UMA ABORDAGEM BAYESIANA PARA O MAPEAMENTO DE QTLS UTILIZANDO O MÉTODO MCMC COM SALTOS REVERSÍVEIS

A bayesian approach to map QTLs using reversible jump MCMC

Joseane Padilha da Silva[1], Roseli Aparecida Leandro[2]

RESUMO

A utilização de metodologias bayesianas tem se tornado frequente nas aplicações em Genética, em particular em mapeamento de QTLs usando marcadores moleculares. Mapear um QTL significa identificar sua localização ao longo do genoma, estimar seus efeitos genéticos: aditivo, dominância, epistasia, etc. A abordagem bayesiana permite combinar a verossimilhança dos dados fenotípicos com distribuições *a priori* atribuídas a todas as quantidades desconhecidas no modelo (número, localização no genoma e efeitos genéticos dos QTLs) de forma a fornecer distribuições *a posteriori* a respeito dessas quantidades. Métodos de mapeamento bayesiano podem incorporar a incerteza relativa ao número desconhecido de QTLs na análise; essa incerteza, no entanto, resulta em complicações na obtenção da amostra da distribuição conjunta *a posteriori*, uma vez que a dimensão do espaço do modelo pode variar. O método MCMC com Saltos Reversíveis (MCMC-SR), proposto por Green (1995), é uma excelente ferramenta para explorar a distribuição conjunta *a posteriori* nesse contexto. Neste trabalho, explora-se o método MCMC-SR, utilizando dados artificiais gerados no *software WinQTLCart*, atribuindo-se diferentes prioris para o número de QTLs.

Termos para indexação: Mapeamento de QTLs, *gibbs-sampler*, *metropolis-hastings*, MCMC com saltos reversíveis.

ABSTRACT

The use of Bayesian methodology in genetic applications has grown increasingly popular, in particular in the analysis of quantitative trait loci (QTL) for studies using molecular markers. In such analyses the objectives are mapping QTLs, estimating their locations in the genome and their genotypic effects (additive, dominance, and epistatic). The Bayesian approach proceeds by setting up a likelihood function for the phenotype and assigning prior distributions to all unknown quantities in the model (number, chromosome, locus, and genetic effects of QTL). These induce a posterior distribution of the unknown quantities that contains all of the available information for inference of the genetic architecture of the trait. Bayesian mapping methods can treat the unknown number of QTL as a random variable, which has several advantages but results in the complication of varying the dimension of the model space. The reversible jump MCMC algorithm (MCMC-RJ), proposed by Green (1995), offers a powerful and general approach to exploring posterior distributions in this setting. The method was evaluated by analyzing simulated data in *WinQTLCart*, attributing different priors distributions on the QTL numbers.

Index terms: Mapping QTLs, gibbs-sampler, metropolis-hastings, reversible jump MCMC**.**

(Recebido em 10 de janeiro de 2007 e aprovado em 11 de julho de 2008)

INTRODUÇÃO

A utilização de metodologias estatísticas tem se tornado um componente indispensável em análises de dados nas mais diversas áreas de conhecimento, sendo imprescindível recorrer ao auxílio delas para desenvolver pesquisas confiáveis e tomar decisões importantes. Dentre as diversas áreas que a estatística pode atuar, destaca-se o ramo da Biologia que estuda a herança e a variação de caracteres quantitativos: a Genética Quantitativa.

Caracteres quantitativos são aqueles cuja expressão fenotípica apresenta variações contínuas, atribuídas à segregação simultânea de muitos genes distribuídos pelo genoma, em regiões definidas como QTLs ("Quantitative Trait Loci"). Com o apoio de modelagens estatísticas pode-se realizar o mapeamento de QTLs. Mapear um QTL significa identificar sua posição no genoma e estimar seus efeitos. É importante salientar que um QTL pode apresentar efeitos de diferentes magnitudes.

Na Agronomia, é economicamente relevante localizar e estimar os efeitos genéticos de QTLs responsáveis pela variação genética de caracteres tais como: produção de grãos, altura da planta, teor de proteína, resistência a doenças, ganho de peso, consumo alimentar, produção de leite etc, que resultam da ação cumulativa de um conjunto de genes.

Conforme Paterson et al. (1991), do ponto de vista prático, os resultados obtidos por meio do mapeamento de QTL podem auxiliar os pesquisadores no melhoramento genético de plantas ou animais e/ou responder questões básicas sobre os processos evolutivos.

[1]USP/ESALQ – Programa de Pós-Graduação em Estatística e Experimentação Agronômica. E-mail: joseane_padilha@hotmail.com
[2]USP/ESALQ – Departamento de Ciências Exatas, C.P. 9 – 13418-900 – Piracicaba, SP – Brasil.

Para que se possa realizar um mapeamento de QTLs, é fundamental que se disponha de um conjunto de marcadores genéticos, pois esses auxiliam na construção da distribuição de probabilidades dos genótipos dos QTLs, uma vez que na prática, eles são desconhecidos.

No presente artigo, o mapeamento de QTLs será feito utilizando a inferência bayesiana. A abordagem bayesiana combina a verossimilhança com a distribuição *a priori* atribuídas a todas as quantidades desconhecidas (número, localização e efeitos genéticos dos QTLs), obtendo-se uma distribuição conjunta *a posteriori* para essas quantidades. Métodos bayesianos de mapeamento podem incorporar a incerteza relativa ao número, desconhecido, de QTLs na análise, implicando em vantagens consideráveis para a modelagem. Entretanto, resulta em complicações na obtenção da amostra aleatória da distribuição conjunta *a posteriori*, uma vez que a dimensão do espaço do modelo pode variar. O algoritmo MCMC com Saltos Reversíveis (MCMC-SR) proposto por Green (1995) é uma excelente ferramenta para obtenção dessa amostra, uma vez que ele permite saltar entre modelos de dimensões diferentes por meio da especificação de distribuições propostas eficientes.

O presente artigo visou a: (i) avaliar o desempenho do método MCMC-SR utilizando vários conjuntos de dados artificiais de populações originadas de cruzamentos controlados, supondo-se que a característica de interesse é afetada por uma quantidade desconhecida de QTLs e que não existe interação entre os QTLs, e nem desses com o ambiente; (ii) avaliar a influência da escolha de distribuições *a priori* para o número de QTLs sobre as inferências *a posteriori* dos efeitos genéticos, localizações e número de QTLs que afetam a expressão fenotípica.

A simulação dos mapas de ligações e das populações de cruzamentos foi realizada no programa *WinQTLCart* versão 2.5 e para obtenção da amostra da distribuição conjunta *a posteriori* utilizou-se módulo *BIM* (*Bayesian Interval Mapping*) do *WinQTLCart*.

MATERIAL E MÉTODOS

Métodos

Modelo QTL

Considerando-se uma população originada de cruzamentos controlados, supondo-se que a característica de interesse é afetada por S QTLs e que não existe interação entre os eles, e nem desses com o ambiente, o valor fenotípico observado para o *i*-ésimo indivíduo, y_i, pode ser descrito através do seguinte modelo de regressão linear (Satagopan et al., 1996):

$$y_i = \mu + \sum_{j=1}^{S} a_j Q_{ij} + \sum_{j=1}^{S} d_j (1 - |Q_{ij}|) + \varepsilon_i \quad (1),$$

em que μ é uma constante, $a = [a_j]_{j=1}^{S}$ e $d = [d_j]_{j=1}^{S}$ são vetores de dimensão $S\,x1$ que denotam o efeito aditivo e o efeito de dominância relativos aos S QTLs e $\mathbf{e} = [\varepsilon_i]_{i=1}^{n}$ é o vetor de dimensão $n\,x1$ de erros, tal que $\mathbf{e} \sim N(\mathbf{0}, I\mathbf{s}^2)$. Considerando-se uma população F_2, Satogopan et al. (1996) assumem a seguinte parametrização para o genótipo do QTL Q_{ij}: se o genótipo for (i) QQ, $Q_{ij} = 1$; (ii) Qq, $Q_{ij} = 0$; (iii) qq, $Q_{ij} = -1$.

A expressão (1) pode ser reescrita, matricialmente, como

$$y_i = \mu + a \otimes Q_i + d \otimes (1 - |Q_i|) + \varepsilon_i \quad (2),$$

em que $Q_i = [Q_{ij}]_{j=1}^{s}$ denota o genótipo dos S QTLs associados ao *i*-ésimo indivíduo e $v_1 \otimes v_2$ denota o produto interno entre os vetores v_1 e v_2.

Ocorre que, na prática os genótipos dos QTLs, Q_{ij}, $j = 1, \ldots, S$, $i = 1, \ldots, n$ não são observáveis. No entanto, com base nas informações de marcadores moleculares pode-se encontrar a distribuição de probabilidade de Q_{ij}. É importante observar que Q_{ij} é uma variável categórica.

Dessa forma, na prática observamos o valor fenotípico $y = [y_i]_{i=1}^{n}$ e um conjunto de dados de marcadores genotípicos $M = [m_{ik}]_{i=1,k=1}^{n,\ K}$ em que n é o número de indivíduos e K é o número de marcadores. Admitindo-se:

(i) A existência de uma mapa de ligação , então as localizações dos marcadores em cada cromossomo é conhecida;

(ii) A existência de apenas um QTL entre os marcadores adjacentes (essa é uma suposição razoável se o mapa for denso);

(iii) A concatenação dos grupos de ligação feita pelos marcadores finais e iniciais de cada grupo;

e considerando-se o modelo dado pela expressão (2), o objetivo é fazer inferências sobre os parâmetros $\theta = (\mu, a, d, \sigma^2)$ e sobre as localizações , $\lambda = [\lambda_j]_{j=1}^{S}$, dos S QTLs no genoma.

Adotando-se a abordagem bayesiana para o processo inferencial torna-se necessário:

(i) a construção da função de verossimilhança;

(ii) a atribuição de distribuições *a priori* para as quantidades desconhecidas de interesse.

(i) Construção da função de verossimilhança

Tem-se que a distribuição conjunta de Y_i e Q_i é dada por:

$$f(y_i, Q_i) = f(y_i \mid Q_i) f(Q_i),$$

ou ainda,

$$f(y_i, Q_i) = f(y_i \mid Q_i) f(Q_i \mid \lambda, M).$$

Considerando-se que o genótipo do *j*-ésimo QTL do *i*-ésimo indivíduo é condicionalmente independente de outros genótipos do mesmo indivíduo, tem-se:

$$f(y_i, Q_i) = f(y_i \mid Q_i) \prod_{j=1}^{S} f(Q_{ij} \mid \lambda, M),$$

Sendo assim, a distribuição marginal de Y_i será dada por:

$$f(y_i) = \sum_{q_{ij} \in \{-1, 0, 1\}} f(y_i \mid Q_{ij}) \prod_{j=1}^{S} f(Q_{ij} = q_{ij} \mid \lambda, M),$$

ou seja, a distribuição marginal de Y_i é dada por uma mistura de distribuições, nesse caso, mistura de distribuições normais, visto que:

$$y_i \mid Q_{ij} \sim N\left(\mu + \sum_{j=1}^{s} a_j Q_{ij} + \sum_{j=1}^{s} d_j (1 - \mid Q_{ij} \mid), \sigma^2\right)$$

e

$$\sum_{q_{ij} \in \{-1, 0, 1\}} f(Q_{ij} = q_{ij} \mid \lambda, M) = 1.$$

A distribuição de probabilidade $f(Q_{ij} \mid \lambda, M)$ será obtida utilizando-se a informação dos marcadores flanqueadores: M_k e M_{k+1} $(M_{k,k+1})$. Sabe-se que, as frequências dos genótipos marcadores MM, Mm e mm em uma população F2 são 1/4, 1/2 e 1/4, respectivamente. Cada componente de $f(Q_i \mid \lambda, M)$ é obtido em termos de recombinação entre os marcadores m_k e m_{k+1} e o *j*-ésimo QTL e o *j*-ésimo QTL e a marca m_{k+1}, a Tabela 1 resume essa informação.

Tabela 1 – Probabilidade do genótipo do QTL dada a informação dos marcadores flanqueadores

m_k	m_{k+1}	$f(q_i = -1 \mid m_k, m_{k+1}, \lambda)$	$f(q_i = 0 \mid m_k, m_{k+1}, \lambda)$	$f(q_i = 1 \mid m_k, m_{k+1}, \lambda)$
-1	-1	$\frac{n_1 n_2}{(1-R)^2}$	$\frac{2t_1 t_2}{(1-R)^2}$	$\frac{2d_1 d_2}{(1-R)^2}$
-1	0	$\frac{n_1 t_2}{R(1-R)}$	$\frac{t_1(n_2 + d_2)}{R(1-R)}$	$\frac{d_1 t_2}{R(1-R)}$
-1	1	$\frac{n_1 d_2}{R^2}$	$\frac{2t_1 t_2}{R^2}$	$\frac{d_1 n_2}{R^2}$
0	-1	$\frac{t_1 n_2}{R(1-R)}$	$\frac{(n_1 + d_1) t_2}{R(1-R)}$	$\frac{t_1 d_2}{R(1-R)}$
0	0	$\frac{2t_1 t_2}{R^2 + (1-R)^2}$	$\frac{(n_1 + d_1)(n_2 + d_2)}{R^2 + (1-R)^2}$	$\frac{2t_1 t_2}{R^2 + (1-R)^2}$
0	1	$\frac{t_1 d_2}{R(1-R)}$	$\frac{(n_1 + d_1) t_2}{R(1-R)}$	$\frac{t_1 n_2}{R(1-R)}$
1	-1	$\frac{d_1 n_2}{R^2}$	$\frac{2t_1 t_2}{R^2}$	$\frac{n_1 d_2}{R^2}$
1	0	$\frac{d_1 t_2}{R(1-R)}$	$\frac{t_1(n_2 + d_2)}{R(1-R)}$	$\frac{n_1 t_2}{R(1-R)}$
1	1	$\frac{d_1 d_2}{(1-R)^2}$	$\frac{2t_1 t_2}{(1-R)^2}$	$\frac{n_1 n_2}{(1-R)^2}$

em que, $n_i = \left(1 - r_i\right)^2$, $t_i = (1 - r_i)$ *e* $d_i = r_i^2$ e r_1 e r_2 são as freqüências de recombinação entre o QTL e as marcas m_k e m_{k+1}.

Portanto, a função de verossimilhança dos dados fenotípicos pode ser expressa como:

$$L(\theta,\lambda \mid y)=\prod_{i=1}^{s}\sum_{q_i} f\left(y_i \mid Q_i=q_i\right) f\left(Q_i \mid \lambda, M\right).$$

Adotando-se a abordagem bayesiana, as inferências são baseadas na distribuição conjunta *a posteriori* das quantidades não observáveis dadas as observáveis, $\pi(\theta,\lambda,Q \mid y,M)$, em que $Q=[Q_{ij}]_{i,k=1}^{n,K}$. **A** distribuição conjunta *a posteriori* $\pi(\theta,\lambda,Q \mid y,M)$ é obtida combinando-se a verossimilhança, $L(\theta,\lambda \mid y)$ com as distribuições *a priori* para os parâmetros, $\pi(\theta,\lambda)$, isto é,

$$\pi(\theta,\lambda,Q \mid y,M) \propto L(\theta,\lambda \mid y)\pi(\theta,\lambda),$$

ou equivalentemente,

$$\pi(\mu,a,d,\sigma^2,\lambda,Q \mid y,M) \propto L(\mu,a,d,\sigma^2,\lambda \mid y)\pi(\mu,a,d,\sigma^2,\lambda)$$

Assumindo-se independência entre os efeitos genéticos e demais parâmetros do modelo, a expressão anterior pode ser reescrita como:

$$\pi(\mu,a,d,\sigma^2,\lambda,Q \mid y,M) \propto L(\mu,a,d,\sigma^2,\lambda \mid y)\pi(\mu)\pi(\lambda)\prod_{i=1}^{n}(a_i)\prod_{i=1}^{n}(d_i)\pi(\sigma^2),$$

em que $\pi(\mu)\prod_{i=1}^{n}(a_i)\prod_{i=1}^{n}(d_i)\pi(\sigma^2)\pi(\lambda)$ são as densidades *a priori* para os parâmetros do modelo.

(ii) Atribuição de distribuições *a priori* para as quantidades desconhecidas de interesse

Segundo Gaffney (2001), as distribuições a priori para as quantidades desconhecidas do modelo podem ser especificadas conforme Tabela 2.

Considerando-se que os S QTLs estão ordenados, ou seja,

$$\lambda_1 < \lambda_2 < \ldots < \lambda_S,$$

e, atribuindo-se distribuição *a priori* Uniforme ao longo do comprimento do genoma e somando uma constante *db* para garantir apenas um QTL por intervalo, isto é, assumindo-se que:

$$\begin{aligned}
&\lambda_1 \sim U \ [0, \text{comprimento do genoma}]\\
&\lambda_2 \sim U \ [\lambda_1 + db, \text{comprimento do genoma}]\\
&\vdots\\
&\lambda_s \sim U \ [\lambda_S - 1 + db, \text{comprimento do genoma}],
\end{aligned}$$

a distribuição conjunta *a posteriori* é dada por:

$$\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{1}{2\sigma^2}\left[y_i-\left(\mu+\sum_{j=1}^{s}a_jQ_{ij}+\sum d_j(1-|Q_{ij}|)\right)\right]^2\right\}*$$

$$*\prod_{i=1}^{n}\prod_{j=1}^{s}P(Q_{ij} \mid \lambda_j, M, Q)*$$

$$\frac{1}{\sqrt{2\pi\tau^2}}\exp\left\{-\frac{1}{2\tau^2}(\mu-\eta)^2\right\}\frac{1}{comp.gen}\frac{1}{comp.gen-(\lambda_1+db)}\cdots$$

$$\frac{1}{comp.gen-(\lambda_{s-1}+db)}*$$

$$\prod_{j=1}^{s}\frac{1}{\sqrt{2\pi\sigma_a^2}}\exp\left\{-\frac{1}{2\sigma_a^2}(a_j-\upsilon_a)^2\right\}\prod_{j=1}^{s}\frac{1}{\sqrt{2\pi\sigma_b^2}}$$

$$\exp\left\{-\frac{1}{2\sigma_d^2}(d_j-\upsilon_d)^2\right\}*\frac{b^a}{(a)}(\sigma^2)^{-(a+s)}\exp\left(-\frac{b}{\sigma^2}\right). \quad (3)$$

Tabela 2 – Especificações das distribuições *a priori* para os parâmetros do modelo.

Parâmetro	Distribuição *a priori*
μ	Normal com média = η e variância = τ^2
Localização - λ	• Uniforme, ao longo do comprimento do genoma. • Poisson ou • Geométrica
Efeito Aditivo – a_j	Normal com média = ν_a e variância = σ_a^2
Efeito Dominância – d_j	Normal com média = ν_d e variância = σ_d^2
Variância - σ^2	Gama inversa com hiperparâmetros b,c.

Na abordagem bayesiana, as inferências sobre as quantidades de interesse são baseadas em suas distribuições marginais *a posteriori*, as quais podem ser obtidas integrando-se a distribuição conjunta *a posteriori* (3). As distribuições marginais *a posteriori* para os parâmetros do modelo nesse caso não podem ser obtidas analiticamente, sendo assim, métodos MCMC, tais como Gibbs sampler e Metropolis-Hastings, serão utilizados na construção de amostras da distribuição conjunta *a posteriori*, que, por sua vez, sob certas condições de regularidade (Casella & George, 1992) fornecerá uma amostra da distribuição marginal *a posteriori* de cada quantidade de interesse de interesse.

Distribuições condicionais completas *a posteriori*

Para a implementação dos algoritmos Gibbs sampler e Metropolis-Hastings é necessário a obtenção das distribuições condicionais completas *a posteriori* para as quantidades de interesse.

Distribuição condicional completa *a posteriori* para μ

$$\mu \mid \lambda, Q, a, d, \sigma^2, y \sim N\left(\frac{\sum^{n}\left(y_i - \sum^{s} a_j Q_{ij} - \sum^{s} d_j (1-\mid Q_{ij} \mid)\right)}{\frac{\sigma^2}{\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)}} + \frac{n}{\tau^2}; \frac{1}{\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)} \right)$$

Distribuição condicional completa *a posteriori* para σ^2.

$$\sigma^2 \mid \lambda, Q, a, d, \mu, y \sim IG\left(b+\frac{n}{2}\ ;\ \frac{\sum^{n}\left(y_i - \mu - \sum^{s} a_j Q_{ij} - \sum^{s} d_j (1-\mid Q_{ij} \mid)\right)^2 + c}{2} \right)$$

Distribuição condicional completa *a posteriori* para $\boldsymbol{a_j}$ *e* $\boldsymbol{d_j}$.

$$a_{j^*} \mid \lambda, Q, \{a_j\}_{j\neq j^*}, d_j, \sigma^2 y \sim N\left(\frac{\sum_{i=1}^{n} Q_{ij^*}\left(y_i - \mu - \sum_{j\neq j^*}^{s} a_j Q_{ij} - \sum_{j=1}^{s} d_j (1-\mid Q_{ij} \mid)\right) - \frac{v_a}{\sigma_a}}{\left(\frac{\sum_{i=1}^{n} Q_{ij^*}^2}{\sigma^2}+\frac{1}{\sigma_a^2}\right)}; \frac{1}{\left(\frac{\sum_{i=1}^{n} Q_{ij^*}^2}{\sigma^2}+\frac{1}{\sigma_a^2}\right)} \right)$$

$$d_{j^*} \mid \lambda, \mu, , \{d_j\}_{j\neq j^*}, \sigma^2, y, Q \sim N(\text{média}, \text{variância}),$$

em que,

$$\text{média} = \left(\frac{\sum_{i=1}^{n} (1-\mid Q_{ij^*} \mid)\left(y_i - \mu - \sum^{s} a_j - \sum_{j\neq j^*}^{s} d_j (1-\mid Q_{ij^*} \mid)\right) + \frac{v_d}{\sigma_d^2}}{\left(\frac{\sum_{i=1}^{n} (1-\mid Q_{ij^*} \mid)^2}{\sigma^2}+\frac{1}{\sigma_d^2}\right)} \right) \quad \text{e}$$

$$\text{variância} = \frac{1}{\frac{\sum_{i=1}^{n} (1-\mid Q_{ij^*} \mid)^2}{\sigma^2}+\frac{1}{\sigma_d^2}}$$

Uma vez obtidas as distribuições condicionais completas *a posteriori* constrói-se sob certas condições de regularidade, uma amostra da distribuição conjunta *a posteriori* conjunta, $\Pi(\theta, \boldsymbol{\lambda}, Q \mid y)$, Cabe ressaltar que, no caso das distribuições condicionais completas a *posteriori* possuírem forma conhecida utiliza-se o método MCMC- *Gibbs-Sampler*. Caso contrário, utiliza-se o algoritmo de *Metropolis-Hastings*.

Entretanto, o método MCMC tradicional é útil somente num contexto em que a dimensão do espaço paramétrico é fixa, pode-se observar em (3) que a dimensão do espaço paramétrico varia de acordo com o número de QTLs presentes no modelo. O método MCMC com Saltos Reversíveis (MCMC-SR) proposto por Green (1995) é uma excelente proposta para a obtenção de distribuições *a posteriori* nesse contexto, uma vez que permite saltar entre modelos de diferentes dimensões.

O Método MCMC com saltos Reversíveis

O método MCMC com saltos reversíveis permite construir uma cadeia de Markov reversível $(X_i)_{i\geq 1}$ com distribuição invariante π, utilizando o algoritmo de Metropolis-Hastings modificado. A modificação consiste na inclusão do Jacobiano da transformação no cálculo da probabilidade de aceitação do algoritmo de Metropolis-Hastings de forma a contemplar a diferença existente na dimensão dos espaços paramétricos envolvidos no movimento. Mais especificamente, suponha que (s, z) é o estado atual da cadeia de Markov denotado por $X^{(t)}$, em que s é a variável indicadora do número de QTL existente no modelo e z é o vetor de parâmetros associado a este modelo e que uma proposta $X^{t=1} = \left(S^{(t+1)}, Z^{(t+1)}\right)$ é gerada para o novo estado da Cadeia de Markov. Com

probabilidade $b_{ss'}$ a proposta $S^{(t+1)}$ é igual a *s'* QTLs. Então, o movimento do modelo contendo s QTLs para o modelo contendo s' QTLs é aceito de acordo com a seguinte probabilidade de aceitação:

$$\alpha_{ss'} = \min\left\{1, \frac{\pi(s', z' \mid y)\, q_{s's}(z', u')\, b_{s's}}{\pi(s, z \mid y)\, q_{ss'}(z, u)\, b_{ss'}} \left| \frac{\partial g_{ss'}(z, u)}{\partial z \partial u} \right| \right\} \quad (4)$$

Deve-se notar que o Jacobiano que aparece na equação (4) é decorrente da transformação determinística usada no mecanismo proposto por Green (1995) para mudança da dimensão do espaço paramétrico. Os movimentos entre modelos são vistos como: (1) passo de nascimento de QTL, nesse caso, propõe-se saltar de um modelo com *s* QTL para um modelo contendo *s'* QTLs, em que *s*<*s'*, com probabilidade b_n; (2) passo morte de QTL, em que deseja-se saltar de um modelo contendo *s* QTLs para um modelo contendo *s'* QTLs, em que *s*>*s'*, com probabilidade b_m; (3) passo de permanência, onde o número de QTLs permanece fixo para o próximo estado da cadeia com probabilidade $b_p = 1 - b_n - bm$. Green (1995) propõe que $b_m + b_n \leq 0{,}9$.

De acordo com Gaffney (2001), não existe restrição quanto ao número de QTLs a serem adicionados ou removidos do modelo. No entanto, salienta que esta escolha pode afetar a taxa de aceitação e, portanto, afetar o desempenho da cadeia gerada e, consequentemente, a estimação dos parâmetros. Sendo assim, restringe-se o passo de nascimento ou morte de apenas um QTL. Obviamente a proposta de nascimento ou morte de um QTL não garante sua aceitação e como em MCMC a razão de aceitação calculada, depende da contribuição da verossimilhança, da razão de prioris para os parâmetros e da densidade proposta. Além disso, é importante salientar, que em mapeamento de QTL o Jacobiano da equação (4) é igual a um.

Indicando a probabilidade de aceitação para o nascimento de um QTL como $\min(1, A)$, então a probabilidade de aceitação para a morte de um QTL será:

$$\min\left(1, \frac{1}{A}\right)$$

Método de Seleção de Modelos – Fator de Bayes

Dado um problema de seleção de modelos, em que temos que comparar o modelo M_1 com o modelo M_2, um critério muito utilizado é o Fator de Bayes (FB), definido como:

$$FB(M_1, M_2) = \frac{\pi(M_1 \mid y) / \pi(M_1)}{\pi(M_2 \mid y) / \pi(M_2)},$$

sendo $\pi(M_i \mid y)$ a probabilidade *a posteriori* condicionada às observações do modelo M_i e $\pi(M_i)$ a probabilidade *a priori* para o modelo M_i, em que i=1,2. A Tabela 3 apresenta a calibração para o Fator de Bayes baseado no valor $FB(M_1, M_2)$ proposta por Jeffreys (1935), citado por Raftery (1995).

Tabela 3 – Desicões sobre a evidência de M_1 em relação a M2.

Valores de $FB(M_1, M_2)$	Conclusão
$1 \leq FB(M_1, M_2) \leq 3$	Evidência a favor de M_1
$3 < FB(M_1, M_2) \leq 10$	Evidência positiva a favor de M_1
$10 < FB(M_1, M_2) \leq 100$	Forte evidência a favor de M_1
$FB(M_1, M_2) > 100$	Evidência decisiva a favor de M_1

Fonte: Jeffreys (1935), citado por Raftery (1995).

Materiais

Para avaliar o desempenho do método MCMC com Saltos Reversíveis, bem como verificar a influência da escolha de diferentes distribuições a priori para o número de QTLs sobre sua estimativa *a posteriori* gerados vários conjuntos de dados utilizando o *software WinQTLCart* versão 2.5. O *WinQTLCart* é um *software* inteiramente gratuito para universidades e pode ser obtido no *site http://statgen.ncsu.edu/qtlcart/WQTLCart.htm.* Este *software* permite simular o mapa genético, a população de mapeamento para o experimento QTL e obter uma amostra da distribuição conjunta *a posteriori*, utilizando o método MCMC-SR. As distribuições a *priori* disponíveis nesse *software* são as apresentadas na Tabela 2.

Para simular os conjuntos de dados artificiais foram construídos vários mapas de ligação e populações de cruzamento. Neste artigo, apresenta-se um conjunto de dados obtidos da seguinte forma: simulou-se um mapa de ligação contendo 10 cromossomos, sendo que cada um desses cromossomos tem comprimento de 200cM; no mapa foram alocados 200 marcadores moleculares: 20 marcadores igualmente espaçados por cromossomo. Posteriormente, foram inseridos 3 QTLs ao longo do mapa simulado: o primeiro foi alocado no cromossomo 3, na posição 131,0cM; o segundo no cromossomo 4, na posição 75,7cM e o terceiro no cromossomo 6, na posição 77,3cM . O

tamanho amostral considerado foi de 300 indivíduos de uma população F2 com herdabilidade de 90%; supondo o modelo (1) foram atribuídos os efeitos aditivos e de dominância dos três QTLs, como segue: (a_1 = 1,1371; d_1 = 0,0929), (a_2 = -0,8868 2; d_2 = -0,2708), (a_3 = 1,0347; d_3 = -0,4275).

RESULTADOS E DISCUSSÃO

Para a construção da amostra da distribuição conjunta *a posteriori* utilizou-se módulo *BIM* do programa *QTLCartographer*, versão 2.5 para Windows. Foi gerada uma cadeia de 100.000, "*burn-in*" de 10.000, "*pré burn-in*" de 10.000 e "thin" de 100 iterações, obtendo-se uma amostra de tamanho 4005. A convergência da cadeia gerada foi monitorada por meio de: (i) análise gráfica; (ii) testes de diagnóstico, tendo sido, neste trabalho, utilizado o teste proposto por Geweke (1992), disponível no CODA (Convergence Diagnosis and Output Analysis) da linguagem de programação R. Foram atribuídas diferentes distribuições *a priori* para o número de QTLs, afetando a característica de interesse e para os demais parâmetros do modelo utilizou-se o *default* do programa. As distribuições *a priori* utilizadas para o número de QTLs foram: (i) a distribuição Geométrica com média 3, 5 e 10; (ii) distribuição de Poisson com média 4, 5 e 10 e a distribuição Uniforme no intervalo [0, comprimento do genoma], ou seja, [0, 200 cM]. As distribuições *a priori* utilizadas estão resumidas na Tabela 4.

Tabela 4 – Distribuições *a priori* para os parâmetros de interesse do modelo QTL.

Parâmetro	Distribuição a priori
μ	$N(0, s_y^{\,2})$
Variância, σ^2	Gama-Inversa(3, σ^2)
Efeito Aditivo, a_j	$N(0, \beta s_y^2)$, em que $\beta \sim N(2,10)$
Efeito de Dominância, d_j	$N(0, \beta s_y^2)$, em que $\beta \sim N(2,10)$
Número de QTLs, *s*	Poisson com média 4, 5 e 10 Geométrica com média 4, 5 e 10 Uniforme[0,200]

De acordo com a Figura 1, a atribuição da distribuição geométrica para o número de QTLs sugere a existência de três QTLs independentemente da média utilizada. O mesmo não é observado quando considera-se a distribuição Poisson com diferentes médias e nem com a distribuição uniforme. Quando a distribuição a *priori* Poisson com média igual a quatro ou cinco QTLs, é utilizada

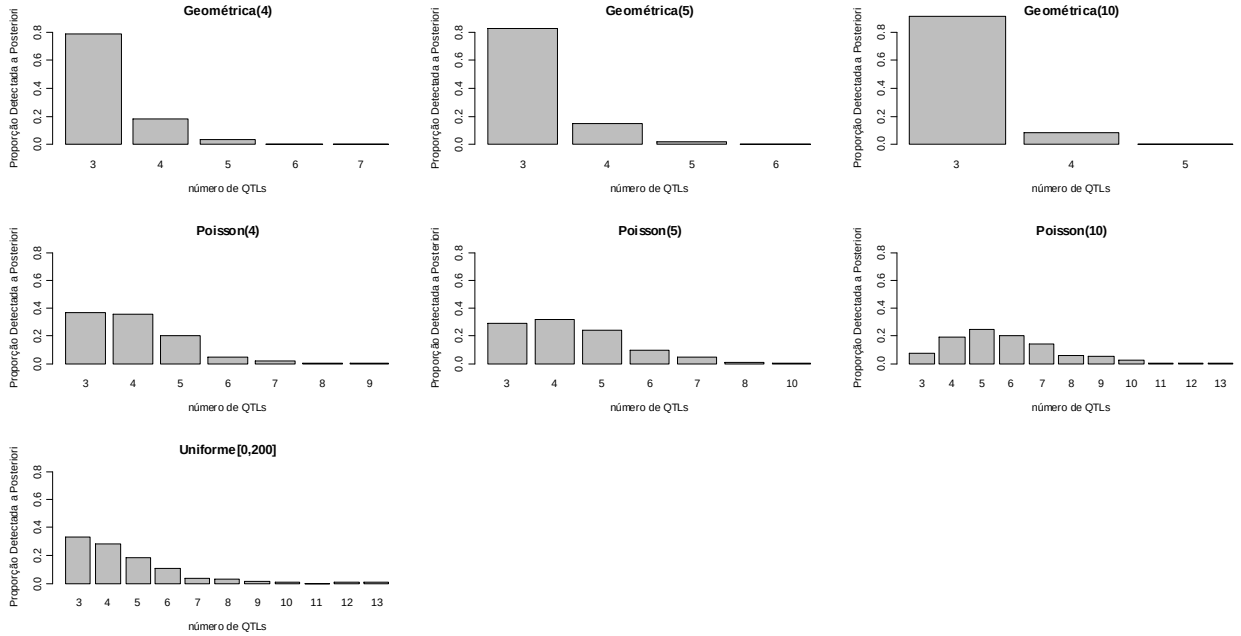

Figura 1 – Proporção *a posteriori* para o número de QTLs detectados,utilizando o método MCMC-SR, de acordo com diferentes distribuições *a priori* atribuídas para o número de QTLs afetando a característica de interesse.

não se observa, claramente, qual o número de QTLs presentes. Deve–se decidir entre três, quatro ou cinco QTLs; já o modelo com média igual a dez QTLs sugere quatro, cinco ou seis QTLs. No caso da distribuição a *priori* uniforme[0,200] a incerteza é em relação a três, quatro ou cinco QTLs presentes no modelo.

Neste contexto, surge a necessidade de calcular o Fator de *Bayes* para selecionar o modelo mais adequado. No caso da distribuição geométrica, nos três casos, comparamos o modelo contendo quatro QTLs (G4) com o modelo contendo três QTLs (G3). Pela Tabela 5, verificam-se que as evidências são fortes ou positivas a favor de G3.

Tabela 5 – Comparações, baseadas no Fator de Bayes, entre os modelos QTLs utilizando distribuição a *priori* geométrica com média de 4, 5 e 10 QTLs.

Média	Comparação	Fator de Bayes	Calibração
4	G3 em relação a G4	5.01	Evidência Positiva a Favor de G3
5	G3 em relação a G4	6.19	Evidência Positiva a Favor de G3
10	G3 em relação a G4	13.24	Forte Evidência a Favor de G3

Quando atribuímos distribuição a *priori* Poisson para o número de QTLs, todas as possíveis comparações (duas a duas) evidenciaram que o modelo contendo três QTLs (P3) é o mais adequado (Tabela 6). Ao utilizar a distribuição a *priori* Poisson com média de dez QTLs, o Fator de Bayes mostra evidências a favor ou positiva a favor do modelo contendo quatro QTLs.

No caso da atribuição de distribuição *a priori* uniforme no intervalo do comprimento do genoma, o Fator de Bayes mostra evidências a favor do modelo que contém três QTLs (U3), conforme mostra a Tabela 7.

Na Tabela 9, apresentamos as estimativas *a posteriori* para os efeitos genéticos e localizações dos QTLs (cromossomo e posição) juntamente com o intervalo HPD 95% para os modelos que indicaram a presença de três QTLs, afetando a característica fenotípica. Nota-se que a escolha das diferentes prioris para o número de QTLs não influencia na estimativa *a posteriori* dos parâmetros. Além disso, observa-se que as estimativas para todos os parâmetros do modelo QTL, exceto para o efeito de dominância do QTL1 (d_1), se aproximaram dos valores simulados. Cabe observar que d_1=0,0929 é um efeito pequeno sobre a característica de interesse. Observa-se também que o modelo QTL com distribuição a *priori P*oisson com média de dez QTLs não apresenta estimativa a *posteriori* razoáveis para os parâmetros, uma vez que detectou várias localizações (cromossomo) para cada QTL, conforme mostra a Tabela 8.

Tabela 6 – Comparações, baseadas no Fator de Bayes, entre os modelos QTLs utilizando distribuição a *priori* Poisson com média de 4, 5 e 10 QTLs.

Média	Comparação	Fator de Bayes	Calibração
4	P3 em relação a P4	1.50	Evidência a favor de P3
	P3 em relação a P5	3.24	Evidência positiva a favor de P3
	P4 em relação a P5	2.16	Evidência a favor de P4
5	P3 em relação a P4	1.53	Evidência a favor de P3
	P3 em relação a P5	2.81	Evidência a favor de P3
	P4 em relação a P5	1.83	Evidência a favor de P4

Tabela 7 – Comparações, baseadas no Fator de Bayes, entre os modelos QTLs utilizando distribuição a *priori* uniforme[0,200].

Distribuição	Comparação	Fator de Bayes	Calibração
Uniforme [0,200]	U3 em relação a U4	1.59	Evidência a favor de U3
	U3 em relação a U5	2.96	Evidência a favor de U3
	U4 em relação a U5	1.87	Evidência a favor de U4

Tabela 8 – Estimativas *a posteriori* para a localização (cromossomo) para o modelo QTL, utilizando o método MCMC-SR, utilizando distribuições *a priori* Poisson com média de dez QTLs.

QTL	Cromossomo
1	1, 2 ou 3
2	3 ou 4
3	4, 5 ou 6
4	6,7,8,9 ou 10

Tabela 9 – Estimativas *a posteriori* e intervalo HPD de 95% para os efeitos genéticos e localizações dos três QTLs simulados, utilizando o método MCMC-SR, de acordo com diferentes distribuições *a priori* atribuídas para o número de QTLs afetando a característica de interesse.

		Efeitos				Localização		
Priori	QTL	Aditivo	Intervalo HPD 95%	Dominância	Intervalo HPD 95%	Crom,	Posição	Intervalo HPD 95%
Geo(4)	1	1.0960	(0,9367 ; 1,2508)	0.02403	(-0,2051 ; 0,2252)	3	130.6523	(128,5305 ; 132,8610)
	2	-1.0269	(-1,1941 ; -0,8602)	-0.20601	(-0,4668 ; 0,0262)	4	74.8527	(72,3352 ; 77,5665)
	3	0.9769	(0,7989 ; 1,1298)	-0.46871	(-0,7051 ; -0,2092)	6	74.8963	(72,6867 ; 77,2241)
Geo(5)	1	1.0898	(0,9286 ; 1,2401)	0.01711	(-0,1967 ; 0,2482)	3	130.7189	(128,4979 ; 132,8902)
	2	-1.0148	(-1,1778 ; -0,8480)	-0.21470	(-0,4218 ; 0,0409)	4	74.7699	(72,4114 ; 77,5636)
	3	0.9785	(0,7858 ; 1,1505)	-0.47259	(-0,7210 ; -0,2237)	6	75.0480	(72,7734 ; 77,2942)
Geo(10)	1	1.0950	(0,9359 ; 1,2539)	0.01870	(-0,2136 ; 0,2299)	3	130.4991	(128,2025 ; 132,7096)
	2	-1.0258	(-1,2022 ; -0,8454)	-0.20600	(-0,4497 ; 0,0430)	4	74.7886	(72,3422 ; 77,2213)
	3	0.9771	(0,8016 ; 1,1388)	-0.47793	(-0,7268 ; -0,2404)	6	74.8907	(72,5887 ; 77,2036)
Poisson(4)	1	1.0950	(0,9502 ; 1,2648)	0.02681	(-0,1893 ; 0,2552)	3	130.5302	(128,3842 ; 132,6656)
	2	-1.0307	(-1,1884 ; -0,8508)	-0.21658	(-0,4645 ; 0,0191)	4	74.7469	(71,8581 ; 76,8538)
	3	0.9794	(0,8134 ; 1,1743)	-0.47442	(-0,7217 ; -0,2479)	6	74.7835	(71,7603 ; 76,9439)
Poisson(5)	1	1.0886	(0,9282 ; 1,2297)	0.02180	(-0,2018 ; 0,2522)	3	130.5144	(128,0946 ; 132,6760)
	2	-1.0343	(-1,2003 ; -0,8506)	-0.19011	(-0,4505 ; 0,0664)	4	74.8017	(72,3357 ; 77,2547)
	3	0.9766	(0,8001 ; 1,1543)	-0.46953	(-0,7036 ; -0,2411)	6	74.8940	(72,5213 ; 77,1217)
U[0,200]	1	1.1014	(0,9443 ; 1,2590)	0.02160	(-0,1735 ; 0,2581)	3	130.5207	(128,1060 ; 132,7823)
	2	-1.0242	(-1,1962 ; -0,8508)	-0.21536	(-0,4536 ; 0,0035)	4	74.7513	(72,3633 ; 77,2479)
	3	0.9743	(0,8058 ; 1,1557)	-0.48705	(-0,7071 ; -0,1839)	6	75.0450	(72,8825 ; 77,3568)

A convergência das cadeias dos parâmetros foi diagnosticada; não existindo evidências contra a convergência das cadeias pelo teste de diagnóstico de Geweke (Geweke, 1992), o qual é um dos testes indicados na literatura para diagnosticar convergência quando trabalha-se com cadeias únicas de tamanho grande.

CONCLUSÕES

Com base nos resultados obtidos, nota-se que o método MCMC-SR implementado software *WinQTLCart*, teve um excelente desempenho. Tanto para as estimativas *a posteriori* para o número de QTLs como para os demais parâmetros de interesse do modelo QTL, o método mostrou-se eficiente. Foi visto que é de extrema importância ao utilizar-se a abordagem bayesiana em mapeamento de QTLs, a escolha cautelosa das distribuições *a priori* para os parâmetros de interesse. Quando o pesquisador for completamente ignorante a despeito do número de QTLs que afetam a característica de interesse , os resultados mostraram que a distribuição geométrica é recomendada.

CONSIDERAÇÕES FINAIS

Com o desenvolvimento deste trabalho, surgiu o questionamento sobre a importância de considerar a inclusão de efeitos de epistasia no modelo QTL para pesquisar traços genéticos complexos. Observa-se que o programa *WinQTLCart* permite simular cruzamentos, considerando o efeito de epistasia, no entanto, o módulo *BIM* do *WinQTLCart* não está preparado para realizar análise sob o enfoque bayesiano, uma vez que não permite atribuir distribuições *a priori* para os efeitos epistásticos.

Nesse contexto, é de extrema necessidade a implementação de programas capazes de realizar o mapeamento bayesiano de QTLs, considerando efeito de epistasia. Entretanto, a inclusão de mais efeitos do modelo QTL aumenta a complexidade dos passos do MCMC-SR e, consequentemente, aumenta a demanda computacional e pode impedir o bom desempenho do algoritmo, é importante pesquisar formas alternativas para simular da distribuição conjunta *a posteriori*.

REFERÊNCIAS BIBLIOGRÁFICAS

CASELA, G.; GEORGE, E.I. Explaining the gibbs sarnpler. **The American Statistician**, Washington, v.46, n.3, p.167-74, 1992.

GAFFNEY, P.J. **An efficient reversible jump markov chain monte carlo approach to detect multiple loci and their effects in inbred crosses**. 2001. 174p. Thesis (Doctor in Philosophy in Statistcs)-University of Winconsin, Madison, 2001.

GEWEKE, J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: BERNARDO, J.M.; BERGER, J.O.; DAWID, A.P.; SMITH, A.F.M. (Ed.). **Bayesian statistics**. Oxford: Clarendon, 1992. v.4.

GREEN, P.J. Reversible jump markov chain monte carlo computation and bayesian model determination. **Biometrics**, Washington, v.82, p.711-732, 1995.

PATERSON, A.H.; DAMON, S.; HEWITT, J.D.; ZAMIR, D.; RAQBINOWITCH, H.D. Mendelian factors underlying quantitative traits in tomato comparison across species generations and environments. **Genetics**, Austin, v.127, p.181-197, 1991.

RAFTERY, A.E. Bayesian model selection in social research. **Sociological Methodology**, Oxford, v.25, p.111-163, 1995.

SATAGOPAN, J.M.; YANDELL, B.S.; NEWTON, M.A.; OSBORN, T.C. A bayesian approach to detece quantitative trait loci using markov chain Monte Carlo. **Genetics**, Austin, p.805-816, 1996.