following formula

$$y_{ij}=\sum_j a_j x_{ij}+\varepsilon_{ij}$$

where a_j is the effect of the j^{th} marker, x_{ij} is the value of the j^{th} marker of the i^{th} line, and ε_{ij} is the random noise. The

algorithm defined in Section II was developed under the assumption of normal noise, but error distributions are generally unknown. In this simulation study, it was of interest to determine if this algorithm was robust to the assumption of normality, and therefore, the error distribution simulated in this work was generated from a gamma distribution.

The gamma distribution is a two-parameter, continuous probability distribution controlled by its shape, α, and scale, ϕ, parameters. We used two different gamma distributions in this study. The first gamma distribution had a shape parameter of 0.5 and a scale parameter of 1, giving an expectation for the variance within each line of 0.5, i.e. a tight variance. The second gamma distribution had a shape parameter of 1 and a scale parameter of 3 giving an expected variance of 3 within each line, i.e. a larger variance for the random noise. To further stress the system, effect sizes, a_j, were selected that ranged from 1 to 9 where an effect size of 1 is very small, effect size of 5 is moderate, and effect size of 9 is large. In order to adequately evaluate the system for complex QTL marker identification the generated data included one QTL up to six QTLs. This simulation study investigated 60 different QTL scenarios.

From the simulation study, the stochastic search identified only the QTL with no false positives for every simulation from 1 to 6 QTL and all effect sizes. Table I illustrates the effectiveness of the proposed approach to find the marker across multiple chromosomes without getting stuck in local troughs of nearby markers.

TABLE I
SUMMARY RESULTS

Ground Truth	Effect Size	Gamma Noise (α,β)	Proposed Results
C1M2	1	0.5,1	C1M2
		1,3	C1M2
C1M2	9	0.5,1	C1M2
		1,3	C1M2
C1M5, C2M15	2,4	0.5,1	C1M5, C2M15
		1,3	C1M5, C2M15
C1M5, C2M15	6,8	0.5,1	C1M5, C2M15
		1,3	C1M5, C2M15
C1M6, C2M15, C3M21	2,4,8	0.5,1	C1M6, C2M15, C3M21
		1,3	C1M6, C2M15, C3M21
C1M2, C1M9, C2M15, C5M31	1,3,5,9	0.5,1	C1M2, C1M9, C2M15, C5M31
		1,3	C1M2, C1M9, C2M15, C5M31
C1M2, C1M5, C1M9, C2M15, C4M27, C5M33	1,2,5,7,8,9	0.5,1	C1M2, C1M5, C1M9, C2M15, C4M27, C5M33
		1,3	C1M2, C1M5, C1M9, C2M15, C4M27, C5M33

A full disclosure of all experimental results is not possible due to the number of experiments conducted and the page constraints. However the interested reader may contact the lead author for the detailed results.

IV. CONCLUSION

The identification of QTL leads to discoveries that can impact society, e.g. identifying drought resistance for a plant or mass marker for cattle that can be used to breed meatier cattle over the use of steroids or bio-feed. Data obtained through QTL experiments are complex and methodologies need to be able to handle the complexities in a robust manner; hence, this research presented an approach based on Bayesian hierarchical model that tolerates complexity well. The Bayesian hierarchical model has shown that it is a flexible model that can incorporate multi-levels of information, e.g. trait values for recombinant plant lines or environmental information or laboratory information, in an experiment. The proposed approach was able to *correctly identify every* QTL for each of the 60 experiments conducted. The algorithm appears to be robust to the assumption of normality, although further investigations might be needed.

Future work will investigate the use of this method to model epistasis among markers and environmental factors.

REFERENCES

[1] Lynch M., Walsh, B., Genetics and Analysis of Quantitative Traits, Sinauer Associates Inc., Sunderland, MA, 1997.
[2] Jing-hu Z., Yuan-zhu X., Bo Z., Ming-gang L., Feng-e L., Jia-lian L.,"Detection of Quantitative Trait Loci Associated with Live Measurement Traits in Pigs", Agricultural Sciences in China 6 (7), 2007, pp. 863-868.
[3] Buitenhuis, A.J., Rodenburg, T.B., Siwek, M., Cornelissen, S.J.B., Nieuwland, M.G.B., Crooijmans, R.P., Groenen, M.A., Koene, P. Bovenhuis, H., van der Poel, J.J., "Quantitative trait loci for behavioural traits in chickens", Livestock Production Science 93 (1), 2005, pp. 95-103.
[4] Séne, M., Causse, M., Damerval, C., Thévenot, C., Prioul, J.L., "Quantitative trait loci affecting amylose, amylopectin and starch content in maize recombinant inbred lines", Plant Physiology and Biochemistry 3 (6), 2000, pp. 459-472.
[5] Lander E.S,. Botstein D.., "Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps", Genetics 121, 1989, pp 185-199.
[6] Luo Z.W., Kearsey M.J., "Interval mapping of quantitative trait loci in an F2 population", Heredity 69, 1992, pp 236¨C242.
[7] Jansen R.C.,"Interval mapping of multiple quantitative trait loci", Genetics 135, 1993, pp. 205-211.
[8] Luo Z.W., Williams J.A. "Estimation of genetic parameters using linkage between a marker gene and a locus underlying a quantitative character in F2 populations", Heredity 70, 1993, pp. 245-253.
[9] Jansen R.C., Stam P., "High resolution of quantitative traits into multiple loci via interval mapping", Genetics 136, 1994, pp. 1447-1455.
[10] Zeng Z.B., "Precision mapping of quantitative trait loci", Genetics 136, 1994, pp. 1457-1468.
[11] Jiang C., Zeng Z.B., "Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines", Genetica 101, 1997, pp. 47-58.
[12] Gao H., Yang R., "Composite interval mapping of QTL for dynamic traits", Chin Sci Bull 51, 2006, pp.: 1857-1862.
[13] Haley C., Knott S., "A simple regression method for mapping quantitative trait loci in line crosses using flanking markers', Heredity 69, 1992,:pp. 315-324.
[14] Jansen R.C. "A general Monte Carlo method for mapping multiple quantitative trait loci", Genetics 142, 1996, pp. 305-311.
[15] Liu J., Mercer J.M., Stam L.F., Gibson G.C., Zeng Z.B., Laurie C.C., "Genetic analysis of a morphological shape difference in the male genitalia of Drosophila simulans and D. mauritiana", Genetics 142, 1996, pp. 1129-1145.

[16] Kao C.H., Zeng Z.B. "General formulae for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm", Biometrics 53, 1997, pp. 653-665.
[17] Weber K., Eisman R., Higgins S., Kuhl L., Patty A., Sparks J., Zeng Z.B. "An analysis of polygenes affecting wing shape on chromosome three in Drosophila melanogaster", Genetics 153, 1999, pp. 773-786.
[18] Kao C.H., "On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci", Genetics 156, 2000, pp. 855-865.
[19] Zeng Z.B., Liu J., Stam L.F., Kao C.H., Mercer J.M., Laurie C.C., "Genetic architecture of a morphological shape difference between two drosophila species", Genetics 154, 2000, pp. 299-310.
[20] Zeng Z.B., Wang T., Zou W., "Modeling quantitative trait loci and interpretation of models", Genetics 169, 2005, pp. 1711-1725.
[21] Satagopan J.M., Yandell B.S., Newton M.A., Osborn T.C., "Markov chain Monte Carlo approach to detect polygene loci for complex traits", Genetics 144, 1996, pp. 805-816.
[22] Sillanpaa M., Arjas E., (1998) "Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data", Genetics 148, 1998, pp. 1373-1388.
[23] Sen S., Churchill G.A.,"A statistical framework for quantitative trait mapping", Genetics 159, 2001, pp. 371-387.
[24] Yandell B.S., Mehta T., Banerjee S., Shriner D., Venkataraman R., Moon J.Y., Neely W.W., Wu H., von Smith R., Yi N., "R/qtlbim: QTL with Bayesian interval mapping in experimental crosses", Bioinformatics 23, 2007, pp. 641-643.
[25] Ball R.D., "Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion", Genetics 159, 2001, pp. 1351-1364.
[26] Broman, K. W., Speed, T. P., "A model selection approach for the identification of quantitative trait loci in experimental crosses", Journal of the Royal Statistical Society, Series B: Statistical Methodology 64, 2002, pp. 641-656
[27] Sillanpaa M.J., Corander J., "Model choice in gene mapping: what and why", Trends in Genetics 18,2002, pp. 301-307.
[28] Boone, E.L., Simmons, S.J., Ye, K., Stapleton, A.E., "Analyzing Quantitative Trait Loci for the Arabidopsis thaliana using Markov Chain Monte Carlo Model Composition with restricted and unrestricted model spaces", Statistical Methodology 3 (1), 2006, pp. 69-78.
[29] Loudet, Chaillou, Camilleri, Bouchez, Vedele, "Bay-0 x Shahdara recombinant inbred lines population: a powerful tool for the genetic dissection of complex traits in Arabidopsis", Theoretical and Applied Genetics 104 (6-7), 2002, pp. 1173-1184.