MOYSÉS NASCIMENTO

O USO DE SIMULAÇÃO DE MONTE CARLO VIA CADEIAS DE MARKOV NO MELHORAMENTO GENÉTICO

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2009

MOYSÉS NASCIMENTO

O USO DE SIMULAÇÃO DE MONTE CARLO VIA CADEIAS DE MARKOV NO MELHORAMENTO GENÉTICO

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de "*Magister Scientiae*".

APROVADA: 20 de fevereiro de 2009.

_____	_____
Prof. Paulo Roberto Cecon (Co-Orientador)	Prof. Luiz Alexandre Peternelli (Co-Orientador)
_____	_____
Prof. Adésio Ferreira	Prof. José Marcelo Soriano Viana

_____

Prof. Cosme Damião Cruz
(Orientador)

AGRADECIMENTOS

Destaquei um agradecimento especial a algumas pessoas que tiveram um importante papel na minha formação, lembrando que estes não são todos.

Em primeiro lugar, gostaria de agradecer a Deus, que me preparou dando toda a força e sabedoria necessária para os estudos e situações adversas deste desafio.

Gostaria também de destacar o trabalho especial que minha avó e mãe tiveram na minha formação. Creio que elas tiveram o papel mais importante de toda a minha história. Fica minha gratidão pelo o apoio, seja na vida acadêmica, profissional ou pessoal.

A mulher da minha vida Ana Carolina companheira, que nunca me abandonou, estando comigo, ora nos momentos difíceis, ora nas vitórias.

Ao meu orientador Cosme Damião Cruz pela ajuda, exemplo e respeito dispensados ao longo da vida minha acadêmica. Obrigado pelo rico tesouro do conhecimento.

Aos professores (amigos) que atuaram na minha formação, em especial a Adésio Ferreira, Luiz Alexandre Peternelli e Mauro C. M. Campos que ajudaram a fortalecer o meu caráter e perfil profissional.

Ao secretário da Pós-Graduação Altino, pelo empenho em sempre ajudar e pelas mensagens de amizade e incentivo.

Aos meus colegas, em especial o pessoal do Laboratório de Bioinformática, pois estes participaram dos momentos mais importantes da minha vida acadêmica.

À banca, composta por Adésio Ferreira, Luiz Alexandre Peternelli, José Marcelo Soriano Viana e Paulo Roberto Cecon que aceitaram o convite que lhes foi feito e, dessa forma, colaboraram para conclusão deste projeto.

À Universidade Federal de Viçosa pela estrutura e oportunidade de desenvolver este projeto.

Ao CNPq pela concessão da bolsa para auxiliar no desenvolvimento deste projeto.

A todos os meus parentes e amigos, dos mais diversos meios. Vocês atuaram precisamente na minha vida, definindo exatamente o que sou.

SUMÁRIO

RESUMO

NASCIMENTO, Moysés, M.Sc., Universidade Federal de Viçosa, fevereiro de 2009. **O Uso de simulação de monte Carlo via cadeias de Markov no melhoramento genético.** Orientador: Cosme Damião Cruz. Co-orientadores: Luiz Alexandre Peternelli e Paulo Roberto Cecon.

Este trabalho teve por objetivo fornecer um referencial teórico e aplicado sobre os principais métodos de simulação de Monte Carlo via cadeias de Markov (*MCMC*), buscando dar ênfase em aplicações no melhoramento genético. Assim, apresentaram-se os algoritmos de Metropolis-Hastings, s*imulated annealing* e amostrador de Gibbs. Os aspectos teóricos dos métodos foram abordados através de uma discussão detalhada de seus fundamentos com base na teoria de cadeias de Markov. Além da discussão teórica, aplicações concretas foram desenvolvidas. O algoritmo de Metropolis-Hastings foi utilizado para obter estimativas das freqüências de recombinação entre pares de marcadores de uma população F_2, de natureza codominante, constituída de 200 indivíduos. O s*imulated annealing* foi aplicado no estabelecimento da melhor ordem de ligação na construção de mapas genéticos de três populações F_2 simuladas, com marcadores de natureza codominantes, de tamanhos 50, 100 e 200 indivíduos respectivamente. Para cada população foi estabelecido um genoma com quatro grupos de ligação, com 100 cM de tamanho cada. Os grupos de ligação possuem 51, 21, 11 e 6 marcadores, com uma distância de 2, 5, 10 e 20 cM entre marcas adjacentes respectivamente, ocasionando diferentes graus de saturação. Já o amostrador de Gibbs foi utilizado na obtenção das estimativas dos parâmetros de adaptabilidade e estabilidade, do modelo proposto por Finlay e Wilkinson (1963), através da inferência bayesiana. Foram utilizados os dados de médias de rendimento de cinco genótipos

avaliados em nove ambientes, provenientes de ensaios em blocos ao acaso com quatro repetições. Em todas as aplicações os algoritmos se mostraram computacionalmente viáveis e obtiveram resultados satisfatórios.

ABSTRACT

NASCIMENTO, Moysés, M.Sc., Universidade Federal de Viçosa, February, 2009. **The Use of Monte Carlo simulation via Markov chains in genetic breeding.** Advisor: Cosme Damião Cruz. Co-Advisors: Luiz Alexandre Peternelli and Paulo Roberto Cecon.

The objective of this work was to provide a theoretical and applied reference on the main Monte Carlo simulation methods via Markov chains (*MCMC*), seeking to focus on applications in genetic breeding. Thus, the algorithms of Metropolis-Hastings, simulated annealing and the Gibbs sampler were presented. The theoretical aspects of the methods were approached through a detailed discussion about their foundations based on the Markov chain theory. Besides the theoretical discussion, concrete applications were developed. The Metropolis-Hastings algorithm was used to achieve estimates from the frequencies of recombination between pairs of markers of a population F_2, of co-dominant nature, with 200 individuals. The s*imulated annealing* was applied to establish a better linking order in the construction of genetic maps of three simulated populations F_2, with markers of co-dominant nature, containing 50, 100 and 200 individuals, respectively. For each population, it was established a genome with four linking groups, each with 100 cM of size. The linking groups present 51, 21, 11 and 6 markers, with a distance of 2, 5, 10 and 20 cM between the adjacent marks, respectively, providing different degrees of saturation. The Gibbs sampler, on the other hand, was used for the achievement of the estimates of the adaptability and stability parameters of the model proposed by Finlay and Wilkinson (1963), through the Bayesian inference. The data of the productivity averages of five genotypes evaluated in nine environments were used, come from essays in randomized blocks with four

replications. In all the applications, the algorithms were computationally viable and achieved satisfactory results.

1. INTRODUÇÃO GERAL

1.1. Contextualização

Desde a década de 90, principalmente devido aos avanços dos recursos computacionais, os métodos de simulação de Monte Carlo via cadeias de Markov (*MCMC*) passaram a ser tema obrigatório para profissionais em diversas áreas. Estes métodos surgiram como alternativa para solução de problemas complexos em inferência estatística (clássica e bayesiana). Dentre estes problemas, dois dos maiores, são os problemas de integração e problemas de otimização (ROBERT e CASELLA, 1999). Estatisticamente, estes problemas podem ser descritos da seguinte forma.

Seja $X = (X_1, \ldots, X_d)$ um vetor aleatório d-dimensional em Λ com distribuição de probabilidade

$$\pi(x) = \begin{cases} c\upsilon(x) & \text{se } x \in \Lambda; \\ 0 & \text{caso contrário,} \end{cases}$$

onde c é uma constante (possivelmente desconhecida).

Problema (i): Deseja-se calcular a quantidade

$$I = E(h(X)) = c\int h(x)\upsilon(x)dx,$$

para alguma função $h : \Lambda \to \Re$.

A obtenção do valor de I pode ser feita analiticamente, entretanto, quando a dimensão do vetor aleatório é grande, a solução torna-se inviável. Para contornar o problema da alta dimensionalidade, diversas técnicas de computação intensiva são citadas na literatura. As opções são:

(i) Integração Numérica (fórmula de Newton-Cotes: fórmula dos trapézios e a fórmula de Simpson): Difícil e imprecisa quando o valor de *d* é grande;
(ii) Simulação de Monte Carlo: Consiste em utilizar um gerador de números aleatórios para obter uma amostra $(X_1,\ldots,X_n)$ independente e identicamente distribuída (i.i.d.) da distribuição de X e estimar *I* pela média amostral

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i).$$

Entretanto, nem sempre é fácil obter uma amostra da distribuição de probabilidade, principalmente quando se trata de um vetor aleatório de variáveis dependentes. Outra situação que inviabiliza a utilização do método ocorre quando se conhece apenas o núcleo da função de probabilidade, isto é, a constante *c* é desconhecida.
(iii) Simulação de Monte Carlo via Cadeias de Markov (*MCMC*): Estes métodos surgem para contornar as dificuldades citadas anteriormente. A proposta é simular π via construção de uma cadeia de Markov em Λ tendo π como sua única distribuição estacionária. Os métodos *MCMC* garantem que, após um tempo suficientemente longo de simulação, elementos de Λ podem ser amostrados com distribuição aproximadamente igual a π.

Problema (ii): Seja Λ um conjunto qualquer e seja f uma função com domínio em Λ ($f:\Lambda\to\Re$). Deseja-se encontrar um ponto $\boldsymbol{x}\in\Lambda$ que minimiza[1] a função f.

Novamente a solução deste problema pode não ser trivial, principalmente quando a cardinalidade de Λ é grande. Deste modo, para obter uma solução, é necessário fazer uso de métodos computacionais. Os algoritmos numéricos para solução deste problema são essencialmente classificados em métodos de programação matemática e métodos probabilísticos. Entre os métodos probabilísticos, um algoritmo que se destaca é conhecido como s*imulated annealing* (KIRKPATRICK, et al. 1983), que na verdade é uma pequena modificação no conhecido algoritmo *MCMC* de Metropolis-Hastings (1970), transformando-o em um algoritmo de otimização.

[1] Pode-se observar que o problema de encontrar um $\boldsymbol{x}\in\Lambda$ que maximize a função $g:\Lambda\to\Re$ recai no problema (ii). Basta ver que maximizar g é o mesmo que minimizar $f=-g$.

1.2. Tema

Segundo Robert e Casella (2008), os métodos de simulação de Monte Carlo via Cadeias de Markov (*MCMC*) são tão antigos quanto os métodos de Monte Carlo, entretanto seu impacto em aplicações estatísticas surgiram a partir da década de 90, excetuando-se em algumas áreas especificas como Estatística Espacial e Análises de Imagens. Os métodos de Monte Carlo surgiram em Los Alamos, Novo México, durante a segunda guerra mundial, em meados de 1950. Estudos provenientes destes métodos resultaram no algoritmo de Metropolis. Enquanto os métodos de Monte Carlo foram utilizados por todo o tempo, os métodos *MCMC* aproximaram-se dos estatísticos a partir do trabalho de Hastings (1970).

O primeiro algoritmo *MCMC* conhecido por algoritmo de Metropolis foi publicado por Metropolis et al. (1953). Este algoritmo surgiu dos estudos feitos pelo mesmo grupo de cientistas criadores do método de Monte Carlo, chamados de cientistas de Los Alamos, que contavam em sua maioria com Físicos trabalhando em física matemática e no projeto da bomba atômica[2] (ROBERT e CASELLA, 2008). Uma interessante variação, do algoritmo de Metropolis et al. (1953) é o *simulated annealing*, desenvolvido por Kirkpatrick et al. (1983).

O algoritmo de Metropolis foi generalizado por Hastings (1970) e Peskun (1973;1981) como uma ferramenta de simulação estatística.

Após trinta anos de relativo esquecimento, os métodos *MCMC* passaram a ser tema obrigatório para profissionais em diversas áreas de conhecimento tais como: estatísticos que utilizam a abordagem bayesiana em análise de dados (GELFAND e SMITH, 1990), profissionais que trabalham em reconstrução de imagens (GEMAN e GEMAN, 1984), Físicos teóricos interessados em problemas fundamentais da mecânica estatística (BINDER e HEERMANN, 1997), Físicos e Engenheiros de Materiais interessados em problemas da matéria condensada (BINDER e HEERMANN, 1997), entre outros.

Recentemente os métodos *MCMC* vêm se destacando na resolução de problemas complexos no melhoramento genético. Como exemplo, pode-se citar a estimação do nível de significância em testes exatos, utilizados na avaliação da hipótese de equilíbrio

[2] O processo de construção da bomba atômica não envolve processo de simulação, embora o posterior desenvolvimento, bomba de hidrogênio faz.

de Hardy-Weinberg (GUO e THOMPSON, 1992; YUAN e BONNEY, 2003; HUBER et al., 2006), na estimação de parâmetros genéticos via inferência bayesiana, possibilitando a obtenção de estimativas pontuais e intervalos de credibilidade para as distribuições a posteriori dos parâmetros (GIANOLA e FERNANDO, 1986), e mapeamento e detecção de QTL (SILVA, 2006).

Entretanto, um referencial teórico e aplicado tratando deste assunto no melhoramento genético não existe. Assim, acredita-se que a elaboração de uma dissertação nesta área dará valiosas contribuições à área científica, estimulando os melhoristas no seu entendimento e aplicação rotineira em seus programas de melhoramento.

Deste modo o presente trabalho visa abordar e aplicar os principais métodos *MCMC* em problemas complexos no melhoramento genético, discutindo os fundamentos teóricos dos algoritmos. Especificamente, serão expostos os algoritmos de Metropolis-Hastings, Amostrador de Gibbs e *simulated annealing*. Os algoritmos serão apresentados em um nível apropriado para alunos e pesquisadores interessados neste assunto com conhecimentos básicos em Probabilidade e Estatística. Além disso, será apresentada uma lista de referências (livros e artigos publicados) que seja apropriada para orientar leitores que desejam fazer um estudo mais aprofundado.

1.3. Organização do Trabalho

Esta dissertação está organizada da seguinte maneira:

1. Na revisão de literatura apresenta-se a teoria de cadeias de Markov em tempo discreto com espaço de estados finitos. Resultados desta teoria são diretamente utilizados na exposição dos algoritmos de Metropolis-Hastings, amostrador de Gibbs e s*imulated annealing*, que também são apresentados na revisão de literatura. Além disso, exemplos simples são apresentados com o objetivo explícito de ilustrar e ressaltar o “funcionamento básico” dos algoritmos. A revisão é terminada com uma pequena introdução à inferência bayesiana;

2. No Capítulo 1, o algoritmo de Metropolis-Hastings é utilizado para estimar a freqüência de recombinação entre pares de marcadores de uma população F_2 simulada;
3. No Capítulo 2 utiliza-se o *simulated annealing* no estabelecimento da melhor ordem de ligação na construção de mapas genéticos de três populações simuladas (F_2);
4. No Capítulo 3, o amostrador de Gibbs foi utilizado na obtenção das estimativas dos parâmetros de adaptabilidade e estabilidade do modelo proposto por Finlay e Wilkinson (1963) através da Inferência Bayesiana;
5. Finalmente, apresentam-se as conclusões do trabalho.

2. REVISÃO DE LITERATURA

2.1. Cadeias de Markov em tempo discreto

Esta seção é dedicada à teoria de cadeias de Markov em tempo discreto. O objetivo principal é apresentar os conceitos e resultados dessa teoria que serão diretamente utilizados na seção seguinte, que tratará dos métodos *MCMC*.

2.1.1. Introdução

Seja Λ um conjunto finito. Um processo estocástico em tempo discreto em Λ é uma seqüência $(X_n)_{n\geq 0}$ de variáveis aleatórias tal que X_n assume valores em Λ para todo $n \geq 0$. O conjunto Λ é chamado de espaço de estados e cada um de seus elementos é chamado de estado (HOEL et al., 1987).

Neste trabalho, o interesse recai em processos com a propriedade que dado o estado presente, os estados passados não influenciam o estado futuro. Esta propriedade é conhecida como propriedade de Markov e processos satisfazendo tal propriedade são chamados de cadeias de Markov. Formalmente, o processo $(X_n)_{n\geq 0}$ é uma cadeia de Markov em Λ se

$$P(X_{n+1} = y \mid X_0 = x_0, \ldots, X_{n-1} = x_{n-1,} X_n = x) = P(X_{n+1} = y \mid X_n = x) \quad (1)$$

Além disso, $(X_n)_{n\geq 0}$ é uma cadeia homogênea se

$$P(X_{n+1} = y \mid X_n = x) = P(X_1 = y \mid X_0 = x) = P(x, y). \quad (2)$$

Para todo $n > 0$ e $x, y \in \Lambda$ (HOEL et al., 1987).

2.1.2. Função de Transição e Distribuição Inicial

Seja $(X_n)_{n\geq 0}$ uma cadeia de Markov homogênea em Λ. A função $P : \Lambda \times \Lambda \to [0,1]$ definida por $P(x, y) = P(X_1 = y \mid X_0 = x)$, representa a função de transição da cadeia (HOEL et al., 1987). A partir da definição tem-se

- $P(x, y) \geq 0$ para todo $x, y \in \Lambda$;
- $\sum_{y\in\Lambda} P(x, y) = 1$ para todo $x \in \Lambda$.

Além disso, segue das equações (1) e (2) que $P(X_{n+1} = y \mid X_0 = x_0, \ldots, X_{n-1} = x_{n-1,} X_n = x) = P(X_{n+1} = y \mid X_n = x)$. Em palavras, se a cadeia esta no estado *x* no tempo *n*, então não importa como ela chegou em *x*, a cadeia possui probabilidade $P(x, y)$ de visitar o estado *y* no tempo *n*+1 (BRÉMAUD, 1999).

A função $\pi_0(x) : \Lambda \to [0,1]$ definida por $\pi_0(x) = P(X_0 = x)$ representa a distribuição inicial da cadeia. Novamente, segue da definição que

- $\pi_0(x) \geq 0$ para todo $x \in \Lambda$;
- $\sum_{x\in\Lambda} \pi_0(x) = 1$.

A distribuição conjunta de $(X_0, X_1, \ldots, X_n)$, pode ser expressa em termos da função de transição e distribuição inicial:

$$P(X_0 = x_0, \ldots, X_n = x_n) = \pi_0(x_0)P(x_0, x_1)P(x_1, x_2) \cdot \ldots \cdot P(x_{n-1}, x_n)$$

para todo $n > 0$ *e* $x_0, \ldots, x_n \in \Lambda$.

Este fato pode ser demonstrado pelo teorema da multiplicação[3], onde tem-se que $P(X_0 = x_0, \ldots, X_n = x_n)$ é igual a

[3] Seja (Ω, F, P) um espaço de probabilidade e sejam $A_0, \ldots, A_n$ eventos em F. Então:

$\pi_0(x_0)P(X_1 = x_1 \mid X_0 = x_0)\cdot\ldots\cdot P(X_n = x_n \mid X_0 = x_0,\ldots,X_{n-1} = x_{n-1})$.

Utilizando-se a propriedade de Markov tem-se

$P(X_0 = x_0,\ldots,X_n = x_n) = \pi_0(x_0)P(x_0,x_1)P(x_1,x_2)\cdot\ldots\cdot P(x_{n-1},x_n)$.

2.1.3. Função de Transição em *m*-Passos

Seja $(X_n)_{n\geq 0}$ uma cadeia de Markov homogênea em Λ com função de transição P e distribuição inicial π_0 e seja *m* um inteiro não-negativo.

A função $P^m : \Lambda\times\Lambda \to [0,1]$ definida por

$$P^m(x,y) = P(X_{n+m} = y \mid X_n = x_n) = P(X_m = y \mid X_0 = x)$$

Representa a função de transição em *m*-passos da cadeia. Em particular

$$P^0(x,y) = \begin{cases} 0 & \text{se } x \neq y; \\ 1 & \text{se } x = y; \end{cases}$$

e

$$P^1 = P.$$

Considere $n, m \geq 0$ e $x, y \in \Lambda$. Como

$$\begin{aligned} P^{n+m}(x,y) &= P(X_{n+m} = y \mid X_0 = x) \\ &= \sum_{z\in\Lambda} P(X_n = z, X_{n+m} = y \mid X_0 = x) \\ &= \sum_{z\in\Lambda} P(X_n = z \mid X_0 = x)P(X_{n+m} = y \mid X_0 = x, X_n = z), \end{aligned}$$

é possível concluir que

$$P^{n+m}(x,y) = \sum_{z\in\Lambda} P^n(x,z)P^m(z,y). \qquad (3)$$

Além disso, segue da equação (3) que

$$\begin{aligned} P^m(x,y) &= \sum_{z\in\Lambda} P(x,z)P^{m-1}(z,y) \\ &= \sum_{z_1\in\Lambda}\ldots\sum_{z_{m-1}\in\Lambda} P(x,z_1)P(z_1,z_2)\cdot\ldots\cdot P(z_{m-1},y). \end{aligned}$$

Agora considere $n > 0$ e $y \in \Lambda$. Como

$P(A_0 \cap\ldots\cap A_n) = P(A_0)P(A_0 \mid A_1)\cdot\ldots\cdot P(A_n \mid \bigcap_{i=1}^{n-1} A_i)$.

$$\begin{aligned}\pi_n(y) &= P(X_n = y) \\ &= \sum_{x\in\Lambda} P(X_0 = x, X_n = y) \\ &= \sum_{x\in\Lambda} P(X_0 = x)P(X_n = y \mid X_0 = x),\end{aligned}$$

segue que

$$\pi_n(y) = P(X_n = y) = \sum_{x\in\Lambda} \pi_0(x)P^n(x,y). \tag{4}$$

De forma alternativa, como

$$\pi_n(y) = \sum_{x\in\Lambda} P(X_{n-1} = x)P(X_n = y \mid X_{n-1} = x),$$

segue também que

$$\pi_n(y) = \sum_{x\in\Lambda} \pi_{n-1}(x)P(x,y). \tag{5}$$

Fechando a subseção, é importante ressaltar que é possível pensar em π_n e P^n, $n > 0$, como vetores e matrizes em Λ e $\Lambda \times \Lambda$ respectivamente. Portanto, na notação matricial, as equações (3), (4) e (5) podem ser descritas como:

- $P^{n+m} = P^n P^m$, onde $n, m \geq 0$;
- $\pi_n = \pi_0 P^n$, onde $n > 0$;
- $\pi_n = \pi_{n-1} P$, onde $n > 0$.

A equação $\pi_n = \pi_{n-1}P$ permite interpretar a matriz P como um operador linear que atua no espaço das distribuições de probabilidade em Λ, atualizando a distribuição marginal da cadeia a cada passo $n > 0$ (HOEL et al., 1987).

2.1.4. Classificação dos Estados

Seja $(X_n)_{n\geq 0}$ uma cadeia de Markov homogênea em Λ com função de transição P e sejam x e y estados não necessariamente distintos.

Defina a variável aleatória T_y por

$T_y = \min\{n > 0; X_n = y\}$

se $X_n = y$ para algum $n > 0$ e por $T_y = \infty$ se $X_n \neq y$ para todo $n > 0$. Em palavras, T_y representa o tempo da primeira visita ao estado *y*. Em particular, se o estado inicial da cadeia é *y*, então T_y representa o tempo de retorno a *y*.

Seja $\rho_{xy} = P(T_y < \infty \mid X_0 = x)$ a probabilidade de uma cadeia visitar o estado *y* em algum tempo finito, dado que ela começou do estado *x*. Um estado *y* é

- Transiente se $\rho_{yy} < 1$;
- Recorrente se $\rho_{yy} = 1$.

Se *y* é um estado transiente, então uma cadeia começando em *y* tem uma probabilidade positiva de nunca retornar a *y*. De fato

$$P(T_y = \infty \mid X_0 = y) = 1 - P(T_y < \infty \mid X_0 = y) = 1 - \rho_{yy} > 0.$$

Se *y* é um estado recorrente, então uma cadeia começando em *y* retorna a *y* com probabilidade 1.

Diz-se que um estado *y* é absorvente se $P(y, y) = 1$. Além disso, todo estado absorvente é recorrente.

Defina a variável aleatória N_y por $N_y = \sum_{n \geq 1} I(X_n = y)$, onde $I(X_n = y)$ denota o indicador do evento $[X_n = y]$. Em palavras, N_y representa o número de visitas ao estado *y*. Pode-se mostrar que a distribuição de $N_y \mid X_0 = x$ é dada por

$$P(N_y = m \mid X_0 = x) = \begin{cases} 1 - \rho_{xy} & \text{se } m = 0; \\ \rho_{xy} \rho_{yy}^{m-1} (1 - \rho_{yy}) & \text{se } m > 0. \end{cases}$$

Além disso, tem-se que

$$G(x, y) = E(N_y \mid X_0 = x) = \sum_{n \geq 1} P^n(x, y).$$

Então $G(x, y)$ denota o número esperado de visitas a *y* para uma cadeia que se inicia em *x* (HOEL et al., 1987) .

O próximo teorema descreve as diferenças fundamentais entre um estado transiente e um estado recorrente.

Teorema 1. (HOEL et al., 1987) (i) *Seja y um estado transiente. Então*

$$P(N_y < \infty \mid X_0 = x) = 1$$

e

$$G(x,y)=\frac{\rho_{xy}}{1-\rho_{yy}}<\infty$$

para todo $x\in\Lambda$.

(ii) *Seja y um estado recorrente. Então*

$$P(N_y<\infty \mid X_0=x)=\rho_{xy}$$

e

$$G(x,y)=\begin{cases}0 & \text{se } \rho_{xy}=0;\\ \infty & \text{se } \rho_{xy}>0,\end{cases}$$

para todo $x\in\Lambda$. *Em particular se* $x=y$, *segue que,* $P(N_y=\infty \mid X_0=y)=1$ *e* $G(x,y)=\infty$.

Demonstração. Seja *y* um estado transiente. Como $0\le\rho_{yy}<1$, segue que

$$P(N_y=\infty \mid X_0=x)=\lim_{m\to\infty}P(N_y\ge m \mid X_0=x)=\lim_{m\to\infty}\rho_{xy}\rho_{yy}^{m-1}=0.$$

Além disso,

$$G(x,y)=\sum_{m\ge 0}mP(N_y=\infty \mid X_0=x)=\sum_{m>0}m\rho_{xy}\rho_{yy}^{m-1}(1-\rho_{yy})=\frac{\rho_{xy}}{1-\rho_{yy}}<\infty.$$

Agora seja *y* um estado recorrente. Como $\rho_{yy}=1$, segue que

$$P(N_y=\infty \mid X_0=x)=\lim_{m\to\infty}P(N_y\ge m \mid X_0=x)=\lim_{m\to\infty}\rho_{xy}=\rho_{xy}.$$

Se $\rho_{xy}>0$, então $P(N_y=\infty \mid X_0=x)>0$ e, portanto $G(x,y)=\infty$. Se $\rho_{xy}=0$, então

$$P(N_y=\infty \mid X_0=x)=\begin{cases}1 & \text{se } m=0;\\ 0 & \text{se } m>0;\end{cases}$$

e portanto $G(x,y)=0$. Em particular, quando $x=y$, segue imediatamente que

$$P(N_y=\infty \mid X_0=y)=\rho_{yy}=1>0 \text{ e } G(y,y)=\infty.$$

Se *y* é um estado transiente, então a cadeia faz somente um número finito de visitas a *y* independente do estado inicial. Além disso, o número médio de visitas a *y* é finito. Suponha que *y* é recorrente. Neste caso, se a cadeia começou em *y*, ela retorna a *y* infinitas vezes. Por outro lado, se a cadeia começou em algum outro estado *x*, pode ser impossível que ocorra uma visita a *y*. Entretanto, se a cadeia alcança *y* pelo menos uma vez, então a cadeia visita *y* infinitas vezes (HOEL et al., 1987).

Corolário 1. Um estado y é transiente se, e somente se, $G(y,y) < \infty$. Um estado y é recorrente se, e somente se, $G(y,y) = \infty$.

Corolário 2. *Se y é um estado transiente, então* $\lim_{n\to\infty} P^n(x,y) = 0$ *para todo* $x \in \Lambda$.

Corolário 3. *Toda cadeia de Markov homogênea com espaço de estados finitos possui pelo menos um estado recorrente.*

2.1.5. Decomposição do Espaço de Estados

Seja $(X_n)_{n\geq 0}$ uma cadeia de Markov homogênea em Λ com função de transição P.

Diz-se que um estado *x* alcança um estado *y* se $P^n(x,y) > 0$ para algum inteiro $n \geq 0$. Para denotar este fato, usa-se o símbolo $x \to y$. A relação $\to$ é reflexiva e transitiva, ou seja:

- $x \to x$;
- se $x \to y$ e $y \to z$, então $x \to z$.

De fato, $x \to x$ pois $P^0(x,x) = 1 > 0$. Além disso, se $x \to y$ e $y \to z$, então existem inteiros positivos $n, m \geq 0$ tais que $P^n(x,y) > 0$ e $P^m(y,z) > 0$. Assim $P^{n+m}(x,z) = \sum_{w\in\Lambda} P^n(x,w)P^m(w,z) \geq P^n(x,y)P^m(y,z)$, portanto $x \to z$.

Diz-se que *x* se comunica com *y* se $x \to y$ e $y \to x$. Para denotar a relação de comunicação entre *x* e *y*, usa-se o símbolo $x \leftrightarrow y$. A relação $\leftrightarrow$ é reflexiva, simétrica e transitiva, ou seja:

- $x \leftrightarrow x$;
- se $x \leftrightarrow y$, então $y \leftrightarrow x$;
- se $x \leftrightarrow y$ e $y \leftrightarrow z$, então $x \leftrightarrow z$.

As duas primeiras propriedades são imediatas da definição. A terceira segue do fato que a relação $\to$ é transitiva (GRIMMETT e STIRZAKER, 1992).

Teorema 2. (HOEL et al., 1987) *Seja x um estado recorrente e suponha que* $x \leftrightarrow y$. *Então y é recorrente e* $\rho_{xy} = \rho_{yx} = 1$.

Um conjunto C de estados é dito ser fechado se $\rho_{xy} = P(T_y < \infty \mid X_0 = x) = 0$, para todo $x \in C$ e $y \notin C$. De forma equivalente, C é fechado se somente se $P^n(x, y) = 0$ para todo $x \in C$ e $y \notin C$ e $n \geq 1$.

Se *C* é fechado, então a cadeia de Markov que começa em *C* permanece em *C* por todo o tempo, com probabilidade 1. Como conseqüência, se $\{a\}$ é um estado absorvente, então $\{a\}$ é fechado.

Diz-se que o conjunto *C* é irredutível se $x \leftrightarrow y$ para todo $x, y \in \Lambda$. A partir do Teorema 2 segue que se *C* é um conjunto fechado e irredutível, todos os estados em *C* são recorrentes ou transientes. O resultado seguinte é uma conseqüência imediata dos Teoremas 1 e 2.

Corolário 4. Seja *C* um conjunto fechado irredutível de estados recorrentes. Então $\rho_{xy} = 1$, $P(N(y) = \infty \mid X_0 = x) = 1$ e $G(x, y) = \infty$, para quaisquer $x, y \in C$.

Uma cadeia de Markov irredutível é uma cadeia cujo espaço de estados é irredutível, isto é, todos os estados se comunicam entre si. Tal cadeia é necessariamente transiente ou recorrente. O Corolário 4 implica, que uma cadeia de Markov recorrente e irredutível, visita todos os estados infinitamente com probabilidade 1.

Teorema 3. (HOEL et al., 1987) Seja *C* um conjunto fechado irredutível finito. Então todo estado em *C* é recorrente.

O próximo teorema afirma que todos os estados que estão em uma mesma classe ou são todos recorrentes ou são todos transientes. Uma classe de estados recorrentes será chamada de classe recorrente. Nomenclatura análoga vale para uma classe de estados transientes.

Teorema 4. (GRIMMETT e STIRZAKER, 1992). *Seja C um conjunto fechado irredutível. Então todos os estados em C ou são recorrentes ou são transientes.*

Demonstração. Considere um par arbitrário $x, y \in C$ e assuma que y é transiente. Por definição existem inteiros $n, m > 0$ tais que $P^n(x, y) > 0$ e $P^m(y, x) > 0$. Para qualquer inteiro $k \leq 1$, temos que $P^{n+k+m}(y, y) \leq P^m(y, x)P^k(x, x)P^n(x, y)$. Assim $\sum_{k \geq 1} P^k(x, x) \leq \frac{1}{P^n(x, y)P^m(y, x)} \sum_{k \geq 1} P^{n+k+m}(y, y) < \infty$. Portanto x também é transiente.

Seja C um conjunto não-vazio de estados. Diz-se que C é fechado se nenhum estado dentro de C alcança um estado fora de C. Formalmente C é fechado se $P^m(x,y) = 0$, para $x \in C$, $y \notin C$ e todo $m \geq 1$.

Se a cadeia começa em uma classe recorrente ela permanece eternamente nessa classe com probabilidade 1, visitando infinita vezes todos estados desta classe. Se a cadeia começa em um estado transiente, então após um tempo aleatório finito, a cadeia alcança uma classe recorrente e permanece para sempre nessa classe, visitando todos os estados infinitas vezes.

Seja x um estado tal que $P^n(x,x) > 0$ para algum $n \geq 1$. O período d_x de x é definido por

$$d_x = m.d.c.\{n \geq 1; P^n(x,x) > 0\}$$

Diz-se que x é um estado periódico se $d_x > 1$ e aperiódico se $d_x = 1$. Em particular se $P(x,x) > 0$, então $d_x = 1$ e x é aperiódico. É possível mostrar que o estado x se comunica com o estado y, $d_x = d_y$. Portanto todos os estados de uma cadeia de Markov irredutível possuem o mesmo período e nesse caso, faz sentido falar que a cadeia possui período d. Em particular, se $d = 1$, diz-se que a cadeia é aperiódica.

2.1.6. Distribuição Estacionária e Teorema Limite

Seja $(X_n)_{n\geq 0}$ uma cadeia de Markov homogênea em Λ com função de transição P. Diz-se que uma função $\pi : \Lambda \to [0,1]$ é uma distribuição estacionária para $(X_n)_{n\geq 0}$ se:

- $\pi(x) \geq 0$, para todo $x \in \Lambda$;
- $\sum_{x\in\Lambda} \pi(x) = 1$;
- $\sum_{x\in\Lambda} \pi(x)P(x,y) = \pi(y)$, para todo $y \in \Lambda$.

Pode-se mostrar que

$$\sum_{x\in\Lambda} \pi(x)P^n(x,y) = \pi(y) \tag{6}$$

Para todo $n \geq 0$ e $y \in \Lambda$ (GRIMMETT e STIRZAKER, 1992).

Quando Λ é finito, é possível pensar em π como um vetor em Λ tal que $\pi = \pi P$. Além disso, na notação matricial a equação (6) pode ser escrita como $\pi = \pi P^n$, para todo $n \geq 0$.

Suponha que a distribuição estacionária π existe e que

$\lim_{n\to\infty} P^n(x, y) = \pi(y)$,

para todo $y \in \Lambda$. Então

1. $\lim_{n\to\infty} \pi_n(y) = \lim_{n\to\infty} P(X_n = y) = \pi(y) \quad \forall\ y \in \Lambda$.
2. π é a única distribuição estacionária da cadeia $(X_n)_{n\geq 0}$.

Teorema 6. (GRIMMETT e STIRZAKER, 1992) *Uma cadeia de Markov homogênea em Λ irredutível possui uma única distribuição estacionária π*, dada por

$$\pi(x) = \frac{1}{E(T_x \mid X_0 = x)}$$

para todo $x \in \Lambda$.

Teorema 7. (HOEL et al., 1987) *Seja $(X_n)_{n\geq 0}$ uma cadeia de Markov homogênea em Λ irredutível, aperiódica e com distribuição estacionário π.*

- $X_n \sim \pi$ *para todo* $n \geq 0$ *se* $X_0 \sim \pi$*;*
- $\lim_{n\to\infty} P^n(x, y) = \pi(y)$;
- *Independente da distribuição de* X_0,

$\pi_n \overset{VT}{\to} \pi$.

Isso significa que[4]

$\lim_{n\to\infty} d_{VT}(\pi_n, \pi) = 0.$

[4] Sejam $\upsilon : \Lambda \to [0,1]$ e $\pi : \Lambda \to [0,1]$ distribuições de probabilidade em Λ. A distância variação total entre υ e π é definida por

$$d_{VT}(\upsilon, \pi) = \frac{1}{2}\sum_{x\in\Lambda} |\upsilon(x) - \pi(x)|.$$

Sejam $\pi, \pi_1, \pi_2, \ldots$ distribuições de probabilidade em Λ. Dizemos que π_n converge para π em variação total quando $n \to \infty$ se

$\lim_{n\to\infty} d_{VT}(\pi_n, \pi) = 0.$

Para denotar o fato que π_n converge para π em variação total quando $n \to \infty$, usa-se o símbolo

$$\pi_n \overset{VT}{\to} \pi.$$

Teorema 8. (BRÉMAUD, 1999) (Teorema Ergódico para Cadeias de Markov). *Considere as mesmas hipóteses do teorema 7. Se* $h : \Lambda \to R$ *é uma função tal que*

$$E_\pi(|\, h(X)\,|) = \sum_{x\in\Lambda} |\, h(x)\,|\pi(x) < \infty,$$

então

$$\frac{1}{n-m}\sum_{i=m+1}^{n} h(X_i) \overset{q.c.}{\to} E_\pi(h(X))$$

quando $n \to \infty$.

As provas dos teoremas 6, 7 e 8 serão omitidas. Leitores interessados nas demonstrações podem consultar Hoel et al. (1972), Grimmett e Stirzaker (1992) e Brémaud (1999). Os resultados serão apresentados através de um exemplo.

Exemplo 1. Considere duas urnas rotuladas por A e B e considere 3 bolas rotuladas por 1, 2 e 3. Inicialmente algumas destas bolas estão na urna A e as restantes estão na urna B. Um inteiro é selecionado aleatoriamente do conjunto $\{1,2,3\}$ e a bola rotulada por este inteiro é removida da urna. Agora seleciona-se aleatoriamente uma das urnas e coloca-se a bola removida dentro da urna selecionada. Este procedimento é repetido indefinidamente, onde as seleções são feitas de forma independente. Seja X_n o número de bolas na urna A no tempo n. A seqüência $(X_n)_{n\geq 0}$ é uma cadeia de Markov homogênea em $\Lambda = \{0,1,2,3\}$, com função de transição.

$$P(x,y) = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/6 & 1/2 & 1/3 & 0 \\ 0 & 1/3 & 1/2 & 1/6 \\ 0 & 0 & 1/2 & 1/2 \end{bmatrix}$$

Percebe-se que cadeia é irredutível e aperiódica. Além disso, uma conta simples ($\pi P = \pi$) revela que a função $\pi = (1/8;3/8;2/8;1/8)$ é uma distribuição estacionária para a cadeia $(X_n)_{n\geq 0}$. Se $\pi_0 = \pi$, então $\pi_n = \pi$ para todo $n > 0$, pois $\pi = \pi P$ e $\pi_n = \pi_{n-1} P$ para to do $n > 0$. Se, porém, $\pi_0 \neq \pi$, então $\pi_n \overset{VT}{\to} \pi$. De fato, se $\pi_0 = (1/4;1/4;1/4;1/4)$.

- $\pi_1 = (0{,}167;0{,}333;0{,}333;0{,}167)$;
- $\pi_3 = (0{,}130;0{,}370;0{,}370;0{,}130)$;

- $\pi_{30} = (0{,}125;0{,}375;0{,}375;0{,}125)$;

Observa-se também que $\lim_{n\to\infty} P^n(x,y) = \pi(y)$

$$P^4 = \begin{bmatrix} 0{,}204 & 0{,}444 & 0{,}297 & 0{,}056 \\ 0{,}148 & 0{,}401 & 0{,}352 & 0{,}099 \\ 0{,}099 & 0{,}352 & 0{,}401 & 0{,}148 \\ 0{,}056 & 0{,}296 & 0{,}444 & 0{,}204 \end{bmatrix} \text{ e } P^{31} = \begin{bmatrix} 0{,}125 & 0{,}375 & 0{,}375 & 0{,}125 \\ 0{,}125 & 0{,}375 & 0{,}375 & 0{,}125 \\ 0{,}125 & 0{,}375 & 0{,}375 & 0{,}125 \\ 0{,}125 & 0{,}375 & 0{,}375 & 0{,}125 \end{bmatrix}.$$

Finalmente, se $h: x \to x^2$, então

$$\frac{1}{n-1000}\sum_{i=1001}^{n} X_i^{\,2} \xrightarrow{q.c.} \sum_{x\in\Lambda} x^2\pi(x) = 3.$$

quando $n \to \infty$.

2.1.7. Cadeias Reversíveis

Seja $(X_n)_{n\geq 0}$ uma cadeia de Markov homogênea em Λ com função de transição P. Diz-se que uma distribuição de probabilidade $\pi: \Lambda \to [0,1]$ é reversível para $(X_n)_{n\geq 0}$ se

$$\pi(x)P(x,y) = \pi(y)P(y,x) \tag{7}$$

para todo $x, y \in \Lambda$. A cadeia é dita reversível se existe uma distribuição reversível para ela.

Teorema 9. (HÄGGSTRÖM, 2001) *Seja $(X_n)_{n\geq 0}$ uma cadeia de Markov homogênea em Λ e função de transição $P(x,y)$. Se π é uma distribuição reversível para $(X_n)_{n\geq 0}$, então π é uma distribuição estacionária para $(X_n)_{n\geq 0}$.*

Demonstração. Como o sistema de equações em (7) é válido, segue que

$$\sum_{x\in\Lambda}\pi(x)P(x,y) = \sum\pi(y)P(y,x) = \pi(y)\sum_{x\in\Lambda}P(y,x) = \pi(y).$$

Portanto π é uma distribuição estacionária para $(X_n)_{n\geq 0}$.

A vantagem do uso do teorema 9 é que em muitas situações o sistema de equações em (7) é bem mais simples de ser resolvido do que o problema de autovetor $\pi = \pi P$.

Exemplo 2. Seja $(X_n)_{n\geq 0}$ uma cadeia de Markov homogênea em $\Lambda = \{0,1,2\}$, com função de transição

$$P = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/6 & 1/6 & 1/3 \\ 0 & 3/4 & 1/4 \end{bmatrix}$$

Usando as equações em (7), obtém-se que

$$\pi(0)\cdot\frac{1}{2} = \pi(1)\cdot\frac{1}{6}$$

$$\pi(1)\cdot\frac{1}{3} = \pi(2)\cdot\frac{3}{4}$$

Logo $\pi(1) = 3\pi(0)$ e $\pi(2) = (4/3)\pi(0)$. Como as componentes de π devem somar 1 para que ela seja uma distribuição de probabilidade em Λ, tem-se que

$$\pi = \left(\pi(0);3\pi(0);\frac{4}{3}\pi(0)\right) = (3/16;9/16;4/16).$$

Logo π é uma distribuição reversível para $(X_n)_{n\geq 0}$. Pode-se ver que π satisfaz a equação $\pi = \pi P$ e, portanto é também uma distribuição estacionária para $(X_n)_{n\geq 0}$.

Exemplo 3. Este exemplo mostra que nem toda distribuição estacionária é também reversível. Seja $(X_n)_{n\geq 0}$ uma cadeia de Markov homogênea em $\Lambda = \{0,1,2\}$, com função de transição

$$P = \begin{bmatrix} 2/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/2 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

Pode-se ver através de $\pi = \pi P$, que $\pi = (16/43;12/43;15/43)$ é uma distribuição estacionária para $(X_n)_{n\geq 0}$, entretanto não é uma distribuição reversível para de $(X_n)_{n\geq 0}$. De fato

$$4/13 = \pi(0)P(0,1) \neq \pi(1)P(1,0) = 3/43.$$

2.1.8. Cadeias de Markov Não-Homogêneas

Toda teoria apresentada até aqui, refere-se a cadeias de Markov homogênea. Estas se caracterizam pelo fato de que a regra probabilística para obter X_{n+1} de X_n não depende do tempo n. Em certas situações esta hipótese é relaxada, permitindo então que as probabilidades de transição mudem com o tempo. Tais cadeias, cujas regras de transição dependem do tempo, são chamadas de cadeias não-homogêneas e são definidas da seguinte forma.

Seja Λ um conjunto finito e seja $(P_n)_{n\geq 1}$, uma seqüência de funções em $\Lambda \times \Lambda$ tal que:

- $P_n(x, y) \geq 0$ para todo $n \geq 1$ e $x, y \in \Lambda$;
- $\sum_{y\in\Lambda} P_n(x, y) = 1$ para todo $n \geq 1$ e $x \in \Lambda$.

Uma cadeia de Markov não-homogênea em Λ é uma seqüência $(X_n)_{n\geq 0}$ de variáveis aleatórias assumindo valores em Λ tal que

$$P(X_{n+1} = y \mid X_0 = x_0, \ldots, X_{n-1} = x_{n-1}, X_n = x) = P(X_{n+1} = y \mid X_n = x) = P_{n+1}(x, y)$$

, para todo $n \geq 0$ e $x_0, \ldots, x_{n-1}, x, y \in \Lambda$. O caso particular de cadeias homogêneas ocorre quando $P_{n+1}(x, y) = P_1(x, y) = P(x, y)$ para todo $n > 0$ (BRÉMAUD, 1999).

2.1.9. Exemplos de cadeia de Markov na genética

Esta subseção tem como objetivo apresentar alguns exemplos de aplicações de cadeias de Markov na genética.

Exemplo 4. (ROSS, 2003) A hipótese de equilíbrio de Hardy-Weinberg entre os alelos é de fundamental importância em estudos genéticos. Esta lei diz que em uma população suficientemente grande e na ausência de seleção, migração e mutação, o equilíbrio é atingido após uma geração de acasalamento ao acaso (“aaa”), de maneira que a relação genotípica torna-se igual ao quadrado da freqüência gênica e, com as sucessivas gerações de acasalamento ao acaso, permanece inalterada. Para ilustrar este fato, será considerada uma população inicial com genótipos *AA*, *Aa* e *aa*, nas freqüências *D*, *H* e *R*, respectivamente. As freqüências alélicas são *p* e *q*, para *A* e *a*, respectivamente.

Considerando que ocorre acasalamento ao acaso entre os indivíduos desta população, pode-se predizer a descendência, conforme ilustrado na Tabela 1.

Tabela 1 - Freqüência genotípica e alélica numa população antes e após acasalamento ao acaso (aaa)

População Inicial (P_o)		População após o "aaa" (P_1)
$AA = D$		$AA = D_1 = p^2$
$Aa = H$	$\Rightarrow$	$Aa = H_1 = 2pq$
$Aa = R$		$Aa = R_1 = q^2$
$f(A) = p = D + H/2$		$f(A) = p_1 = p$
$f(a) = q = R + H/2$		$f(a) = q_1 = q$

Assim, pode-se conhecer as freqüências genotípica e alélica que ocorrerão numa geração futura, derivada de sucessivos acasalamentos ao acaso numa população inicial, a partir da sua freqüência alélica (*p* e *q)* original. Este conhecimento preditivo permite aos pesquisadores estabelecer estratégias de melhoramento e manipulação de população, bem como reconhecer a dinâmica evolutiva da espécie em determinadas regiões. O exposto pode ser facilmente demonstrado se considerados os cruzamentos, na população original, ilustrados na Tabela 2.

Tabela 2 - Relação dos possíveis acasalamentos numa população e predição da descendência resultante do acasalamento ao acaso

Cruzamentos em P_o	Descendência em P_1			
	Freq.	AA	Aa	aa
$AA \times AA$	D^2	D^2	-	-
$AA \times Aa$	$2DH$	DH	DH	-
$AA \times aa$	$2DR$	-	$2DR$	-
$Aa \times Aa$	H^2	$H^2/4$	$H^2/2$	$H^2/4$
$Aa \times aa$	$2HR$	-	HR	HR
$aa \times aa$	R^2	-	-	R^2
Total	1,0	$(D+H/2)^2$ p^2	$2(D+H/2)(R+H/2)$ $2pq$	$(R+H/2)^2$ q^2

Assim, demonstra-se que a freqüência genotípica da descendência pode ser predita por meio do conhecimento da freqüência alélica na população genitora. Com o acasalamento ao acaso, a freqüência alélica não se altera, ou seja:

$f(A \text{ em } P1) = p_1 = D + ½\ H = p^2 + ½\ 2pq = p$

$f(a \text{ em } P1) = q_1 = R + ½\ H = q^2 + ½\ 2pq = q$

A relação genotípica da descendência é, portanto, dada por $(p_A + q_a)^2$.

Considere uma população em equilíbrio de Hardy-Weinberg, isto é, a freqüência dos genótipos *AA*, *Aa* e *aa* estão estabilizadas em *p*, *r* e *q*, respectivamente. Deseja-se acompanhar a história genética de um único indivíduo e de seus descendentes (por simplicidade assume-se que cada indivíduo gere apenas um descendente). Assim para um determinado individuo, denota-se X_n o estado genético de seu descendente na *n*-ésima geração. A função de transição da cadeia de Markov é dada por

$$P(x, y) = \begin{bmatrix} p+\frac{r}{2} & q+\frac{r}{2} & 0 \\ \frac{p}{2}+\frac{r}{4} & \frac{p}{2}+\frac{q}{2}+\frac{r}{2} & \frac{q}{2}+\frac{r}{4} \\ 0 & p+\frac{r}{2} & q+\frac{r}{2} \end{bmatrix}$$

Pode-se representação a função de transição no diagrama apresentado a seguir.

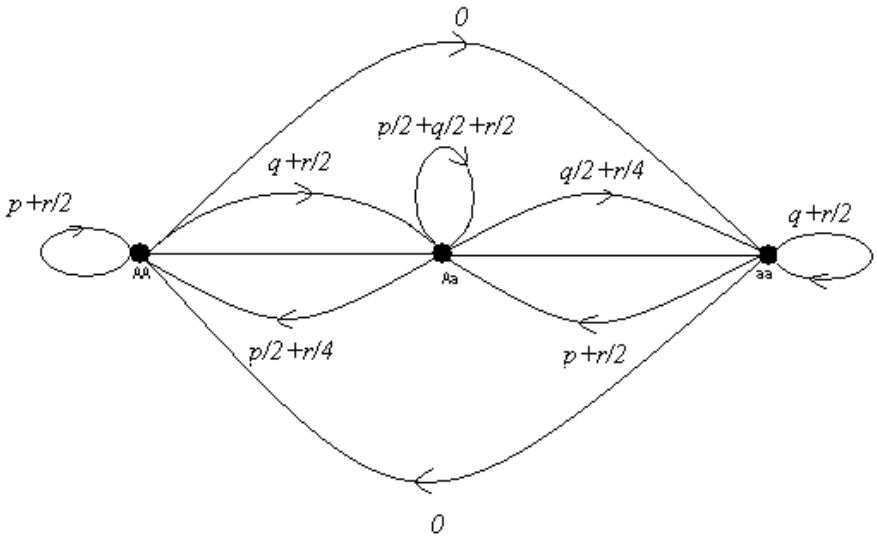

Figura 1 - Diagrama das probabilidades de transição da cadeia.

A partir do diagrama pode-se ver que a cadeia é irredutível, isto é, todos estados genéticos se comunicam entre si ($\leftrightarrow$), e, além disso, pelo fato de a cadeia não possuir nenhum ciclo ela é dita aperiódica.

Exemplo 5. (HOEL et al., 1987) Considere um gene representado por *a* alelos, onde cada alelo pode ser considerado normal ou mutante. Considere uma célula com um gene composto por *m* alelos mutantes e *a*-*m* alelos normais. Sabe-se que, antes da célula dividir-se em duas células filhas, o gene duplica-se. O gene correspondente, pertencente a uma das células filhas, é composto de *a* alelos escolhidos aleatoriamente em 2*m* alelos mutantes e 2(*a*-*m*) alelos normais. Suponha que fixe uma linha de descendentes para um dado gene. Seja X_0 o número de alelos mutante inicialmente presentes, e seja X_n o número presente no *n*-ésimo gene descendente. Então $X_{n(n>0)}$, é uma cadeia de Markov em $\Lambda = \{0,1,2,\ldots,a\}$ e sua função de transição é dada por

$$P(x,y) = \frac{\binom{2x}{y}\binom{2a-2x}{a-y}}{\binom{2a}{a}},$$

percebe-se que os estados 0 e *a* são estados absorventes da cadeia.

2.2. Métodos de Simulação de Monte Carlo via Cadeias de Markov (MCMC)

2.2.1. Introdução

A abordagem habitual da teoria de cadeias de Markov inicia-se com a função de transição da cadeia, definida por $P(x,y) \quad \forall \; x,y \in \Lambda$. A função de transição denota a probabilidade de a cadeia mover-se para o estado *y* dado que se encontra no estado *x* no tempo anterior.

O maior interesse da teoria de simulação de Monte Carlo via Cadeias de Markov é determinar sob quais condições existe a distribuição estacionária π e quais condições fazem com que a função de transição convirja para a distribuição estacionária. Sabe-se,

da discussão apresentada na seção anterior, que a distribuição estacionária satisfaz $\pi = \pi P$ e que $P^n(x, y)$ representa a função de transição em *n*-passos da cadeia, deste modo, deseja-se que $\lim_{n\to\infty} P^n(x, y) = \pi(y)$, isto é, quando n tende ao infinito, elementos da função de transição possam ser amostrados de uma distribuição aproximadamente igual à distribuição estacionária.

Seja $X = (X_1, X_2, \ldots, X_d)$ um vetor aleatório *d*-dimensional assumindo valores em Λ com distribuição de probabilidade π, dada por:

$$\pi(x) = \begin{cases} c\upsilon(x) & \text{se } x \in \Lambda; \\ 0 & \text{caso contrário.} \end{cases}$$

Onde *c* é uma constante desconhecida ou difícil de ser calculada. Em outras palavras conhece-se somente o núcleo de $\pi(x)$. Para gerar amostras de $\pi(x)$, os métodos *MCMC* encontram e utilizam a função de transição $P(x, y)$ que converge para $\pi(x)$ na *n*-ésima iteração. O processo é iniciado em um estado arbitrário *x* e após um número suficientemente longo de simulação as observações geradas são aproximadamente iguais a distribuição alvo $\pi(x)$. O problema então se resume a encontrar uma função de transição $P(x, y)$ apropriada.

2.2.2. Algoritmo de Metropolis-Hastings

O algoritmo de Metropolis-Hastings é, sem dúvida, o mais fundamental dos métodos *MCMC*, pois todos os outros são derivações dele. O algoritmo originalmente foi proposto por Metropolis et al. (1953) e generalizado por Hastings em 1970.

O algoritmo usa a mesma idéia dos métodos de aceitação rejeição, isto é, um valor é gerado de uma distribuição auxiliar e aceito com uma dada probabilidade, para mais detalhes sobre estes métodos ver Ehlers (2004) e Robert e Casella (1999).

Suponha que $X = (X_1, \ldots, X_d)$, seja um vetor aleatório discreto *d*-dimensional com distribuição de probabilidade π, cujo espaço amostral é um conjunto finito Λ. O objetivo é mostrar que o algoritmo de Metropolis-Hastings é capaz de obter amostras da distribuição do vetor X. É importante ressalvar que a exposição restrita ao caso em que Λ é finito, não perde em generalidade. O algoritmo tal como será apresentado funciona

perfeitamente e sem modificações fundamentais mesmo em uma situação onde Λ é um conjunto infinito (ROBERT e CASELLA, 1999)

O algoritmo propõe uma cadeia de Markov $(X_n)_{n\geq 0}$ em Λ com distribuição estacionária π. Os ingredientes básicos são:

- Uma função de transição auxiliar $q(x, y)$ tal que

- $0 \leq q(x, y) \leq 1$, para todo $(x, y) \in \Lambda \times \Lambda$;

- $\sum_{y \in \Lambda} q(x, y) = 1$, para todo $x \in \Lambda$.

- Uma função $\alpha(x, y)$ tal que

- $0 \leq \alpha(x, y) \leq 1$, para todo $(x, y) \in \Lambda \times \Lambda$;

- $\alpha(x, x) = 1$, para todo $x \in \Lambda$.

Pode-se pensar em $q(x, y)$ como a função de transição auxiliar de uma cadeia de Markov auxiliar em Λ. Além disso, é necessário que essa cadeia seja irredutível e aperiódica (condições de regularidade) para que a cadeia do algoritmo também seja. Estas condições, discutidas anteriormente, garantem a convergência da cadeia para a distribuição estacionária. Geralmente essas condições são satisfeitas se $q(x, y)$ possui densidade positiva no mesmo suporte de $\pi(x)$ e também quando o suporte da distribuição é restrito (exemplo: distribuição uniforme definida em um intervalo (a, b) definido (CHIB e GREEBERG, 1995).

Considere $q(x, y)$ como uma densidade da qual seja possível gerar candidatos. Deste modo a densidade geradora de candidatos é denotada por $q(x, y)$. Se $q(x, y)$ satisfaz a condição de reversibilidade definida em 2.1.7 como $\pi(x)q(x, y) = \pi(y)q(y, x)$ para todo $x, y \in \Lambda$, então π é distribuição estacionária única de $(X_n)_{n\geq 0}$. Entretanto, geralmente esta condição não é satisfeita. Pode-se encontrar

$$\pi(x)q(x, y) > \pi(y)q(y, x) \tag{8}$$

Neste caso, percebe-se que a freqüência que cadeia se move de *x* para *y* é maior que de *y* para *x*. Uma forma conveniente de correção é reduzir o número que movimentos de *x* para *y* introduzindo a probabilidade $\alpha(x, y) < 1$ com que o movimento é feito. Se o movimento não é realizado, a cadeia retorna *x* como valor a distribuição alvo π. Então as transições de *x* para *y* são feitas de acordo com a função de transição de Metropolis-Hastings (P_{MH})

$$P_{MH} = q(x,y)\alpha(x,y), \quad \text{se } x \neq y,$$

onde $\alpha(x,y)$ será definida adiante.

Considerando novamente a inequação (8), percebe-se também que o movimento de *y* para *x* não é feito com freqüência suficiente. Deste modo deve-se definir $\alpha(y,x)$ tão grande quanto possível, desde que se trate de uma probabilidade, o valor máximo é igual a 1. A partir destas informações define-se a probabilidade de movimento $\alpha(x,y)$ impondo que a condição de reversibilidade seja satisfeita,

$$\begin{aligned}\pi(x)q(x,y)\alpha(x,y) &= \pi(y)q(y,x)\alpha(y,x) \\ &= \pi(y)q(y,x).\end{aligned}$$

Pode-se ver que

$$\alpha(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}.$$

Se o sinal da inequação (8) for revertido, isto é, $\pi(x)q(x,y) > \pi(y)q(y,x)$, assume-se $\alpha(x,y) = 1$ e deriva $\alpha(y,x)$ como demonstrado acima.

A idéia principal é introduzir as probabilidades $\alpha(x,y)$ e $\alpha(y,x)$ de forma garantir que os dois lados da inequação estejam em equilíbrio satisfazendo a condição de reversibilidade.

Assim, para que $P_{MH}(x,y)$ possua a condição de reversibilidade, a probabilidade de mudança $\alpha(x,y)$ deve ser definida no conjunto

$$\alpha(x,y) = \min\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right), \quad \text{se } \pi(x)q(x,y) > 0.$$

Para completar a definição da função de transição do algoritmo, é necessário considerar a possibilidade que a cadeia permaneça em *x*. Denotando este evento por $r(x)$ a probabilidade associada é dada por

$$r(x) = 1 - q(x,y)\alpha(x,y).$$

Pode-se definir a cadeia do algoritmo da seguinte maneira. Quando $X_n = x$, simule um vetor "candidato" $Y \sim q(x,y)$. Supondo $Y = y$, faça

$$X_{n+1} = \begin{cases} y & \text{com probabilidade } \alpha(x,y); \\ x & \text{com probabilidade } 1-\alpha(x,y). \end{cases}$$

A função de transição da cadeia de Metropolis-Hastings é dada por

$$P_{MH}(x,y) = \begin{cases} q(x,y)\alpha(x,y) & \text{se} \quad x \neq y; \\ 1 - \sum q(x,y)\alpha(x,y) & \text{se} \quad x = y. \end{cases}$$

O fato de $P_{MH}(x,y)$ ser reversível por construção somado ao Teorema 9, nos faz concluir que $\pi(x)$ é distribuição estacionária da cadeia (CHIB e GREENBERG, 1995).

Algoritmo 1. O algoritmo de Metropolis-Hastings pode ser descrito assim:

1. Escolha uma função de transição $q(x,y)$;
2. Escolha $X_0 \in \Lambda$;
3. Para $n \geq 0$ e $X_n = x$ simule $Y \sim q(x,y)$ e lance uma $U \sim (0,1)$. Supondo que $Y = y$, faça

$$X_{n+1} = \begin{cases} y & \text{se} \quad U < \alpha(x,y); \\ x & \text{caso contrário.} \end{cases}$$

4. $n \leftarrow n+1$ e retorne para o passo 3.

Pode-se observar que o fato de eventualmente conhecer somente o núcleo da distribuição π (ou seja, $\pi(x) \propto \upsilon(x)$) não representa nenhum impedimento quanto à utilização do algoritmo pois

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right) = \min\left(1, \frac{\upsilon(y)q(y,x)}{\upsilon(x)q(x,y)}\right),$$

pode ainda ser calculada exatamente para todo $(x,y) \in \Lambda \times \Lambda$.

Um caso particular é quando q é simétrica, ou seja, $q(x,y) = q(y,x)$ para $(x,y) \in \Lambda \times \Lambda$, é conhecido como algoritmo de Metropolis. Observa-se nesse caso

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\right).$$

Outro caso particular ocorre quando q é escolhida tal que $q(x,y) = q'(y)$, ou seja a distribuição proposta depende apenas do valor gerado naquela iteração. Assim

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)q'(x)}{\pi(x)q'(y)}\right) = \min\left(1, \frac{\pi(y)/q'(y)}{\pi(x)/q'(x)}\right).$$

2.2.3. Amostrador de Gibbs

Esta subseção tem objetivo de apresentar outro método *MCMC* conhecido como Amostrador de Gibbs, para amostrar da distribuição de interesse π. O método é simplesmente um caso particular do algoritmo de Metropolis-Hastings. A discussão exige a seguinte notação:

- $x^{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$;
- $X^{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$
- $\pi_i(x_i \mid x^{-i}) = P(X_i = x_i \mid X^{-i} = x^{-i})$

As distribuições $\pi_i(\cdot \mid x^{-i})$ são conhecidas como distribuições condicionais completas. Uma característica que torna o método interessante é que apenas essas distribuições são utilizadas na simulação. Assim, mesmo desejando simular valores de uma distribuição de alta dimensão, as simulações são feitas através de uma distribuição unidimensional (CASELLA e GEORGE, 1992).

A cadeia do algoritmo é definida da seguinte maneira: Se

$X_n = x = (x_1, \ldots, x_d)$,

escolha uniformemente um índice em $\{1, \ldots, d\}$. Se o índice escolhido foi i, então simule um valor

$X \sim \pi_i(x_i \mid x^{-i})$.

Se $X = x$, então o vetor candidato é dado por

$y = (x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_d)$

onde

$$q(y \mid x) = \frac{\pi_i(x_i \mid x^{-i})}{d}.$$

Uma conta revela que $\alpha(x, y) = 1 \quad \forall \quad (x, y) \in \Lambda \times \Lambda \quad \alpha(x,y) = 1$. De fato

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right);$$

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)\pi_i(x_i \mid x^{-i})}{\pi(x)\pi_i(x \mid x^{-i})}\right);$$

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)\pi(x)P(X^{-i} = x^{-i})}{\pi(x)\pi(y)P(X^{-i} = x^{-i})}\right);$$

$$\alpha(x, y) = 1.$$

Portanto, quando utiliza-se o amostrador de Gibbs, o vetor candidato é sempre aceito como o próximo estado da cadeia.

Algoritmo 2. O algoritmo amostrador de Gibbs pode ser descrito assim:

1. Escolha X_0 em Λ;
2. Para $n \geq 0$ e $X_n = (X_1, \ldots, X_d)$, escolha uniformemente um índice i em $\{1, \ldots, d\}$ e simule

   $X \sim \pi_i(x_i \mid x^{-i})$;
3. Se $X = x$, então

   $X_{n+1} = (x_i, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_d)$;
4. Faça $n \leftarrow n+1$ e retorne a passo 2.

De maneira alternativa, é possível atualizar componente a componente do estado da cadeia do algoritmo de forma seqüencial. Portanto, alternativamente, o amostrador de Gibbs pode ser descrito tal como segue:

Algoritmo 3. Algoritmo amostrador de Gibbs seqüencial:

1. Escolha $X_0 = (X_{1.0}, \ldots, X_{d.0})$ em Λ;
2. Obtenha o estado $X_{n+1} = (X_{1.n+1}, \ldots, X_{d.n+1})$ de $X_n = (X_{1.n}, \ldots, X_{d.n})$ via simulação seqüencial dos valores

   $X_{1.n+1} \sim \pi_1(x_1 \mid x_{2.n}, \ldots, x_{d.n})$

   $X_{2.n+1} \sim \pi_1(x_2 \mid x_{1.n+1}, x_{3.n}, \ldots, x_{d.n})$

   $\vdots$

   $X_{d.n+1} \sim \pi_d(x_d \mid x_{1.n+1}, \ldots, x_{d-1.n})$
3. Faça $n \leftarrow n+1$ e retorne a passo 2.

Embora o amostrador de Gibbs seja um caso especial do algoritmo de Metropolis-Hastings, Casella e George (1992) fazem algumas considerações sobre o método:

1. A taxa de aceitação do algoritmo é igual a 1. Assim, todo o valor gerado é aceito;
2. O uso do amostrador de Gibbs implica em limitações na escolha da distribuição candidata ($X \sim \pi_i(x_i \mid x^{-i})$) e requer um conhecimento a priori sobre as propriedades probabilísticas de π. Em outras palavras, é necessário o conhecimento da distribuição;
3. O amostrador é por construção um algoritmo multidimensional.

2.2.4. Avaliação da Convergência

Seja π uma distribuição de interesse com suporte em Λ. Na subseção 2.2.2 foi mostrado que o algoritmo de Metropolis-Hastings é capaz de obter amostras em Λ tal que cada valor sorteado é proveniente de uma distribuição aproximadamente igual a π. Da exposição feita até aqui, é possível ver que um valor da distribuição π é obtido somente quando o tempo de simulação da cadeia do algoritmo tende ao infinito. Entretanto, na prática, isso normalmente não é satisfeito uma vez que o processo de simulação é interrompido após um tempo n suficientemente longo e finito. Após este tempo *n*, os valores amostrados são considerados como sendo provenientes da distribuição de interesse π, ou seja, a distribuição marginal da cadeia no tempo *n* (denotada por π_n) está "próxima" da distribuição de interesse π. A dificuldade é determinar quanto tempo é necessário esperar para que π_n esteja "próxima" de π. Não existe nenhuma resposta simples para este problema e muito esforço tem sido feito para responder questões relacionadas a convergência dos métodos *MCMC.*

Existem duas maneiras de abordar o problema. A primeira é mais teórica e tenta medir distâncias e estabelecer cotas entre a distribuição marginal da cadeia no tempo *n* e a distribuição estacionária π. Este tipo de abordagem, é discutida nos artigos de Meyn e Tweedie (1994) e Rosenthal (1995).

A segunda abordagem do problema de convergência dos métodos *MCMC* utiliza uma perspectiva estatística. A estratégia é analisar as propriedades das saídas da cadeia do algoritmo. A dificuldade é devido ao fato dessa abordagem ser totalmente empírica, desta forma ela nunca garante formalmente a convergência (HÄGGSTRÖM, 2001).

Neste trabalho, utilizou-se a segunda abordagem para avaliar a convergência dos algoritmos.

2.2.5. *Simulated annealing*

O próximo método discutido, s*imulated annealing* (KIRKPATRICK et al. 1983), corresponde a um conhecido método *MCMC* (*Markov Chain Monte Carlo*, especificamente o Algoritmo de Metropolis-Hastings), modificado de forma a se tornar um algoritmo de otimização.

Considere Λ um conjunto enumerável. O problema é encontrar um vetor $x \in \Lambda$, que minimize ou maximize uma função de interesse $f : \Lambda \to \Re$. A idéia fundamental do método de s*imulated annealing* é emprestada da física. Em física da matéria condensada, *annealing* é um processo térmico utilizado para minimizar a energia livre de um sólido. Informalmente o processo pode ser descrito em duas etapas: (i) aumentar a temperatura do sólido até ele derreter; (ii) Diminuir lentamente a temperatura até as partículas se organizarem no estado de mínima energia do sólido. Esse processo físico pode ser simulado no computador usando o algoritmo de Metropolis-Hastings.

Suponha que o estado atual do sólido é x, e que a energia desse estado é $H(x)$. Um estado candidato y, de energia $H(y)$, é gerado aplicando uma pequena perturbação no estado x. A regra de decisão para aceitar o estado candidato utiliza a seguinte probabilidade

$$\alpha_T(x,y) = \begin{cases} 1 & \text{se} \quad H(y) - H(x) \leq 0; \\ \exp\left(-\dfrac{H(y) - H(x)}{T} \right) & \text{se} \quad H(y) - H(x) > 0. \end{cases}$$

Onde T denota a temperatura. Observe que podemos reescrever essa probabilidade da seguinte maneira

$$\alpha_T(x,y) = \min\left\{1, \exp\left(-\frac{H(y)-H(x)}{T}\right)\right\}.$$

Se o resfriamento é realizado lentamente, o sólido atinge o equilíbrio térmico a cada temperatura. Do ponto de vista de simulação, isso significa gerar muitas transições a uma certa temperatura T (ROBERT e CASELLA, 1999).

Seja x_0 um valor inicial, c_0 o parâmetro de controle inicial e L_0 o número inicial de iterações utilizados para um mesmo valor de c_0. O s*imulated annealing* pode ser descrito assim:

Algoritmo 4. *Simulated Annealing*

1. Escolha $n = 0$, $x = x_0 \in \Lambda$, c_0 e L_0;
2. Faça *i* de 1 até L_n
   - Gere y na vizinhança de x e gere uma variável aleatória $X \sim U(0,1)$;
   - Se $f(y) \leq f(x)$, então $x \leftarrow y$;
   - Se $f(y) > f(x)$ e $U < \exp\left(-\frac{f(y)-f(x)}{c_n}\right)$, então $x \leftarrow y$;
   - Fim do faça;
3. $n \leftarrow n+1$;
4. Defina c_n e L_n, e volte até o passo 2 até um critério de parada.

Onde L_n é o número de transições da cadeia em cada temperatura (c_n).

A partir da idéia do algoritmo percebe-se que seqüência $(c_n)_{n\geq 0}$ deve ser escolhida tal que $c_n \to 0$ lentamente quando $n \to \infty$.

2.2.6. Exemplos

Os exemplos aqui apresentados têm como objetivo ilustrar o funcionamento básico dos algoritmos.

Exemplo 6. Considere uma variável aleatória discreta $X \sim \pi(1/3;1/2;1/6)$. Estime a quantidade $I = E(X^2)$ via algoritmo de Metropolis-Hastings**.**

Antes de aplicar o algoritmo de Metropolis-Hastings a fim de obter amostras de π, vamos exibir explicitamente neste exemplo todos os ingredientes do algoritmo, ou seja, as funções $q(x,y)$, $\alpha(x,y)$ e $P(x,y)$ da cadeia do algoritmo. Aqui $\Lambda = \{1,2,3\}$ e q foi escolhida como

$$q(x,y) = \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 4/9 & 4/9 & 1/9 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}.$$

Usando o fato que $\alpha(x,y) = \min\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)},1\right)$, obtém-se

$$\alpha(x,y) = \begin{bmatrix} 1 & 1 & 2/3 \\ 3/4 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Agora pode-se exibir explicitamente a função de transição do algoritmo

$$P(x,y) = \begin{bmatrix} 1/3 & 1/2 & 1/6 \\ 1/3 & 5/9 & 1/9 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}.$$

Observa-se que $\pi = \pi P$, ou seja, π é de fato a única distribuição estacionária da cadeia. A estimação de *I* é feita por

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

O valor exato de I é $23/6 \approx 3{,}83$ e a estimativa obtida pelo algoritmo de Metropolis-Hastings foi de 3,82. A Figura 1 apresenta um gráfico indicando a convergência do estimador $\hat{I}_n$.

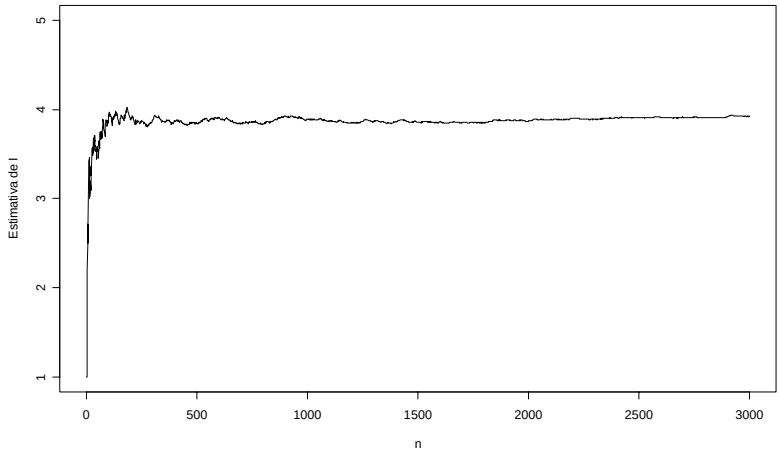

Figura 1. Convergência do estimador $\hat{I}_n$.

Exemplo 7. Seja $X = (X_1, X_2)$ um vetor bidimensional discreto, com distribuição de probabilidade conjunta $\pi(x)$ dada por

$\pi(x_1, x_2)$	0	1
0	1/3	1/6
1	1/6	1/3

Obter uma estimativa para o valor esperado da função *f* definida por

$$h(X_1, X_2) = \frac{1}{1+e^{-(X_1+X_2)}},$$

Ou seja, estimar um valor para a quantidade

$$I = E_\pi[h(X_1, X_2)] = \sum_{x\in\Lambda} \pi(x_1, x_2) \frac{1}{1+e^{-(x_1+x_2)}}.$$

Para que seja possível o uso do amostrador de Gibbs, é necessário determinar as distribuições condicionais completas de X. Desta forma

$$\pi_1(x_1 \mid X_2 = x_2) = \begin{cases} (2/3, 1/3) & \text{se} \quad x_2 = 0; \\ (1/3, 2/3) & \text{se} \quad x_2 = 1. \end{cases}$$

e

$$\pi_2(x_2 \mid X_1 = x_1) = \begin{cases} (2/3, 1/3) & \text{se} \quad x_1 = 0; \\ (1/3, 2/3) & \text{se} \quad x_1 = 1. \end{cases}$$

Assim, obtém-se todos os ingredientes necessários para implementação do algoritmo. O valor exato de I é 0,704 e a estimativa obtida pelo método foi $\hat{I} = 0{,}706$.

Exemplo 8. Encontre o mínimo da função $f(x, y) = -e^{(x^2+y^2)}$**,** fazendo uso do *simulated annealing*.

Para solução deste problema utilizou-se $x_0 = (2,3)$ como valor inicial, $c_0 = 1$ e $L_0 = 10$.

O ponto mínimo obtido pelo simulated annealing foi de $(-0{,}0057, -0{,}0042)$. A Figura 2 apresenta um gráfico indicando o *simulated annealiang* capturando o mínimo da função *f*.

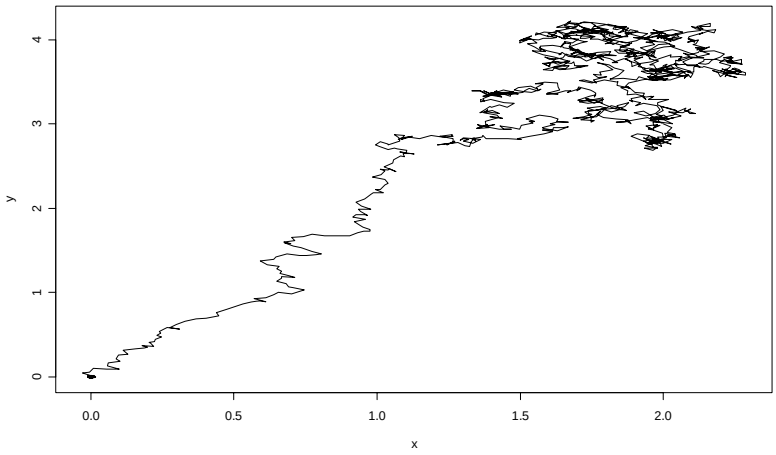

Figura 2. *Simulated annealing* capturando o mínimo da função *f*.

2.3. Introdução a Inferência Bayesiana

Estatística é uma área de conhecimento que lida com problemas nos quais quantidades aleatórias estão envolvidas. Particularmente na Inferência Estatística o interesse recai numa quantidade desconhecida e não observada (θ), onde θ assume valores no conjunto Θ. Essa quantidade pode ser um escalar, um vetor, ou mesmo uma matriz. O principal problema da área consiste em descrever a incerteza sobre θ (PAULINO et al., 2003).

Na Inferência Clássica θ é apenas um parâmetro desconhecido e a única fonte de informação relevante sobre este parâmetro é a informação probabilística de quantidades aleatórias "observáveis" associadas a ele. Por outro lado, na Inferência Bayesiana, a abordagem é um pouco diferente. A diferença essencial é que θ é pensado como uma quantidade aleatória, tal como os "observáveis" associados a ele, e assim outras fontes de informação são consideradas.

Denote por H a informação inicial disponível sobre θ. Assuma que essa informação possa ser expressa em termos probabilísticos através de uma distribuição de probabilidade em θ, genericamente denotada por $\pi(\theta \mid H)$. Se a informação contida em H e suficiente, então a descrição da incerteza sobre θ está completa (PAEZ E GAMERMAN, 2005).

Entretanto, na maioria dos casos a informação inicial H não é suficiente. Nesse caso a informação inicial precisa ser aumentada e a principal ferramenta utilizada nessa tarefa é a experimentação. Assuma que um vetor $X = (X_1, X_2, \ldots, X_n)$ de quantidades aleatórias relacionadas à θ pode ser observado. Este vetor proporciona informação adicional sobre θ. Assume-se também que a distribuição amostral de *X* dado θ e *H*, denotada por $f(x \mid \theta, H)$, é conhecida.

Desta forma, a informação sobre θ esta resumida pela distribuição $\pi(\theta \mid x, H)$. Pode-se utilizar o teorema de Bayes[5] para relacionar $\pi(\theta \mid x, H)$ com $\pi(\theta \mid H)$ e $f(x \mid \theta, H)$. De fato

$$\pi(\theta \mid x, H) = \frac{f(\theta, x \mid H)}{f(x \mid H)} = \frac{f(x \mid \theta, H)\pi(\theta \mid H)}{\int_{\Theta} f(x \mid \theta, H)\pi(\theta \mid H)d\theta}$$

Para simplificar a notação, vamos omitir a dependência em *H* visto que ela aparece em todos os termos. Além disso, observa-se que a função no denominador não

[5] Teorema de Bayes: Suponha que eventos $C_1, C_2, \ldots C_k$ formem uma partição de Ω e que suas probabilidades sejam conhecidas. Suponha ainda que para um evento *A*, se conheçam as probabilidades $P(A \mid C_i)$ para todo $i = 1,2,\ldots,k.$ Então para qualquer j,

$$P(C_j \mid A) = \frac{P(A \mid C_j)P(C_j)}{\sum_{i=1}^{k} P(A \mid C_i)P(C_i)}.$$

depende de θ, portanto é só uma constante em relação a $\pi(\theta \mid x)$. Assim, pode-se reescrever o teorema de Bayes da seguinte maneira

$$\pi(\theta \mid x, H) = k \cdot f(x \mid \theta)\pi(\theta) \propto f(x \mid \theta)\pi(\theta).$$

A equação acima proporciona uma regra para atualizar probabilidades sobre θ, partindo de $\pi(\theta)$ e chegando a $\pi(\theta \mid x)$. Daí a razão para chamar $\pi(\theta)$ de distribuição a "priori" e $\pi(\theta \mid x)$ de distribuição "a posteriori". A função $f(x \mid \theta)$ é conhecida como função de verossimilhança de θ correspondente à amostra observada $X = x$ (PAEZ e GAMERMAN, 2005).

A distribuição a posteriori descreve completamente a incerteza sobre θ após a observação dos dados, levando em conta a distribuição a priori. Isso representa uma distinção importante entre a Inferência clássica e a Bayesiana, visto que na abordagem clássica a incerteza sobre θ é descrita via o cálculo exato ou estimação (o que é mais comum) do erro padrão de um estimador pontual proposto de forma criteriosa para θ. Outra observação é que a distribuição a posteriori depende dos dados somente através de $f(x \mid \theta)$.

Uma prática comum é considerar o valor esperado ou o valor mediano da posteriori como o estimador pontual Bayesiano. Outro estimador comumente empregado é a moda da posteriori, isto é, o valor que maximiza $\pi(\theta \mid x)$. Esses estimadores são bastante intuitivos, visto que o máximo de informação que temos a respeito de um parâmetro a ser estimado é a sua distribuição a posteriori, e que qualquer distribuição de probabilidade pode ser sumarizada pontualmente por uma dessas três quantidades.

Os estimadores Bayesianos geralmente são descritos em termos do valor esperado condicional de alguma função h definida em Θ, ou seja, são obtidos por

$$E_{\theta|x}[h(\theta) \mid x] = \int_{\Theta} h(\theta)\pi(\theta \mid x)d\theta$$

Dessa forma, quando utiliza-se o valor esperado da posteriori como estimador, tem-se que avaliar $E_{\theta|x}(\theta \mid x)$. Se escolhe-se a mediana da posteriori tem-se a necessidade de avaliar

$$E_{\theta|x}(I_{[\theta \le c]} \mid x) = 0.5.$$

Quando o interesse recai na obtenção de uma estimativa por intervalo para θ, é necessário a resolução das seguintes equações

$$\begin{cases} E_{\theta|x}(I_{[\theta \le a]} \mid x) = l \\ E_{\theta|x}(I_{[\theta \le b]} \mid x) = l \end{cases}$$

para $0 < l < 0.5$.

Após apresentação de alguns estimadores, percebe-se que nem sempre é possível avaliar analiticamente as integrais envolvidas. Mesmo quando não envolve integrais, a abordagem bayesiana pode ser uma alternativa difícil quando a distribuição $\pi(\theta \mid x)$ é complexa. Por exemplo, quando utiliza-se a moda a posteriori como estimador, é necessário obter

$$\arg\max_{\theta \in \Theta} \{\pi(\theta \mid x)\}.$$

Ou ainda, na determinação de uma região de credibilidade pelo método da máxima densidade a posteriori, ou seja, na especificação do conjunto

$$C = \{\theta; \pi(\theta \mid x) \ge c\},$$

é requerido solucionar a equação $\pi(\theta \mid x) = c$ para o valor de *c* que satisfaz a equação

$$P(\theta \in C \mid x) = P(\pi(\theta \mid x) \ge c \mid x) = \alpha,$$

onde α é um nível de credibilidade pré-determinado. Além disso, em algumas situações, existe a necessidade da especificação completa da distribuição a posteriori. Contudo em poucas situações isso é possível sem o conhecimento da constaste normalizadora *k* da densidade a posteriori. Desta forma, nessas situações é necessário resolver a equação

$$k^{-1} = \int_{\Theta} f(x \mid \theta)\pi(\theta) d\theta = E_{\theta}[\pi(x \mid \theta)].$$

REFERÊNCIAS BIBLIOGRÁFICAS

BINDER, K.; HEERMANN, D. **Monte Carlo Simulation in Statistical Physics.** 3rd ed. Springer, Berlin, 1997.

BRÉMAUD, P. **Markov chains: Gibbs fields, Monte Carlo simulation and queues.** Springer-Verlag, New Yorq, 1999, 441p.

CASELLA, G.; GEORGE, E. Explaining the Gibbs Sampler. **The American Statistician**, v.46, p. 167-157, 1992.

CHIB, S.; GREENBERG, E. Understanding the Metropolis-Hastings Algorithm. **The American Statistician**, v.49, p.327-335, 1995.

EHLERS. R. S. **Métodos computacionalmente extensivos em estatística.** Versão nº 2. 2004. Disponível em http://leg.ufpr.br/~ehlers/notas/mci.pdf. Acessado em: novembro 2008.

GELFAND, A.; SMITH, A. Sampling based approaches to calculating marginal densities. **J. American Statist. Assoc.**, v.85 p.398–409, 1990.

GEMAN, S.; GEMAN, D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. **IEEE Trans. Pattern Anal. Mach. Intell.**, v.6 p.721–741, 1984.

GIANOLA, D.; FERNANDO R. Bayesian methods in animal breeding theory**, J. Anima. Sci.** v.63, p.217-244, 1986.

GRIMMETT, G.; STIRZAKER, D. **Probability and Random Processes.** 3rd ed. Oxford university press, New York, 1992, 541p.

GUO, S. W.;THOMPSON, E. A. Performing the exact test of Hardy–Weinberg proportion for multiple alleles**. Biometrics**, v.48, p.361–372, 1992.

HÄGGSTRÖM, O. **Finite Markov Chains and Algorithmic Applications.** Matematisk statistik, Chalmers högskola och Göteborgs universitet (January 2001).

HASTINGS, W. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. **Biometrika**, v. 57, p.97-109, 1970.

HOEL, P.; PORT, S.; STONE, C. **Introduction to Stochastic Processes.** Waveland Press, 1987. 203p.

HUBER, M., CHEN, Y., DINWOODIE, I., DOBRA, A., NICHOLAS, M. Monte Carlo algorithms for Hardy-Weinberg proportions. **Biometrics**, v.62, p.49-53, 2006.

KIRKPATRICK, S.; GELATT, C.; VECCHI, M. Optimization by Simulated Annealing. **Science**, v. 220, p.671-680, 1983.

METROPOLIS, N.; ROSENBLUTH, A.; ROSENBLUTH, M.; TELLER, A.; TELLER, E. Equations of state calculations by fast computing machines. **J.Chem. Phys.**, v.21, p. 1087–1092, 1953.

MEYN, S.; TWEEDIE, R. Computable Bounds for Convergence Rates of Markov Chains. **Annals of Applied Probability**, v.4, p.124-148, 1994.

PAEZ, M.; GAMERMAN, D. **Modelagem de Processos Espaço-Temporais**, 11a. ESTE 2005, 102p.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. **Estatística Bayesiana**. Lisboa: Fundação Calouste Gulbenkian, 2003, 446p.

PESKUN, P. Guidelines for chosing the transition matrix in Monte Carlo methods using Markov chains. **Journal of Computational Physics**, v. 40, p.327–344,1981.

PESKUN, P. Optimum Monte Carlo sampling using Markov chains. **Biometrika**, v.60, p.607–612, 1973.

ROBERT, C.; CASELLA, G. **A History of Markov Chain Monte Carlo - Subjective Recollections from Incomplete Data**. Disponível em http://arxiv.org/abs/0808.2902v1. Acessado em Julho 2008.

ROBERT, C.; CASELLA, G. **Monte Carlo Statistical Methods**. Springer-Verlag, New York 1999. 645p.

ROSENTHAL, J. Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo. **Journal of American Statistical Association**, v.90, p.558-566, 1995.

ROSS, S. **An Introduction to Probability Models**, 8th ed. Academic Press, 2003. 754p.

SCHUSTER, I.; CRUZ, C. D. **Estatística genômica - Aplicada a populações derivadas de cruzamentos controlados.** 2 ed. Editora UFV, Viçosa, 2008. 568p.

SILVA, J. P. **Uma abordagem bayesiana para mapeamento de QTLs utilizando o método MCMC com saltos reversíveis.** Piracicaba, SP: ESALQ\USP 2006. 81f. Dissertação (Mestrado Estatística e Experimentação Agronômica) – Escola Superior de Agricultura Luiz de Queiroz - USP.

YUAN, AO; BONNEY, G. E. Exact test of Hardy–Weinberg equilibrium by Markov chain Monte Carlo. **Mathematical Medicine and Biology**, v. 20, p.327–340, 2003.

CAPÍTULO 1

ESTIMAÇÃO DE FREQÜÊNCIA DE RECOMBINAÇÃO VIA ALGORITMO DE METROPOLIS-HASTINGS

RESUMO

O objetivo deste trabalho foi avaliar a eficiência da utilização do algoritmo de Metropolis-Hastings para estimar a freqüência de recombinação entre dois locos. Para avaliar a capacidade do algoritmo foi simulada uma população F_2, de natureza codominante, constituída de 200 indivíduos. Para esta população foi gerado um genoma com um grupo de ligação, com 50 cM de tamanho. O grupo de ligação possui três marcas, com uma distância variável entre marcas adjacentes. A estimação da freqüência de recombinação através do algoritmo de Metropolis-Hastings obteve resultados equivalentes aos encontrados pelo método da máxima verossimilhança. O algoritmo convergiu para a distribuição de interesse independentemente do valor inicial da cadeia. Além disso, o algoritmo de Metropolis-Hastings é de fácil implementação para o caso estudado, uma vez que não existe a necessidade de derivações complexas.

1. INTRODUÇÃO

O mapeamento genético facilita o trabalho de melhoramento, uma vez que uma ou mais marcas do genótipo podem estar associadas a um ou mais genes controladores de características qualitativas e quantitativas (*QTL*). Desse modo, tendo-se o genótipo mapeado, o trabalho de melhoramento pode ser otimizado, tanto na eficiência do programa, quanto na velocidade de obtenção de ganhos, pois é possível a realização de seleção com base nos marcadores (BHERING, 2008).

Após a primeira etapa do mapeamento, que consiste em selecionar marcadores moleculares que apresentam polimorfismo, é necessário obter estimativas da freqüência de recombinação entre pares de locos.

A estimação da freqüência de recombinação entre pares de locos é realizada por meio do método da máxima verossimilhança (FISHER, 1921). Este método é, sem dúvida, o mais popular dentre os métodos de estimação. No mapeamento genético, o método é empregado tanto na obtenção de estimativas de freqüência de recombinação quanto no cálculo de estimativas de parâmetros no mapeamento de *QTL*'s (SCHUSTER e CRUZ, 2008).

Um pré-requisito para o uso da técnica da máxima verossimilhança é o conhecimento da função de distribuição de probabilidade. Considera-se a distribuição multinomial na estimação da freqüência de recombinação.

Em muitas situações a solução da equação de verossimilhança pode ser obtida analiticamente. Entretanto em situações mais complexas a solução analítica é impraticável, devendo o resultado ser obtido a partir de aproximações numéricas (BOLFARINE e SANDOVAL, 2000).

Para contornar o problema, tem-se como opção o uso do método gráfico (SCHUSTER e CRUZ, 2008), o qual atribui diferentes valores para o parâmetro, dentro do intervalo 0 a 0,5, na função de verossimilhança, encontrando assim, o ponto de máximo da função. Podem-se encontrar casos onde o método gráfico deixa de ser interessante devido à dificuldade da análise visual em planos ou superfícies. Nestes casos, métodos iterativos como o de Newton-Raphson e Algoritmo *EM* (Esperança e Maximização) surgem como alternativas na estimação dos parâmetros. Entretanto, para utilização destes, é necessário atribuir um valor inicial, e uma escolha ruim pode fazer com que o algoritmo convirja para um valor que não seja o máximo da função de verossimilhança. Uma alternativa para contornar estes problemas e obter estimativas para a freqüência de recombinação é o uso do algoritmo de Metropolis-Hastings (1970), uma vez que nos métodos *MCMC*, independentemente do estado inicial da cadeia, o algoritmo sempre converge para a distribuição de interesse. Desta forma, tem-se como objetivo avaliar a eficiência do algoritmo de Metropolis-Hastings para obter estimativas das freqüências de recombinação entre pares de marcadores.

2. MATERIAL E MÉTODOS

Para gerar os dados utilizados na estimação da freqüência de recombinação entre pares de marcas, foi utilizado o módulo de simulação do aplicativo computacional GQMOL (CRUZ, 2007), que permite gerar informações sobre genomas, genótipos, genitores, indivíduos de diferentes tipos de populações e dados de características quantitativas. Foram simulados genomas parentais e uma população F_2 constituída por 200 indivíduos, com 1 grupo de ligação, construída com base em marcadores codominantes.

Foi tomada como referência uma espécie diplóide fictícia com $2n = 2x = 2$ cromossomos, cujo comprimento total do genoma foi de 50 cM. Foi gerado o genoma com nível de saturação de três marcas moleculares. O genoma simulado é apresentado na Figura 1.

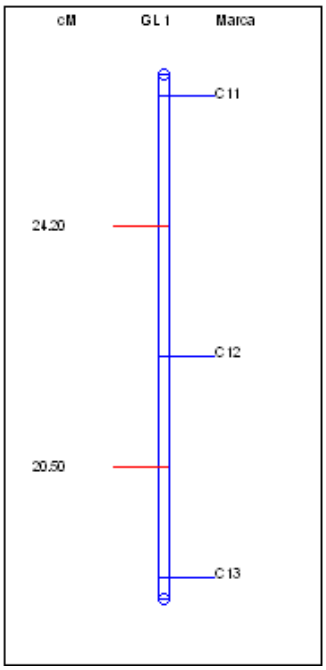

Figura 1 - Genoma simulado com saturação de 3 marcas no grupo de ligação com distância variável.

2.1. Método da Máxima Verossimilhança

Observando-se n indivíduos dentre os quais n_i pertencem à classe genotípica i, onde $i \in \{1,2,\ldots,9\}$ então o vetor aleatório $N = (n_1, n_2, \ldots, n_9)$ tem distribuição multinomial com parâmetros $n, p_1, p_2, \ldots, p_9$ e portanto

$$\pi(N \mid p) = \frac{n!}{n_1! n_2! \ldots n_9!} p_1^{n_1} p_2^{n_2} \ldots p_9^{n_9},$$

em que

$$p_1 = p_9 = \tfrac{1}{4}(1-r)^2; \qquad p_2 = p_4 = p_6 = p_8 = \tfrac{1}{2} r(1-r); \qquad p_3 = p_7 = \tfrac{1}{4} r^2;$$

$$p_5 = \tfrac{1}{2}(1-r)^2 + \tfrac{1}{2} r^2.$$

Apesar de haver nove classes, verifica-se que algumas delas possuem a mesma frequência esperada, ficando assim reduzidas a quatro.

A função de verossimilhança correspondente à amostra aleatória observada é dada por

$$L(r;N)=\prod_{i=1}^{n}\pi(n_i \mid r)$$

$$L(r;N)=\lambda\left[\tfrac{1}{4}(1-r)^2\right]^{n_1+n_9}\left[\tfrac{1}{2}r(1-r)\right]^{n_2+n_4+n_6+n_8}\left[\tfrac{1}{2}(1-r)^2+\tfrac{1}{2}r^2\right]^{n_5}\left[\tfrac{1}{4}r^2\right]^{n_3+n_7}$$

, em que

$$\lambda=\frac{n!}{n_1!n_2!\ldots n_9!}.$$

O logaritmo natural da função de verossimilhança de r, denotado por $l(r;N)=\ln L(r;N)$ é conhecido como função suporte e tem como objetivo facilitar a derivação. Denotando $A=n_1+n_9$, $B=n_2+n_4+n_6+n_8$, $C=n_5$ e $D=n_3+n_7$, tem-se

$$\ln L(r;N)=\ln(\lambda)+2A\ln(1-r)+B\ln(r(1-r)+C\ln(1-2r+2r^2)+2D\ln(r)$$

O estimador de máxima verossimilhança é obtido através da função escore definida como a derivada da função suporte

$$\frac{\partial l(r;N)}{\partial r}=\frac{-2A}{1-\hat{r}}+\frac{B(1-2\hat{r})}{\hat{r}(1-\hat{r})}+\frac{-2C(1-2\hat{r})}{1-2\hat{r}+2\hat{r}^2}+\frac{2D}{\hat{r}}=0.$$

Fazendo as simplificações necessárias, obtém-se a equação polinomial, cujas raízes devem ser obtidas.

$$2D+B-\hat{r}(2A+4B+2C-6D)+\hat{r}^2(4A+6B+6C+8D)-4\hat{r}^3(A+B+C+D)=0$$

As raízes deste polinômio podem ser encontradas analiticamente através do dispositivo prático de Briot-Ruffini. Este método baseia-se na lei da divisão, aplicada a um polinômio quando dividido por um binômio da forma $(x-a)$. Desta forma, sua utilização está condicionada ao conhecimento de ao menos uma raiz do polinômio. O conhecimento de possíveis raízes pode ser obtido utilizando o teorema de raízes racionais[6]. Entretanto, para grandes valores de a_0 e a_n existe uma infinidade de possíveis raízes, tornando a resolução analítica do problema uma tarefa árdua.

---

[6] Se o número racional $\frac{p}{q}$, com q e p, primos entre si, é uma raiz da equação polinomial com coeficientes inteiros

$$a_n x^n+a_{n-1}x^{n-1}+\ldots+a_2x^2+a_1x+a_0=0,$$

então, p é divisor de a_0 e q é divisor de a_n.

Assim, nestes casos o resultado deve ser obtido a partir de métodos gráficos, aproximações numéricas e até métodos *MCMC* (BOLFARINE e SANDOVAL, 2000).

2.2. Método Gráfico

Como discutido anteriormente, encontrar o ponto de máximo de uma função de verossimilhança pode ser tarefa bastante difícil. Nos casos mais simples, quando um ou dois parâmetros estão sendo estimados, os estimadores de máxima verossimilhança podem ser obtidos plotando-se em duas ou três dimensões o plano, ou superfície, de verossimilhança em função de parâmetros de verossimilhança (SCHUSTER e CRUZ, 2008).

Desta forma, onde não seja possível encontrar o máximo da verossimilhança analiticamente a análise gráfica dos valores de verossimilhança (atribuindo-se diferentes valores para o parâmetro) pode ser usada para encontrar o ponto máximo da função. O pico da curva indica o valor que, neste trabalho, é a frequência de recombinação que melhor explica os dados (SCHUSTER e CRUZ, 2008).

Entretanto, quando usa-se a função de verossimilhança os valores observados são muito pequenos e não permitem observar pequenas diferenças. Assim, na prática utiliza-se o valor do *LOD*, que é definido como o logaritmo na base 10, da razão entre a verossimilhança considerando o valor atual de r e a verossimilhaça para $r = 0{,}5$ (considerando ausência de ligação). Matematicamente

$$LOD = Log_{10}\left[\frac{L(r;N)}{L(r=0{,}5;N)}\right].$$

O logaritmo é utilizado, pois os resultados encontrados em ambas as funções de verossimilhança é muito pequeno, o que torna difícil o estabelecimento da superioridade do valor de probabilidade encontrado. Um valor de *LOD* igual 3 significa que o valor de verossimilhança, considerando o valor atual de *r*, é 1000 vezes maior que a verossimilhança no caso onde $r = 0{,}5$.

Apesar da grande utilidade deste método, em algumas situações o método deixa de ser interessante devido à dificuldade da análise visual em planos ou superfícies. Além disso, quando existe interesse de estimar mais de dois parâmetros o método gráfico torna-se impraticável.

2.3. Método iterativo de Newton-Raphson

O método de Newton-Raphson é um algoritmo apropriado para encontrar raízes (ou zeros) de funções. O método baseia-se na expansão em série de Taylor, utilizando a seguinte igualdade para o ponto em que $\theta = \theta_0$.

$$f(\theta) = f(\theta_0) + \frac{(\theta - \theta_0)}{1!} f^{'1}(\theta_0) + \frac{(\theta - \theta_0)^2}{2!} f^{'2}(\theta_0) + \ldots + \frac{(\theta - \theta_0)^n}{n!} f^{'n}(\theta_0) + R(\theta),$$

onde

$$R_n(\theta) = \frac{f^{'(n+1)}(c)}{(n+1)}(\theta - \theta_0)^{n+1}.$$

Ignorando os termos com grau igual e superior a dois, cujos valores, por serem muitos pequenos, podem ser negligenciados, tem-se que a função de θ quando $\theta = \theta_0$ é:

$$f(\theta) = f(\theta_0) + \frac{(\theta - \theta_0)}{1!} f^{'1}(\theta_0).$$

Para $f(\theta) = 0$, pode-se obter θ pela seguinte equação:

$$\theta = \theta_0 - \frac{f(\theta_0)}{f^{'1}(\theta_0)}.$$

A obtenção da raiz da função escore (ou maximização da função suporte) através do método de Newton-Raphson para o problema de estimação da freqüência de recombinação é feita da seguinte forma:

1. Fixa-se um valor real $\xi > 0$;
2. Atribui-se um valor arbitrário r_0 para r;
3. Para $k \geq 0$, faça

$$r_{k+1} = r_k - \frac{l^{'}(r;N)}{l^{''}(r;N)}$$

4. O processo iterativo se encerra quando $|r_k - r_{k+1}| < \xi$. Caso contrário volta para o passo anterior.

A seqüência $(r_k)_{k>0}$ converge para $\hat{r}$ quando $k \to \infty$, se r_0 é escolhido próximo de r. Para obtenção de um valor inicial (r_0), pode-se fazer uso do gráfico de $l^{'}(r;N)$ ou $l(r;N)$ para auxiliar na escolha de r_0.

2.4. Algortimo de Metropolis-Hastings

Este é sem dúvida o mais importante dos métodos *MCMC*, pois todos os outros são casos especiais dele. O algoritmo foi inicialmente proposto por METROPOLIS (1953) e generalizado por HASTINGS (1970). A idéia do algoritmo é simular uma cadeia de Markov $(X_n)_{n\geq 0}$ em Λ com distribuição estacionária π. Isto é, valores de Λ após um tempo suficientemente longo de simulação são amostrados de uma distribuição aproximadamente igual a π.

Neste algoritmo um valor é gerado a partir de uma distribuição auxiliar $q(r, r^*)$ e aceito com uma dada probabilidade $\alpha(r, r^*)$. Esse mecanismo de correção garante a convergência da cadeia para a distribuição de equilíbrio, que, neste caso, é a distribuição de interesse (EHLERS, 2003; PAULINO et al., 2003). Suponha que no instante *t* a cadeia esteja no estado r e um valor r^* é gerado de uma distribuição proposta $q(r, r^*)$. O novo valor r^* é aceito com probabilidade

$$\alpha\left(r, r^{'}\right) = \min\left(\frac{\pi(r^*)q(r^{*'}, r)}{\pi(r)q(r, r^*)}, 1\right),$$

onde π é a distribuição de interesse.

Uma característica importante é que só é necessário conhecer π parcialmente, isto é, a menos de uma constante. No caso da estimação da freqüência de recombinação pode-se negligenciar o valor de λ. Isto é fundamental em aplicações Bayesianas, onde não se conhece completamente a distribuição a posteriori.

O algoritmo de Metropolis-Hastings para o caso da estimação da freqüência de recombinação pode ser descrito como:

Algoritmo 1.

1. Escolhe-se uma função de transição auxiliar $q(r, r^*)$;
2. Escolha $X_0 \in \Lambda$;
3. Para $n \geq 0$ e $X_n = r$ simule $X_{n+1} \sim q(r, r^*)$ e lance uma $U \sim (0,1)$. Supondo que $X_{n+1} = r^*$, faça

$$X_{n+1} = \begin{cases} r^* & \text{se } U < \alpha(r, r^*); \\ r & \text{caso contrário.} \end{cases}$$

4. $n \leftarrow n+1$ e retorne para o passo 3.
5. Interrompa o processo utilizando um critério de parada.

Neste trabalho tomou-se a distribuição $U(0,1)$ como função de transição auxiliar. Por se tratar de uma distribuição de probabilidade simétrica, a probabilidade de aceitação se simplifica para

$$\alpha(r, r^*) = \min\left(1, \frac{\pi(N \mid r^*)}{\pi(N \mid r)}\right)$$

$$\alpha(r, r^*) = \min\left\{1, \left[\frac{(1-r^*)}{(1-r)}\right]^{2A} \left[\frac{2r(1-r^*)}{2r(1-r)}\right]^{B} \left[\frac{2(1-r^*)^2 + 2r^{*2}}{2(1-r)^2 + 2r^2}\right]^{C} \left[\frac{r^{*2}}{r^2}\right]^{D}\right\}$$

.

2.5. Intervalos de confiança para freqüência de recombinação

Os intervalos de confiança para a freqüência de recombinação podem ser construídos das seguintes maneiras.

- Baseado na técnica de bootstrap: São obtidas *b* estimativas de *r* utilizando reamostragem. O intervalo de confiança que contenha 95% dos valores de *r* será aquele limitado pelos percentis 2,5 e 97,5 da distribuição empírica dos dados;
- Baseado na curva logaritmo da verossimilhança (intervalo suporte): Obtido a partir da curva dos valores de *LOD*, em função dos valores da freqüência de recombinação entre pares de locos. O intervalo é delimitado pelos pontos correspondentes da freqüência de recombinação obtidos a partir da redução em uma unidade do *LOD* representativo do ponto máximo da curva.

Segundo SCHUSTER e CRUZ (2008), estudos realizados através de simulação de dados têm mostrado que os intervalos de confiança obtidos pela metodologia bootstrap são os que possuem melhor amplitude e maior probabilidade de acerto.

3. RESULTADOS E DISCUSSÃO

A partir das informações relativas aos locos gênicos (A/a, B/b e C/c), obtidas da população F_2 simulada derivada de um duplo-hererozogoto em aproximação (Tabela 1), obtem-se, através do método da máxima verossimilhança, os polinômios que possibilitam a estimação dos valores das freqüências de recombinação de r_{AB} e r_{BC}, que representam a distância entre os locos A/a e B/b e B/b e C/c, respectivamente. São eles

$$-800\hat{r}_{AB}^3 + 1120\hat{r}_{AB}^2 - 584\hat{r}_{AB} + 92 = 0 .$$

$$-800\hat{r}_{BC}^3 + 1062\hat{r}_{BC}^2 - 550\hat{r}_{BC} + 75 = 0 .$$

Tabela 1 - Valores observados e esperados da segregação de dois locos codominantes

Genótipo	Obs.	Genótipo	Obs.	Frequência Esperada genes ligados
AABB	28	BBCC	34	$\frac{1}{4}(1-r)^2$
AABb	14	BBCc	19	$\frac{1}{2}r(1-r)$
AAbb	5	BBcc	2	$\frac{1}{4}r^2$
AaBB	26	BbCC	21	$\frac{1}{2}r(1-r)$
AaBb	68	BbCc	56	$\frac{1}{2}(1-r)^2 + \frac{1}{2}r^2$
Aabb	30	Bbcc	15	$\frac{1}{2}r(1-r)$
aaBB	1	bbCC	0	$\frac{1}{4}r^2$
aaBb	10	bbCc	16	$\frac{1}{2}r(1-r)$
Aabb	18	bbcc	37	$\frac{1}{4}(1-r)^2$

Obs.: Observação.

Através do teorema das raízes racionais, tomando como base os valores de a_0 e a_n, percebe-se a existência de inúmeras possíveis raízes para estes polinômios, o que torna a solução analítica do problema pouco interessante.

A partir do método gráfico constata-se que os valores de $\hat{r}_{AB}$ e $\hat{r}_{BC}$ são de 0,271 e 0,205, isto é, 27,1 e 20,5 centimorgans, respectivamente. Estes valores são correspondes aos valores de *LOD* iguais a 9,317 e 21,12. As Figuras 3 e 4 apresentam graficamente todos os valores de *LOD* obtidos através de valores de *r* atribuídos entre 0 e 0,5 com incremento de 0,01.

As soluções obtidas através do método iterativo de Newton-Raphason, considerando um valor de $\xi = 0{,}001$, para ambos pares de locos, foram de $\hat{r}_{AB} = 0{,}271$ e $\hat{r}_{BC} = 0{,}205$, valores exatamente iguais ao resultado obtido pelo método gráfico. As Tabelas 2 e 3 apresentam os valores obtidos a partir do processo iterativo, admitindo-se como valor inicial $r_0 = 0{,}25$, para ambos os pares de locos. Verifica-se a partir destas tabelas que o método de Newton-Raphason convergiu com apenas três iterações.

A Figura 2 apresenta o mapa de ligação construído com base nas estimativas obtidas por estes métodos.

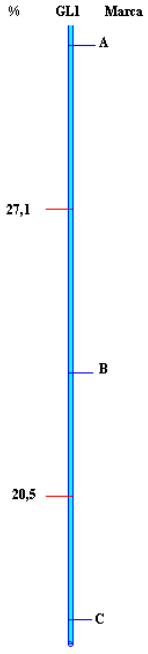

Figura 2 – Mapa de ligação com saturação de 3 marcas no grupo de ligação com distância variável.

Tabela 2 – Valores da freqüência de recombinação ($\hat{r}_{AB}$) obtidos até a convergência do algoritmo de Newton-Raphson

Iteração	r_{AB}	$\xi = \mid r_k - r_{k+1} \mid$
0	0,2500	9,1828
1	0,2701	9,3174
2	0,2714	9,3179
3	0,2714	9,3179

Tabela 3 – Valores da freqüência de recombinação ($\hat{r}_{BC}$) obtidos até a convergência do algoritmo de Newton-Raphson

Iteração	r_{BC}	$\xi = \mid r_k - r_{k+1} \mid$
0	0,2500	20,3600
1	0,1990	21,1100
2	0,2050	21,1200
3	0,2050	21,1200

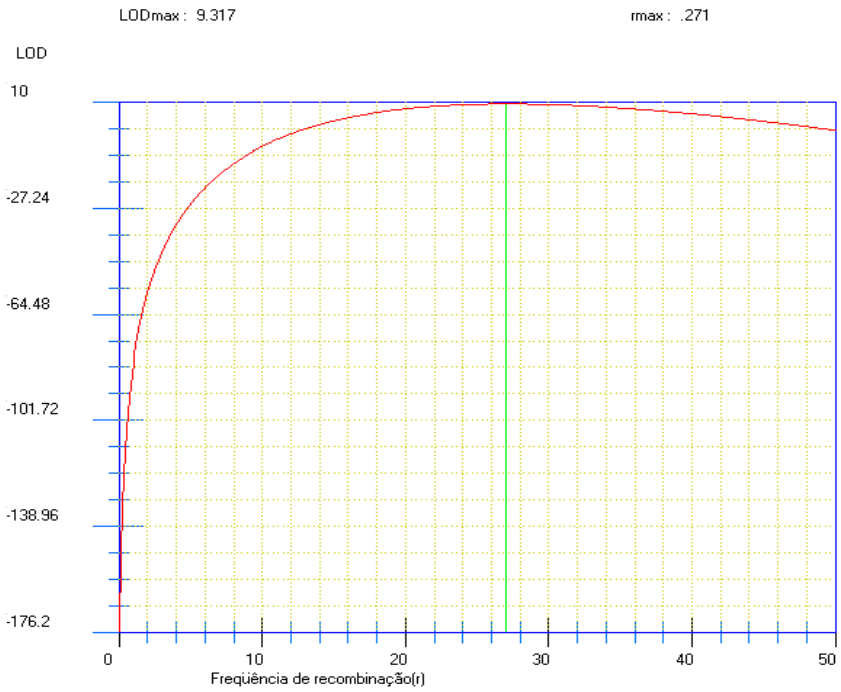

Figura 3 - Valores de *LOD* obtidos através de valores de r_{AB} entre 0 e 0,5 com incremento de 0,01.

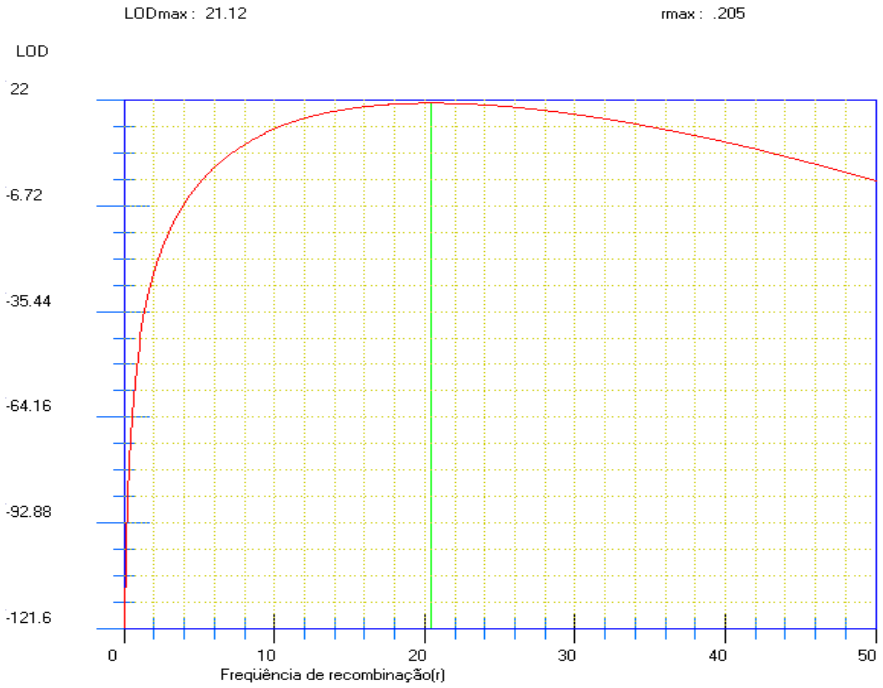

Figura 4 - Valores de *LOD* obtidos através de valores de r_{BC} entre 0 e 0,5 com incremento de 0,01.

Os intervalos de confiança obtidos através da técnica bootstrap para cada par de locos foram de $IC_{95\%}(r_{AB}) = [0{,}232;\ 0{,}315]$ e $IC_{95\%}(r_{BC}) = [0{,}161;0{,}259]$. Já os intervalos suportes (confiança) baseados na curva do *LOD*, são apresentados nas Figuras 5 e 6 e são dados numericamente por $IC(r_{AB}) = [0{,}160;0{,}256]$ e $IC(r_{BC}) = [0{,}216;0{,}336]$.

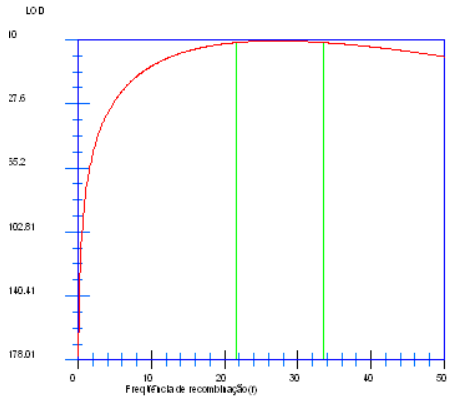

Figura 5 – Intervalo suporte (r_{AB}).

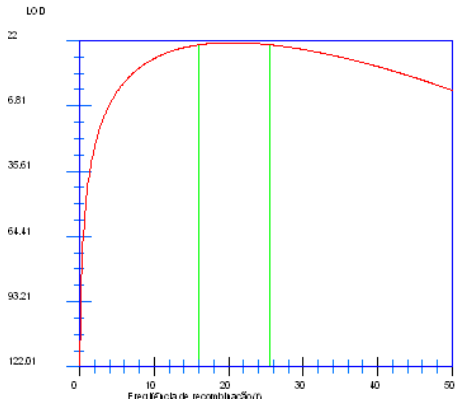

Figura 6 - Intervalo suporte (r_{BC}).

O algoritmo de Metropolis-Hastings foi implementado utilizando-se o programa *R* versão 2.7.1 (R Development Core Team). Considerou-se, em todas as análises efetuadas, um número fixo de 2.000 iterações. Além disso, desprezaram-se as 50 primeiras iterações para fazer inferência sobre a freqüência de recombinação. Para obtenção dos valores de $\hat{r}_{AB}$ e $\hat{r}_{BC}$ foram realizadas 100 simulações com 2000 iterações cada. Nas Figuras 8, 10, 12 e 14 são apresentados os histogramas dos valores estimados da freqüência de recombinação, para cada par de locos.

Com o objetivo de mostrar a diminuição da influência do valor inicial da cadeia ao longo do processo de simulação, foram utilizados, como valores iniciais da cadeia, para cada par de locos, os valores de $r_0 = 0{,}05$ e $r_0 = 0{,}45$. Percebe-se a partir das Figuras 7, 9, 11 e 13 que a cadeia converge rapidamente para um valor em torno da média da cadeia.

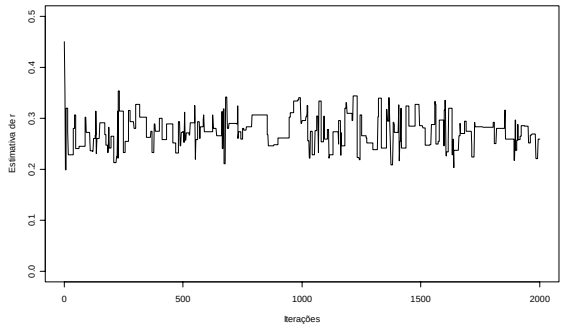

Figura 7 – Cadeia de 2000 valores gerados de r_{AB} ($r_0 = 0{,}45$).

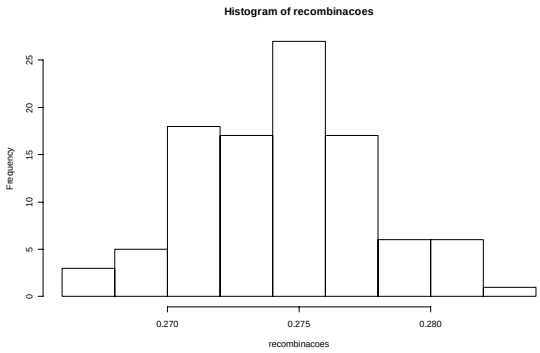

Figura 8 - Histograma dos $\hat{r}_{AB}$ obtidos em 100 simulações com 2000 iterações cada ($r_0 = 0{,}45$).

Os valores de $\hat{r}_{AB}$ e $\hat{r}_{BC}$ obtidos a partir de 100 simulações com 2000 iterações para o caso onde $r_0 = 0{,}45$ foram de 0,274 e 0,207, respectivamente. Isto é 27,4 e 20,7 centimorgans.

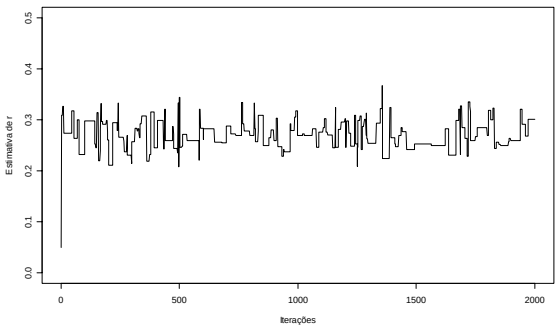

Figura 9 – Cadeia de 2000 valores simulados de r_{AB}, com $r_0 = 0{,}05$.

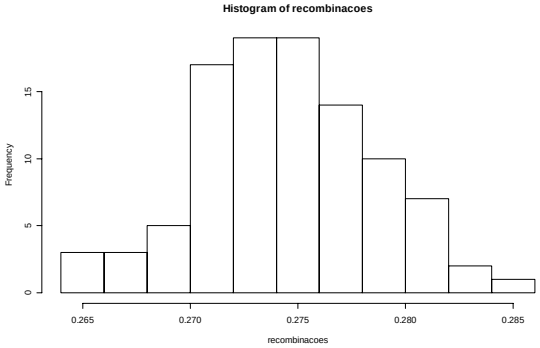

Figura 10 - Histograma dos $\hat{r}_{AB}$ obtidos em 100 simulações com 2000 iterações cada ($r_0 = 0{,}05$).

Já para o caso onde $r_0 = 0{,}05$, os valores de $\hat{r}_{AB}$ e $\hat{r}_{BC}$ obtidos foram os mesmos encontrados para o caso anterior, isto é, 27,4 e 20,7 centimorgans, respectivamente.

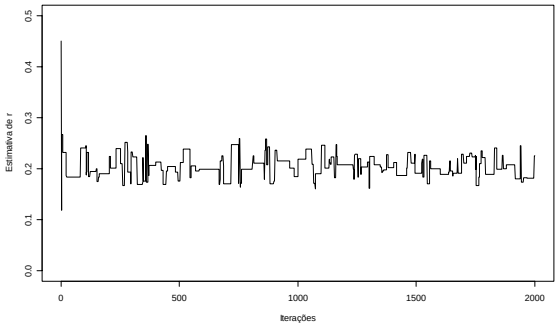

Figura 11 - Cadeia de 2000 valores gerados de r_{BC} ($r_0 = 0{,}45$).

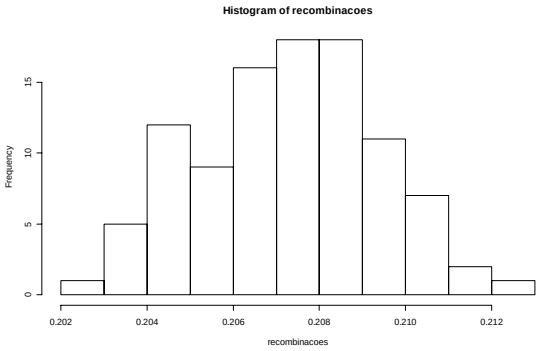

Figura 12 - Histograma dos $\hat{r}_{BC}$ obtidos em 100 simulações com 2000 iterações cada ($r_0 = 0{,}45$).

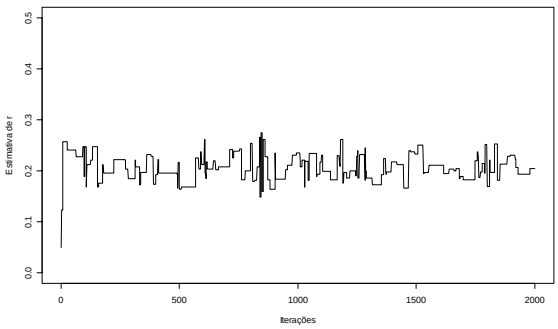

Figura 13 – Cadeia de 2000 valores simulados de r_{BC}, com $r_0 = 0{,}05$.

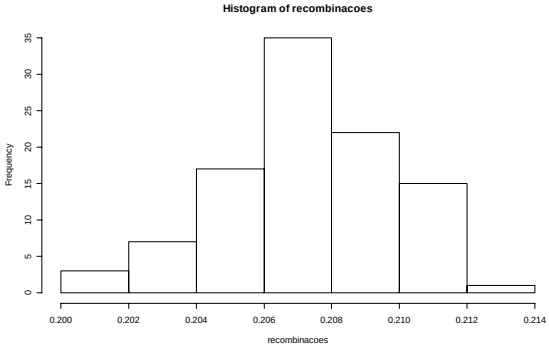

Figura 14 - Histograma dos $\hat{r}_{BC}$ obtidos em 100 simulações com 2000 iterações cada ($r_0 = 0{,}05$).

Percebe-se que os valores estimados da freqüência de recombinação através do algoritmo de Metropolis-Hastings encontram-se dentro dos intervalos de confiança bootstrap e intervalo suporte calculados anteriormente.

4. CONCLUSÕES

1. O algoritmo de Metropolis-Hastings proporcionou resultados semelhantes aos obtidos pelo método da máxima verossimilhança;
2. O algoritmo de Metropolis-Hastings é de fácil implementação para o caso da estimação da freqüência de recombinação, uma vez que não existe a necessidade de derivações complexas.

REFERÊNCIAS BIBLIOGRÁFICAS

BHERING, L. L.; CRUZ, C. D.; GOD, P. I. V. G. Estimativa de frequência de recombinação no mapeamento genético de famílias de irmãos completos. **Pesquisa Agropecuária Brasileira**, v. 43, p. 363-369, 2008.

BOLFARINE, H.; SANDOVAL, M. C. **Introdução a inferência estatística.** 1 ed. Editora SBM, Rio de Janeiro-RJ, 2001, 125p.

CRUZ, C. D. **Programa para análises de dados moleculares e quantitativos – GQMOL**. Viçosa: UFV, 2007.

EHLERS. R. S. **Métodos computacionamente extensivos em estatística**. Versão nº 2. 2004. Disponível em http://leg.ufpr.br/~ehlers/notas/mci.pdf. Acessado em: novembro 2008.

FISCHER, R. A. On the mathematical fundations of statistics. **Phil. Trans. Royal Soc. Lond**, v. 222, p.309-368, 1921.

HASTINGS, W. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. **Biometrika**, v.57, p.97-109, 1970.

METROPOLIS, N.; ROSENBLUTH, A.; ROSENBLUTH, M., TELLER, A.; TELLER, E. Equation of State Calculations by Fast Computing Machine. **Journal of Chemical Physics**, v.21, p.1087- 1091, 1953.

PAULINO, C. D.; TURKMAN, M. A.; MURTEIRA B. **Estatística Bayesiana**. Lisboa: Fundação Calouste Gulbenkiman, 2003. 429p.

R DEVELOPMENT CORE TEAM (2007). **R: A language and environment for statistical computing**. R Fundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, disponível em: http://r-project.org

SCHUSTER, I.; CRUZ, C. D. **Estatística genômica - Aplicada a populações derivadas de cruzamentos controlados**. 2 ed. Editora UFV, Viçosa, 2008. 568p.

CAPÍTULO 2

O USO DO ALGORITMO DO *SIMULATED ANNEALING* NA CONSTRUÇÃO DE MAPAS DE LIGAÇÃO

RESUMO

Este trabalho teve como objetivo avaliar e comparar a eficiência, no estabelecimento da melhor ordem de ligação na construção de mapas genéticos, do algoritmo *simulated annealing***.** Os resultados obtidos foram comparados com o algoritmo de delineação rápida em cadeia. Para avaliar e comparar a capacidade do algoritmo foram simuladas três populações F_2, com marcadores codominantes de tamanhos 50, 100 e 200 respectivamente. Para cada população foi estabelecido um genoma com quatro grupos de ligação, com 100 cM de tamanho cada. Os grupos de ligação possuem 51, 21, 11 e 6 marcas, respectivamente, com uma distância de 2, 5, 10 e 20 cM entre marcas adjacentes, ocasionando diferentes graus de saturação. Para grupos de ligação muito saturados, distância adjacente entre marcas de 2 cM, e com maior número de marcas, 51 marcas, o método baseado em simulação estocástica, *simulated annealing*, apresentou ordens com distância (*SARF*) iguais ou menores que o método delineação rápida em cadeia. Nos demais casos, ambos os métodos foram equivalentes, apresentando mesmo valor de *SARF*.

1. INTRODUÇÃO

A construção criteriosa de mapas genéticos é de fundamental importância para o mapeamento de locos que controlam caracteres quantitativos (*QTL*), e é uma das aplicações com maior potencial de uso para o entendimento da arquitetura genética desses caracteres. Uma das etapas mais importantes na construção de mapas de ligação é a ordenação dos marcadores genéticos dentro de cada grupo de ligação (MOLLINARI et al., 2008).

Uma vez estimadas as frações de recombinação entre cada par de marcadores, e já discriminados os grupos de ligação, deve-se determinar a melhor ordem para os marcadores dentro de cada grupo (CARNEIRO e VIERA, 2002). Quando se tem apenas dois marcadores, apenas uma ordem é possível. No caso de três, WEIR (1996) sugere que a ordenação seja feita considerando, como critério de avaliação, a menor soma dos coeficientes de recombinação adjacentes (*SAR – sum of adjacent recombination coefficients).* O problema surge quando o interesse é estabelecer a melhor ordem a partir de um grande número de marcas, pois tem-se, para n marcadores, $\frac{n!}{2}$ possíveis ordens (SHUSTER e CRUZ, 2008). Nota-se então que, para grande número de locos, este não é um problema trivial de ser resolvido, uma vez que, no caso de 12 marcas, por exemplo, tem-se 239500800 possíveis ordens tornando este problema impossível de ser resolvido analiticamente.

Para este propósito vários métodos são citados na literatura: delineação rápida em cadeia (DOERGE, 1996); seriação (BUETOW e CHAKRAVARTI, 1987a;b); *simulated annealing* (KIRKPATRICK et al., 1983); ramos e conexões (THOMPSON, 1987). A delineação rápida em cadeia consiste, com base na matriz de recombinações de todos os

pares de marcas, na obtenção de uma ordem preliminar para os locos. Em seguida, tentam-se inversões sucessivas em triplas marcas, a fim de minimizar a soma das recombinações adjacentes (*SARF*). A seriação é um método simples, no qual um conjunto de regras é proposto com base nas frações de recombinação entre dois locos (LIU, 1998). O método de ramos e conexões é firmado na estrutura das árvores, sendo o número de recombinantes calculado para cada ramo. O método de simulação estocástica, s*imulated annealing* é na verdade um conhecido método *MCMC* (especificamente o Algoritmo de Metropolis-Hastings), modificado de forma a se tornar um algoritmo de otimização. Para obtenção da solução de ordenação através destes métodos, vários critérios podem ser utilizados: soma mínima das frações de recombinação adjacentes (*SARF - sum of adjacent recombination fractions*) (FALK, 1989); produto mínimo das frações de recombinação adjacentes (*PARF* - "*product of adjacent recombination fractions*") (WILSON, 1988) e soma máxima dos LOD Scores adjacentes (*SALOD* - "*sum of adjacent LOD Scores*") (WEEKS e LANGE, 1987).

Os diversos softwares existentes utilizam diferentes métodos para solução do problema de ordenação, dentre eles podem-se citar: PGRI (LU e LIU, 1995) utiliza os métodos s*imulated annealing* e/ou ramos e conexões - GMENDEL (LIU e KNAPP, 1990) utiliza o s*imulated annealing* - CARTHAGENE (SCHIEX e GASPIN, 1997) possui como opções para solução do problema de ordenação, os métodos s*imulated annealing* e algoritmos genéticos - GQMOL (CRUZ, 2007) faz uso do método da delineação rápida em cadeia. Apesar da grande quantidade de métodos com propósito de fornecer a solução do problema de ordenação, é raro trabalho que forneça uma comparação entre estes métodos. MOLLINARI et al. (2008), compararam os métodos delineação rápida em cadeia e seriação, e concluiu que eles apresentam resultados semelhantes.

Assim, este trabalho tem como objetivo avaliar a eficiência, no estabelecimento da melhor ordem de ligação na construção de mapas genéticos, dos métodos *simulated annealing* e comparar este algoritmo com o método de delineação rápida em cadeia**.** O problema de ordenação de marcas é descrito como o problema do menor caminho. O trabalho é desenvolvido de forma que qualquer pesquisador interessado seja capaz de reproduzi-lo e utiliza-lo em sua pesquisa.

2. MATERIAL E MÉTODOS

Para apresentar uma situação concreta e comparar a eficiência dos métodos foram simuladas três populações F_2, com marcadores codominantes de tamanhos 50, 100 e 200, respectivamente. Para cada população, foi gerado um genoma com quatro grupos de ligação, com 100 cM de tamanho cada. Os grupos de ligação possuem 51, 21, 11 e 6 marcas, com uma distância de 2, 5, 10 e 20 cM entre marcas adjacentes, ocasionando diferentes graus de saturação. Os grupos são compostos por:

- Primeiro grupo de ligação: marcador 1 (m_1), marcador 2 (m_2), ..., marcador 51 (m_{51}), com intervalos entre marcas adjacentes de 2 cM;
- Segundo grupo de ligação: marcador 52 (m_{52}), marcador 53 (m_{53}),..., marcador 72 (m_{72}), com intervalos entre marcas adjacentes de 5 cM;
- Terceiro grupo de ligação: marcador 73 (m_{73}), marcador 74 (m_{74}),..., marcador 83 (m_{83}), com intervalos entre marcas adjacentes de 10 cM;
- Quarto grupo de ligação: marcador 84 (m_{84}), marcador 85 (m_{85}),..., marcador 89 (m_{89}), com intervalos entre marcas adjacentes de 20 cM.

Utilizou-se o módulo de "Simulação de genoma complexo" do aplicativo computacional GQMOL (CRUZ, 2007) para obtenção destas populações.

2.1. Descrição do problema

O problema de ordenação de marcas, fazendo as analogias necessárias ao problema do menor caminho, pode ser descrito da seguinte forma: seja $I = \{1, \ldots, k\}$ um conjunto de índices e seja $M = \{m_i : i \in I\}$ um conjunto de marcadores indexados por I.

Considere que D_{ij} representa a distância entre o marcador m_i e o marcador m_j e defina Λ como o conjunto de todas as possíveis permutações dos elementos do conjunto M. Um elemento de M será denotado por $x_m = (m_{\sigma_1}, \ldots, m_{\sigma_k})$, onde $(\sigma_1, \ldots \sigma_k)$ é uma permutação dos elementos do conjunto I. Uma permutação $x_m \in \Lambda$ pode ser entendida como uma ordem para passar por todos os marcadores. O problema consiste em encontrar uma ordem que minimize a distância necessária para passar por todos os marcadores apenas uma única vez, sem necessidade de retornar à origem.

Seja f a função que associa a cada ordem $x_m \in \Lambda$ a distância total percorrida (*SARF - sum of adjacent recombination frequencie*), ou seja, $f(x_m) = \sum_{i=1}^{K-1} D_{\sigma_i, \sigma_{i+1}}$. O interesse é encontrar a ordem $x_m \in \Lambda$ que minimiza a função f. Para obter uma aproximação numérica da solução deste problema, serão utilizados os algoritmos s*imulated annealing* e delineação rápida em cadeia.

2.2. *Simulated annealing*

O *simulated anneling* é uma pequena modificação no conhecido algoritmo *MCMC* de Metropolis-Hastings (1970), que o transforma em um algoritmo de otimização conhecido como s*imulated annealing* (KIRKPATRICK, et al. 1983). A idéia fundamental deste método é emprestada da física. Em física da matéria condensada, annealing é um processo térmico utilizado para minimizar a energia livre de um sólido. Informalmente o processo pode ser descrito em duas etapas: (i) aumentar a temperatura do sólido até ele derreter; (ii) Diminuir lentamente a temperatura até as partículas se organizarem no estado de mínima energia do sólido. Esse processo físico pode ser simulado no computador usando o algoritmo de Metropolis. Suponha que o estado atual do sólido é x, e que a energia desse estado é $H(x)$. Um estado candidato y, de energia $H(y)$, é gerado aplicando uma pequena perturbação no estado x. A regra de decisão para aceitar o estado candidato utiliza a seguinte probabilidade

$$\alpha_T(x, y) = \min\left(1, \exp\left(-\frac{H(y) - H(x)}{T}\right)\right),$$

onde *T* denota a temperatura. Se o resfriamento é realizado lentamente, o sólido atinge o equilíbrio térmico a cada temperatura. Do ponto de vista de simulação, isso significa gerar muitas transições a uma certa temperatura T (ROBERT e CASELLA, 2004).

Para o problema de ordenação de marcadores, faz-se a seguinte analogia:

- As soluções do problema de ordenação (otimização), ou seja, os elementos $x_m \in \Lambda$ são equivalentes aos estados físicos x;
- A função $f : \Lambda \to \Re$ (*SARF*) é equivalente à função energia do sólido, $H(x)$;
- Uma ordem candidata y_m de distância dada por $f : \Lambda \to \Re$ é equivalente a um estado candidato y de energia $H(y)$;
- Um parâmetro de controle $c > 0$ é equivalente à temperatura.

Seja x_{m_0} uma ordem inicial, c_0 o parâmetro de controle inicial e L_0 o número inicial de iterações utilizadas para um mesmo valor de c_0. O s*imulated annealing* pode ser descrito da seguinte forma:

1. Escolha $n = 0$, $x_m = x_{m_n} \in \Lambda$, c_0 e L_0;
2. Faça *i* de 1 até L_n
3. Gere y_m na vizinhança de x_m e gere uma variável aleatória $X \sim U(0,1)$;
4. Se $f(y_m) \leq f(x_m)$, então $x_m \leftarrow y_m$;
5. Se $f(y_m) > f(x_m)$ e $U < \exp\left(-\dfrac{f(y_m) - f(x_m)}{c_n}\right)$, então $x_m \leftarrow y_m$;
6. Fim do faça;
7. $n \leftarrow n + 1$;
8. Defina c_n e L_n, e volte até o passo 2 até um critério de parada.

Onde L_n é o número de transições da cadeia em cada temperatura (c_n).

2.3. Delineação rápida em cadeia

O algoritmo da delineação rápida em cadeia (DOERGE, 1996), consiste numa maneira simples para a ordenação de marcadores moleculares dentro dos grupos de ligação. Este algoritmo pode ser descrito da seguinte forma:

1. Verifica-se qual par de marcadores (m_i, m_j) possui a menor estimativa de frações de recombinação entre cada par de marcadores. Esses marcadores iniciarão a cadeia;
2. Verifica-se qual é o marcador não mapeado (m_k) que apresenta a menor estimativa de frações de recombinação com um dos marcadores terminais. Posiciona-se este marcador ao lado daquele com o qual apresentou a menor fração de recombinação;
3. Repete-se o procedimento até que todos os marcadores sejam adicionados à cadeia;
4. Em seguida, tentam-se inversões sucessivas em duplas e triplas marcas, a fim de minimizar a soma das recombinações adjacentes (*SARF*).

Os resultados obtidos pelo método do s*imulated annealing* foram comparados com o fornecido pelo método delineação rápida em cadeia, o critério utilizado para obtenção da solução é a soma mínima das frações de recombinação adjacentes (*SARF - sum of adjacent recombination fractions*).

3. RESULTADOS E DISCUSSÃO

Os resultados obtidos a partir do software GQMOL, que encontra a solução do problema através do método delineação rápida em cadeia, são apresentados nas figuras 2, 3 e 4.

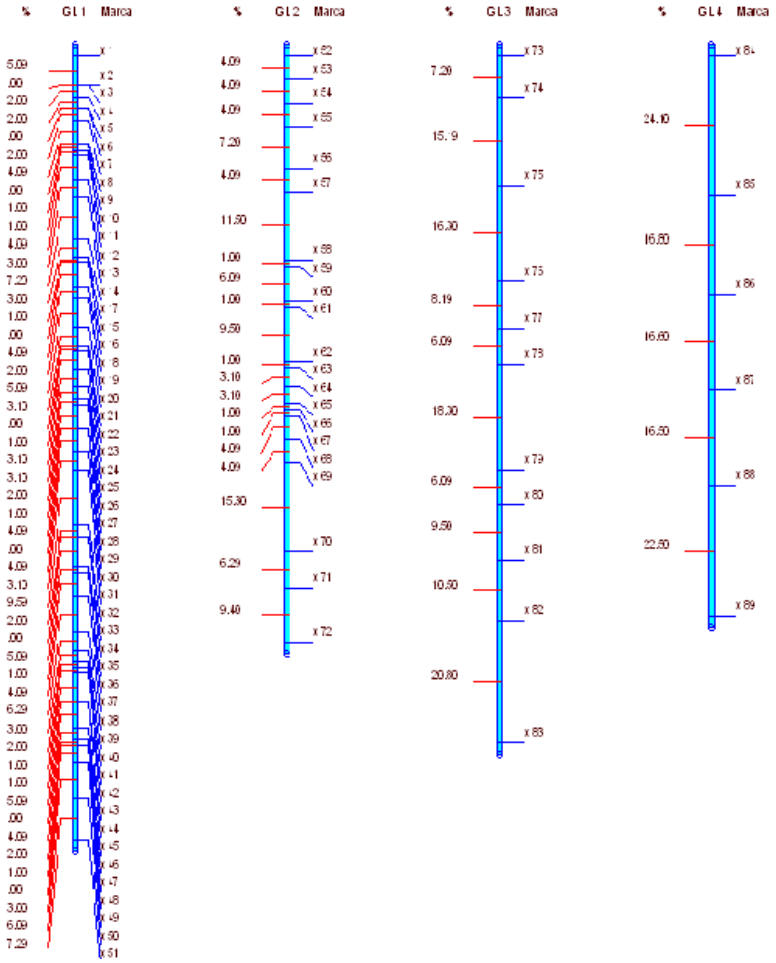

Figura 2 - Solução obtida para o problema de ordenação de marcas através do método delineação rápida em cadeia (GQMOL), para população com 50 indivíduos.

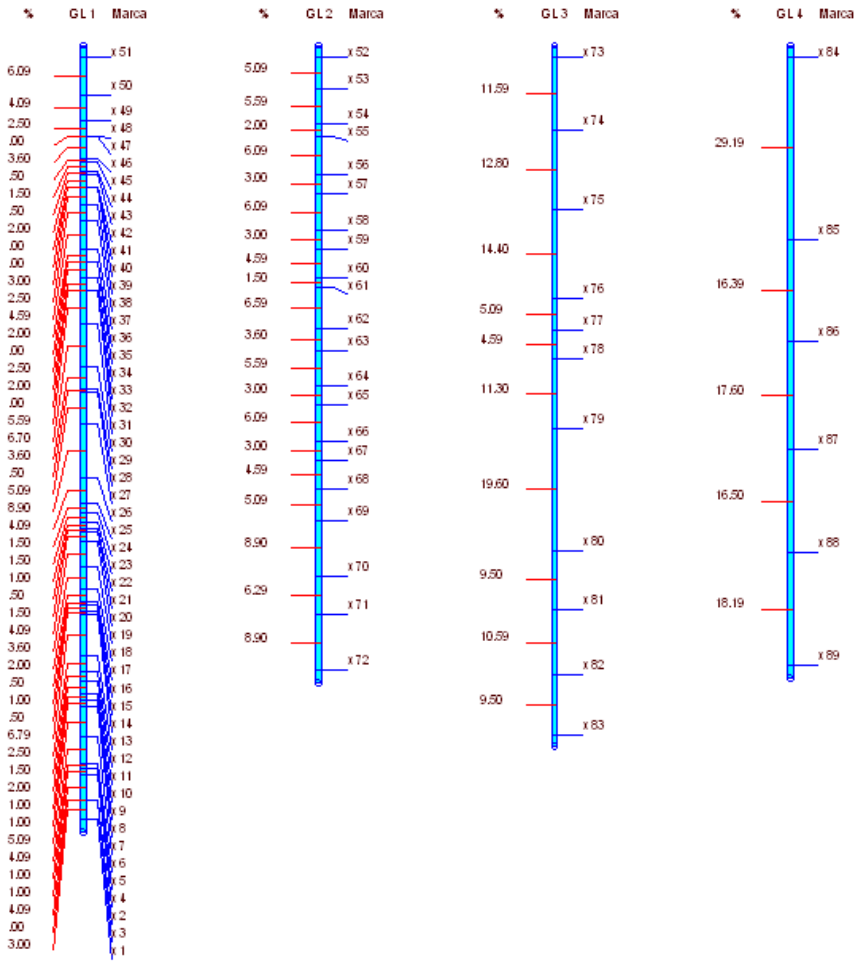

Figura 3 - Solução obtida para o problema de ordenação de marcas através do método delineação rápida em cadeia (GQMOL), para população com 100 indivíduos.

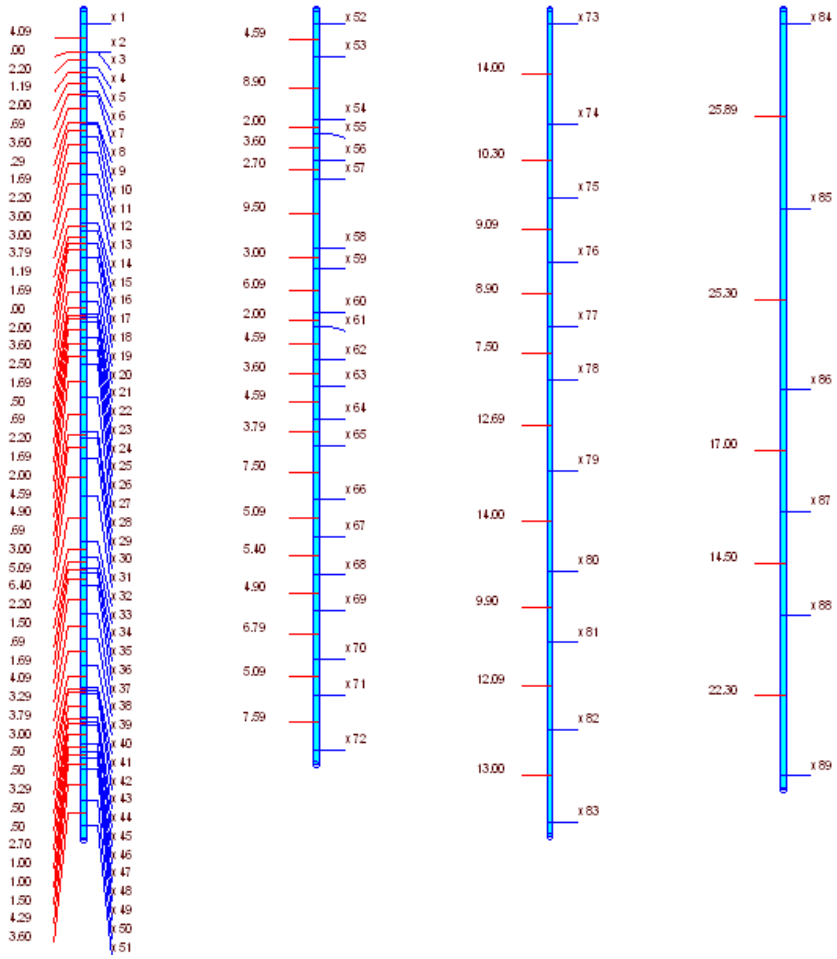

Figura 4 - Solução obtida para o problema de ordenação de marcas através do método delineação rápida em cadeia (GQMOL), para população com 200 indivíduos.

Para obter uma aproximação numérica da solução do problema de ordenação dos marcadores, utilizando o algoritmo s*imulated annealing*, é necessário definir um sistema de vizinhança em Λ, isto é, uma permutação candidata de marcadores. Adotou-se um sistema em que o vizinho típico (ordem candidata) de uma ordem $x_m = \left(m_{\sigma_1}, \ldots, m_{\sigma_i}, m_{\sigma_{i+1}}, \ldots, m_{\sigma_{j-1}}, m_{\sigma_j}, \ldots, m_{\sigma_k}\right)$ foi definido como $y_m = \left(m_{\sigma_1}, \ldots, m_{\sigma_i}, m_{\sigma_{j-1}}, m_{\sigma_{j-2}} \ldots, m_{\sigma_{i+1}}, m_{\sigma_j}, \ldots, m_{\sigma_k}\right)$. A Figura 1 apresenta um gráfico de um vizinho típico "candidato" de uma ordem $x_m \in \Lambda$.

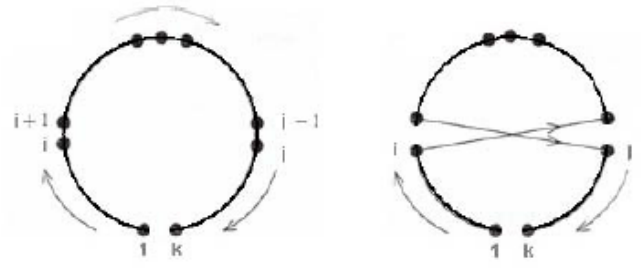

Figura 1 - Vizinho "candidato" de uma ordem $x \in \Lambda$.

Durante a aplicação do algoritmo, optou-se por escolher uniformemente uma ordem y_m no conjunto das possíveis ordens. O algoritmo foi implementado na linguagem de programação *R* versão 2.7.1 (R Development Core Team). O parâmetro de controle na *n*-ésima iteração do algoritmo, denotado por c_n, foi calculado com base na expressão,

$$c_n = \frac{A}{\ln(m+1)^2} \ .$$

Onde m é o número de iterações do algoritmo, *A* uma constante escolhida de forma conveniente.

A escolha de *A* é feita de forma que o algoritmo do s*imulated annealing* escape dos mínimos locais da função de interesse (*SARF*), e alcance o mínimo global. Portanto, a constante *A* deve ser escolhida de forma que todas as ordens iniciais sejam aceitas. Neste trabalho, utilizou-se 2 como o valor desta constante.

A seguir são apresentados os resultados obtidos para população constituída de 50 indivíduos.

Para o grupo de ligação 1, o s*imulated annealing* obteve como solução numérica a ordem, $m_1, m_2, \ldots, m_{14}, m_{16}, m_{17}, m_{15}, m_{18}, m_{19}, \ldots, m_{50}, m_{51}$, que possui uma distância total (*SARF*) de 132,9 cM, de tamanho menor que o da ordem fornecida pelo método da delineação rápida em cadeia, a qual possui 135,0 cM, dada pelos marcadores apresentados na Figura 2. Para o segundo, terceiro e quarto grupos de ligação às soluções obtidas pelo s*imulated annealing* são as mesmas obtidas pelo método implementado no programa GQMOL e também são apresentadas na figura 2. Estas ordens possuem distâncias de 101,1 , 118,2 e 96,5 respectivamente. A figura 5 mostra

a evolução das distâncias total a cada iteração do algoritmo nos grupos de ligação analisados.

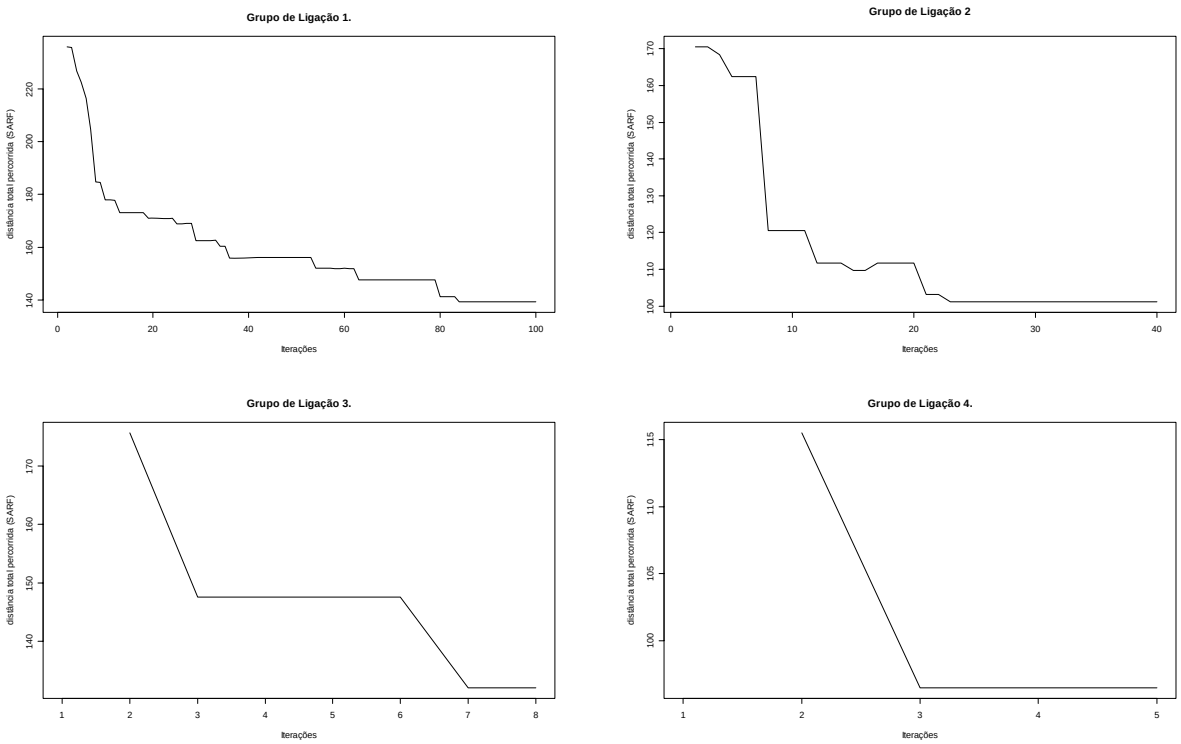

Figura 5 - Evolução das distâncias total a cada iteração do algoritmo, para populações de 50 indivíduos, nos grupos de ligação 1, 2, 3 e 4.

Para a população constituída de 100 indivíduos, a solução obtida para o primeiro grupo de ligação é dada pela seguinte ordem: $m_1, m_2, m_3, m_5, m_4, m_6, m_7, \ldots, m_{19}, m_{20}, \ldots, m_{50}, m_{51}$, esta ordem possui um *SARF* de 122,7 cM. Comparando-se a solução do primeiro grupo de ligação, com a obtida pelo algoritmo delineação rápida em cadeia (Figura 3), percebe-se que a ordem não é a mesma, entretanto as duas tem o mesmo valor de *SARF* igual a 122,7 cM. As soluções obtidas para os grupos de ligação 2, 3 e 4 obtidas pelo s*imulated annealing* são as mesmas encontradas pelo método delineação rápida em cadeia, e possuem distância total de 98,7, 109,00 e 99,90 cM cada, estas ordens são apresentadas na Figura 3. A figura 6 mostra a evolução das distâncias total a cada iteração do algoritmo nos grupos de ligação analisados.

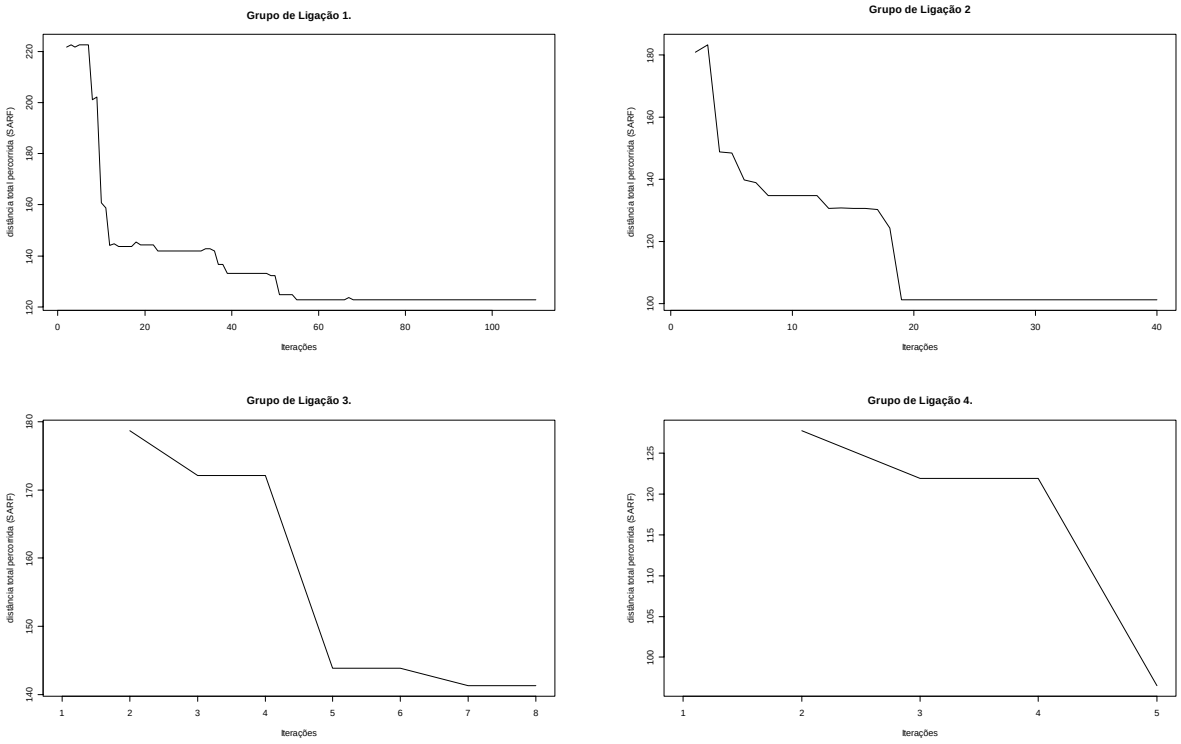

Figura 6 - Evolução das distâncias total a cada iteração do algoritmo, para populações de 100 indivíduos, nos grupos de ligação 1, 2, 3 e 4.

Considerando a população de 200 indivíduos, o primeiro grupo de ligação teve como solução numérica obtida através do método de otimização estocástica a ordem, $m_1, m_3, m_2, m_4, m_5, \ldots, m_{33}, m_{38}, m_{34,} m_{35}, m_{36}, m_{37}, m_{39}, m_{41}, m_{40,} m_{42,} m_{44}, m_{43}, m_{45}, m_{46}, \ldots, m_{50}, m_{51}$, que tem uma distância total de 111,50 cM, menor que a solução obtida pelo método implementado no programa GQMOL, que tem como valor de *SARF* 112,00 cM e a ordem numérica é apresentada na Figura 4. Para os demais grupos de ligação as soluções obtidas pelos dois métodos são equivalentes e podem ser vistas na Figura 4, estas ordens possuem a distância total de 101,40, 111,50 e 105,00 cM respectivamente. A Figura 7 mostra a evolução das distâncias total a cada iteração do algoritmo nos grupos de ligação analisados.

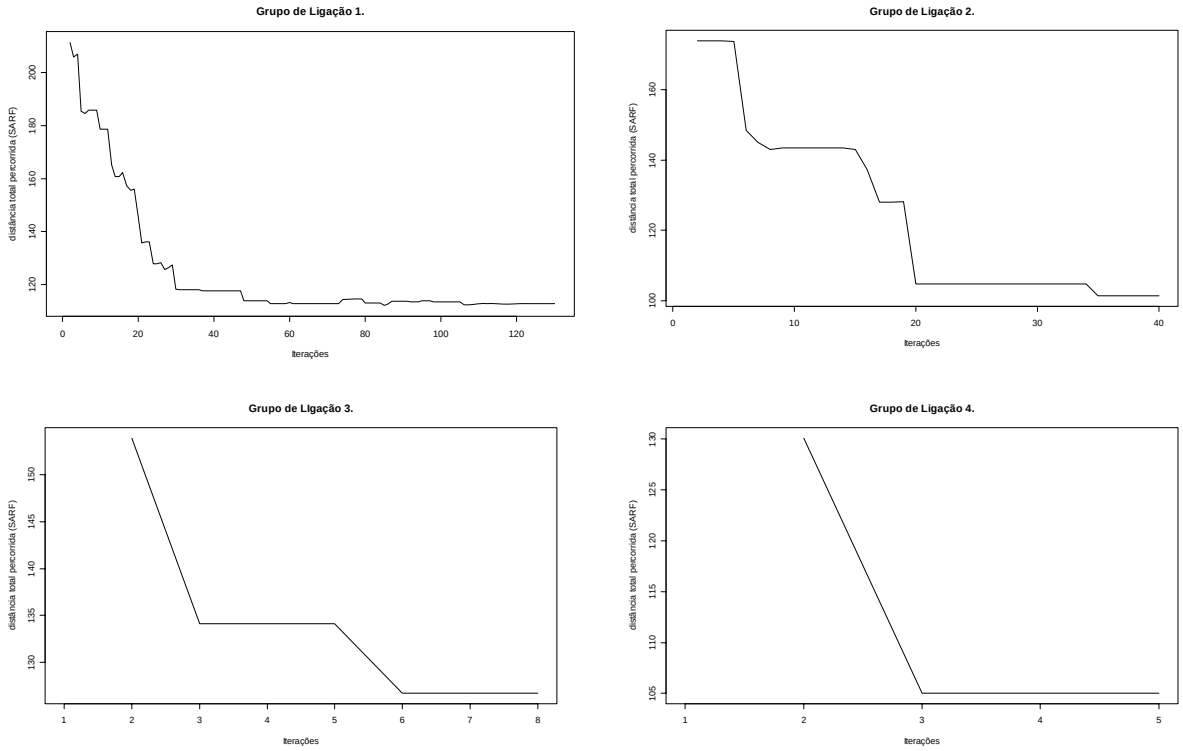

Figura 7 - Evolução das distâncias total a cada iteração do algoritmo, para populações de 200 indivíduos, nos grupos de ligação 1, 2, 3 e 4.

Através das Figuras 5, 6 e 7, nota-se que o número de iterações necessárias para que o algoritmo obtenha um resultado satisfatório, depende do número de marcadores no estudo, pois quanto maior o número de marcas no grupo de ligação maior o número de iterações.

A partir dos resultados, percebe-se que para grupos de ligação mais saturados, isto é com distâncias entre marcas adjacentes menores, 2 cM, o algoritmo do s*imulated annealing* obteve resultados semelhantes ou melhores que o método de delineação rápida em cadeia. Este melhor desempenho é também explicado pelo número de marcadores, pois quanto maior o número de marcadores, mais eficiente é o método de simulação estocástica em relação ao método da delineação rápida em cadeia. Isto se deve ao fato que o método utilizado no trabalho analisa um maior número de possíveis ordens. Já para os demais grupos de ligação que possuem níveis de saturação menores e consequentemente menor número de marcadores o algoritmo obteve resultados iguais ao método delineação rápida em cadeia.

Observa-se também que o número de indivíduos que constituem a população não influência no resultado obtido pelo algoritmo, uma vez que, a ordenação é feita levando em consideração as freqüências de recombinação, que já estão calculadas para cada par de marcadores. Assim o número de indivíduos influencia na precisão das estimativas e não na ordenação, podendo levar a construção de mapas de ligação imprecisos. A partir do trabalho de MOLLINARI et al. (2008), onde se conclui que os métodos delineação rápida em cadeia e seriação são equivalentes, pode-se supor que o método do *simulated annealing* é também superior ao método da seriação.

4. CONCLUSÕES

1. Para grupos de ligação muito saturados, distância adjacente entre marcas de 2 cM, e com maior número de marcas, 51 marcas, o método baseado em simulação estocástica, *simulated annealing*, apresentou ordens com distância (*SARF*) iguais ou menores que o método delineação rápida em cadeia.
2. Nos demais casos, ambos os métodos foram equivalentes, apresentando mesmo valor de *SARF*.
3. O número de indivíduos na população não influencia no ordenamento e sim nas estimativas das freqüências de recombinação.

REFERÊNCIAS BIBLIOGRÁFICAS

BUETOW, K. H.; CHAKRAVARTI, A. Multipoint gene mapping using seriation. I. General methods. **American Journal Human Genetics**, Chicago, v.41, p.180-188, 1987a.

BUETOW, K.H.; CHAKRAVARTI, A. Multipoint gene mapping using seriation. I. Analysis of simulated and empirical data. **American Journal Human Genetics**, Chigaco, v.41, p.189-201, 1987b.

CARNEIRO, N. S.; VIEIRA, M. L. C. Mapas genéticos em plantas. **Bragantia**, Campinas, v.61, p.89-100, 2002.

CRUZ, C. D. **Programa para análises de dados moleculares e quantitativos – GQMOL**. Viçosa: UFV, 2007.

DOERGE, R. Constructing genetic maps by rapid chain delineation. **Journal of Quantitative Trait Loci**, v.2, p.121-132, 1996.

FALK, C.T. A simple scheme for preliminary ordering of multiple loci: application to 45 CF families. **Prog Clin Biol Res.**, v.329, p.17-22, 1989.

ELSTON, R.C.; SPENCE, M.A.; RODGE, S.E.; MCCLUE, J.W. Multipoint mapping and linkage based upon affected pedigree members. In: Genetics Analysis Worshop 6. **Proceedings of a workshop**, Long Beach, Mississippi, p.17-22, 1988.

KIRKPATRICK, S.; GELATT, C.D.; VECCHI, M.P. Optimization by simulated annealing. **Science**, v.220, p.671-680, 1983.

LIU, B. H.; KNAPP, S. J. (1990) gmendel: A program for Mendelian segregation and linkage analysis of individual or multiple progeny populations using log-likelihood ratios. **J. Hered.,** v.81, p.407.

LIU, B. H. **Statistical genomics.** New York: CRC, 1998.

LU Y.Y.; LIU, B. H. **PGRI, a software for plant genome research.** Plant Genome III, San Diego, California, 1995.

MOLLINARI, M.; MARGARIDO, G. R. A.; GARCIA A. A. F. Comparação dos algoritmos delineação rápida em cadeia e seriação, para a construção de mapas genéticos. **Pesquisa agropecuária brasileira**, Brasília, v.43, p.505-512 (2008).

R DEVELOPMENT CORE TEAM (2007). R: **A language and environment for statistical computing**. R Fundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, disponível em: http://r-project .org.

ROBERT, C.; CASELLA, G. **Monte Carlo Statistical Methods.** 2ed. Springer, 2004. 645p.

SCHIEX, T.; GASPIN, C. **CARTAGENE: Constructing and joining maximum likelihood genetic maps.** In: Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology, Porto Carras, Halkidiki, Greece, p 258–267, 1997.

SCHUSTER, I.; CRUZ, C. D. **Estatística genômica - Aplicada a populações derivadas de cruzamentos controlados**. 2 ed. Editora UFV, Viçosa, 2008. 568p.

THOMPSON, E.A. Crossover counts and likelihood in multipoint linkage analysis. **IMA-Journal of Mathematical Applied Medicine Biology**, Oxford, v.4, p.93-108, 1987.

WEEKS, D.; LANGE, K. Preliminary ranking procedures for multilocus ordering. **Genomics**, p.236-242, 1987.

WEIR, B. **Genetic data analysis**. 2.ed. Sunderland: Sinauer Associates, 1996. 447p.

WILSON, S.R. A major simplification in the preliminary ordering of liked loci. **Genetic Epidemiology**, p.75-80, 1988.

CAPÍTULO 3

ESTIMAÇÃO DOS PARÂMETROS DE ADAPTABILIDADE E ESTABILIDADE: UMA ABORDAGEM BAYESIANA

RESUMO

Este trabalho teve como objetivo ilustrar a obtenção das estimativas dos parâmetros de adaptabilidade e estabilidade, do modelo proposto por FINLAY e WILKINSON (1963), através da inferência bayesiana, evidenciando seus passos, bem como a utilização do método *MCMC* amostrador de Gibbs. Além disso, comparou-se as estimativas obtidas pela inferência bayesiana com o método dos mínimos quadrados ordinários. Para ilustrar e comparar os resultados obtidos a partir da metodologia bayesiana, foram utilizados os dados de médias de rendimento de cinco genótipos avaliados em nove ambientes, provenientes de ensaios em blocos ao acaso com quatro repetições. Verificou-se que a estimação pontual por ambas as metodologias são equivalentes, porém, os intervalos de credibilidade apresentaram menores amplitudes, indicando uma maior precisão na estimação bayesiana. O uso do sofware WINBUGS, que utiliza o amostrador de Gibbs, proporcionou resultados satisfatórios tanto na estimação de parâmetros quanto na simplificação na derivação da técnica bayesiana.

1. INTRODUÇÃO

A partir de meados da década de 80 observa-se em aplicações no domínio da Estatística um enorme crescimento da inferência bayesiana. Isso se deve ao avanço dos recursos computacionais, os quais permitem resolver dificuldades inerentes à utilização desta metodologia (PAULINO et al., 2003).

Na abordagem bayesiana a incerteza sobre o parâmetro de interesse é representada através de modelos probabilísticos, desta forma, o parâmetro é tratado como variável aleatória e não apenas como uma quantidade observável como na abordagem clássica.

No melhoramento genético a inferência bayesiana teve como seu precursor Daniel Gianola que na década de 90 desenvolveu diversos trabalhos no melhoramento animal. Atualmente a inferência bayesiana não se restringe apenas a aplicações em genética animal, mas sim em todas as áreas da genética. Dentre os diversos estudos, no melhoramento de plantas, podem-se citar os trabalhos de REIS et al. (2008), onde utilizaram-se da metodologia bayesiana para estimar o coeficiente de endogamia e a taxa de fecundação cruzada de uma população diplóide por meio do modelo aleatório de COCKERHAM para freqüências alélicas. SILVA (2004), aplicou o amostrador de Gibbs para o ajuste de um Modelo Linear Generalizado Misto usando a inferência bayesiana em um ensaio envolvendo dados de contagens de tubérculos graúdos em batata, visando à obtenção de estimativas de parâmetros genéticos como herdabilidades, componentes da variância e valores genéticos. Além destas aplicações no melhoramento de plantas, a inferência bayesiana vem sendo frequentemente empregada em estudos de evolução como pode ser verificado no trabalho de O' HARA et al. (2008) onde realizaram uma

revisão bibliográfica com objetivo de dar uma visão geral do uso de métodos bayesianos em genética evolutiva.

Entretanto, a utilização da abordagem bayesiana por vezes envolve derivações complexas que necessitam a utilização de recursos computacionais. Dentre estes, uma classe de técnicas que se destaca para solução dos problemas são os métodos de simulação de Monte Carlo via Cadeias de Markov (*MCMC*).

Este trabalho tem por objetivo ilustrar a obtenção de estimativas dos parâmetros de adaptabilidade e estabilidade, utilizados para estudo da interação genótipos x ambientes, a partir do modelo proposto por FINLAY e WILKINSON (1963), fazendo uso da inferência bayesiana, evidenciando seus passos, e a utilização do método *MCMC* amostrador de Gibbs através do software WINBUGS.

2. MATERIAL E MÉTODOS

Para ilustrar a aplicação da metodologia bayesiana na estimação dos parâmetros de adaptabilidade e estabilidade, foram utilizados os dados de médias de rendimento de cinco genótipos avaliados em nove ambientes, provenientes de ensaios em blocos ao acaso com quatro repetições. Estes dados apresentados na Tabela 1 podem ser encontrados em CRUZ, REGAZZI e CARNEIRO (2004).

Tabela1 – Médias de rendimento de cinco genótipos avaliados em nove ambientes

Genótipos	Ambientes								
	A1	A2	A3	A4	A5	A6	A7	A8	A9
1	2,0	6,4	7,3	3,8	3,1	5,9	5,4	8,3	7,9
2	3,7	6,7	8,4	3,6	4,1	8,1	5,8	6,7	5,5
3	3,1	6,6	8,1	3,8	4,7	6,3	6,3	7,1	5,7
4	2,4	6,1	8,6	2,8	4,2	5,3	5,9	5,2	4,5
5	4,9	4,9	6,3	3,8	4,0	3,8	4,3	4,4	3,8
I_j	-2,1	0,82	2,41	-1,76	-1,3	0,56	0,22	1,01	0,16

Fonte: Cruz, Regazzi e Carneiro (2004).

O método proposto por FINLAY e WILKINSON (1963), baseia-se em analise de regressão, que mede a resposta de cada genótipo às variações ambientais.

Para um experimento com *g* genótipos, *a* ambientes e *r* repetições define-se o seguinte modelo estatístico, $Y_{ij} = \beta_{0i} + \beta_{1i} I_j + \psi_{ij}$

em que:

Y_{ij} : média de genótipo *i* no ambiente *j*;

β_{0i} : coeficiente linear referente ao *i*-ésimo genótipo (intercepto);

β_{1i} : coeficiente de regressão, o qual mede a resposta do *i*-ésimo genótipo à variação do ambiente;

I_j : índice ambiental codificado $\left(I_j = \frac{\sum_j Y_j}{g} - \frac{\sum_i \sum_j Y_{ij}}{ga} \right)$;

ψ_{ij} : erros aleatórios;

As estimativas de I_j indicam a qualidade do ambiente, onde valores negativos de I_j identificam ambientes desfavoráveis e valores positivos de I_j, ambientes favoráveis.

2.1. Estimação dos parâmetros de adaptabilidade e estabilidade via método dos mínimos quadrados ordinários

No método usual, para obtenção dos parâmetros de um modelo de regressão, é utilizado o método de mínimos quadrados ordinários (*MQO*), que consiste em adotar como estimativas dos parâmetros, os valores que minimizam a soma de quadrados dos desvios (erros). Deste modo, os estimadores de mínimos quadrados são dados por:

Estimador de β_{0i} :

$$\hat{\beta}_{0i} = \overline{Y}_{i\cdot}.$$

Estimador de β_{1i} :

$$\hat{\beta}_{1i} = \frac{\sum_j Y_{ij} I_j}{\sum_j I_j^2}.$$

Estimador de σ_i^2 :

$$\hat{\sigma}_i^2 = \frac{\left[\sum_j Y_{ij}^2 - \frac{\left(\sum Y_{ij} \right)^2}{a} \right] - \hat{\beta}_{1i} \sum_j Y_{ij} I_j}{a-2}.$$

Além da estimação pontual é possível obter intervalos de confiança para os parâmetros de interesse. Pode-se pensar na amplitude do intervalo de confiança estimado como uma medida da qualidade da reta de regressão estimada (MONTGOMERY e

PECK, 1992). Os intervalos de confiança para os parâmetros β_{0i}, β_{1i} e σ_i^2 com 100 (1-α) % de confiança de são dados por

$$\hat{\beta}_{0i} - t_{(\alpha/2,a-2)}\sqrt{\frac{\hat{\sigma}_i^2}{a}} \leq \beta_{0i} \leq \hat{\beta}_{0i} + t_{(\alpha/2,a-2)}\sqrt{\frac{\hat{\sigma}_i^2}{a}};$$

$$\hat{\beta}_{1i} - t_{(\alpha/2,a-2)}\sqrt{\frac{\hat{\sigma}_i^2}{\sum_j I_{ij}^2}} \leq \beta_{1i} \leq \hat{\beta}_{1i} + t_{(\alpha/2,a-2)}\sqrt{\frac{\hat{\sigma}_i^2}{\sum_j I_{ij}^2}};$$

$$\frac{(a-2)\hat{\sigma}_i^2}{\chi^2_{(\alpha/2,a-2)}} \leq \sigma_i^2 \leq \frac{(a-2)\hat{\sigma}_i^2}{\chi^2_{((1-(\alpha/2),a-2)}};$$

respectivamente.

2.2. Estimação dos parâmetros de adaptabilidade e estabilidade via inferência bayesiana

Para estimação dos parâmetros de adaptabilidade e estabilidade via inferência bayesina é necessário atribuir distribuições a priori para os parâmetros de interesse. Para cada genótipo foram consideradas as seguintes distribuições para β_{0i}, β_{1i} e $\frac{1}{\sigma_i^2}$,

$$P_i(\beta_{0i}) \sim N(\mu_{0i}, \sigma_{0i}^2),$$

$$P_i(\beta_{1i}) \sim N(\mu_{1i}, \sigma_{1i}^2),$$

$$P_i\left(\frac{1}{\sigma_i^2}\right) \sim Ga(\alpha, \beta),$$

em que, $N(\mu_{0i}, \sigma_{0i}^2)$ e $N(\mu_{1i}, \sigma_{1i}^2)$ denotam distribuições normais com médias μ_{0i} e μ_{1i} e variâncias σ_{0i}^2 e σ_{1i}^2, respectivamente, e $Ga(\alpha, \beta)$ denota a distribuição Gama com média $\frac{\alpha}{\beta}$ e variância $\frac{\alpha}{\beta^2}$ (EHLERS, 2007). Neste trabalho, por motivo de

simplicidade, utilizou-se a precisão $\tau = \left(\frac{1}{\sigma_i^2}\right)$ ao invés da variância (σ_i^2). Desta forma a estimativa de (σ_i^2) é obtida através de $\hat{\sigma}_i^2 = \frac{1}{\tau}$.

Além disso, é assumida a independência a priori entre os parâmetros. Assim, as distribuições a prioris conjunta para cada genótipo são dadas por

$$P_i(\beta_{0i}, \beta_{1i}, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_{0i}^2}\exp\left\{-\frac{1}{2\sigma_{0i}^2}(\beta_{0i}-\mu_{0i})^2\right\} \times \frac{1}{\sqrt{2\pi}\sigma_{1i}^2}\exp\left\{-\frac{1}{2\sigma_{1i}^2}(\beta_{1i}-\mu_{1i})^2\right\} \times$$
$$\frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma_i^2}\right)^{\alpha+1}\exp\left\{-\frac{\beta}{\sigma_i^2}\right\}$$
$$\propto \exp\left\{-\frac{1}{2\sigma_{0i}^2}(\beta_{0i}-\mu_{0i})^2\right\} \times \exp\left\{-\frac{1}{2\sigma_{1i}^2}(\beta_{1i}-\mu_{1i})^2\right\}$$
$$\times\left(\frac{1}{\sigma_i^2}\right)^{\alpha+1}\exp\left\{-\frac{\beta}{\sigma_i^2}\right\}.$$

Considerando o modelo estatístico $Y_{ij} = \beta_{0i} + \beta_{1i}I_j + \psi_{ij}$, cada observação Y_{ij} tem distribuição $Y_{ij} \sim N(\beta_{0i} + \beta_{1i}I_j, \sigma_i^2)$ e suas funções de verossimilhança são dadas por

$$L(\beta_{0i}, \beta_{1i}, \sigma_i^2; y_{ij}) = \prod_{j=1}^{9}\frac{1}{\sqrt{2\pi}\sigma_i^2}\exp\left\{-\frac{1}{2\sigma_i^2}\left(y_{ij}-(\beta_{0i}+\beta_{1i}I_j)\right)^2\right\}$$
$$= \frac{1}{\left(\sqrt{2\pi}\sigma_i\right)^9}\exp\left\{-\frac{1}{2\sigma_i^2}\sum_{j=1}^{9}\left(y_{ij}-(\beta_{0i}+\beta_{1i}I_j)\right)^2\right\} \quad i = 1,2,3,4,5.$$

A distribuição a posteriori dos parâmetros é dada combinando-se a distribuição priori conjunta com a função de verossimilhança, desta forma

$$P_i(\beta_{0i},\beta_{1i},\sigma_i^2 \mid y_{ij}) \propto L(\beta_{0i},\beta_{1i},\sigma_i^2; y_{ij}) \times P_i(\beta_{0i},\beta_{1i},\sigma_i^2)$$

$$\propto \frac{1}{\left(\sqrt{2\pi}\sigma_i\right)^n} \exp\left\{-\frac{1}{2\sigma_i^2}\sum_{j=1}^{9}\left(y_{ij}-(\beta_{0i}+\beta_{1i}I_j)\right)^2\right\} \times$$

$$\exp\left\{-\frac{1}{2\sigma_{0i}^2}(\beta_{0i}-\mu_{0i})^2\right\} \times \exp\left\{-\frac{1}{2\sigma_{1i}^2}(\beta_{1i}-\mu_{1i})^2\right\} \times$$

$$\left(\frac{1}{\sigma_i^2}\right)^{\alpha+1} \exp\left\{-\frac{\beta}{\sigma_i^2}\right\} \qquad i = 1,2,3,4,5.$$

A partir deste ponto, apesar de estarmos trabalhando com um modelo relativamente simples, percebe-se que a obtenção de estimavas dos parâmetros não é tarefa fácil, uma vez que para obtenção das mesmas é necessário derivar as distribuições marginais a posteriori que são obtidas a partir das seguintes integrais.

$$P_i(\beta_{0i} \mid Y_{ij}) = \iint P_i(\beta_{0i},\beta_{1i},\sigma_i^2 \mid y_{ij}) d\beta_{1i} d\frac{1}{\sigma_i^2};$$

$$P_i(\beta_{1i} \mid Y_{ij}) = \iint P_i(\beta_{0i},\beta_{1i},\sigma_i^2 \mid y_{ij}) d\beta_{0i} d\frac{1}{\sigma_i^2};$$

$$P_i(\frac{1}{\sigma_i^2} \mid Y_{ij}) = \iint P_i(\beta_{0i},\beta_{1i},\sigma_i^2 \mid y_{ij}) d\beta_{0i} d\beta_{1i}.$$

Para contornar este problema e obter amostras das distribuições marginais a posteriori, o que possibilita a obtenção de amostras para as quais é possível fazer inferência sobre os parâmetros de interesse, é necessário fazer uso dos métodos *MCMC*. O método simulação de Monte Carlo via cadeias de Markov mais popular em aplicações bayesianas é o amostrador de Gibbs e uma explanação completa deste algoritmo pode ser encontrada em CASELLA e GEORGE (1992).

Na inferência Bayesiana, os intervalos para os parâmetros do modelo (intervalos de credibilidade) são obtidos diretamente da distribuição a posteriori dos parâmetros. Seja $\theta_i = (\beta_{0i},\beta_{1i},\sigma_i^2)$ o vetor de parâmetros a serem estimados. Fixando uma probabilidade $1-\alpha$, o intervalo de credibilidade para θ_i com probabilidade de cobertura $1-\alpha$ é dado por (θ_*,θ^*), tal que

$$\int_{-\infty}^{\theta^*} P_i(\theta_i = (\beta_{0i}, \beta_{1i}, \sigma_i^2) \mid y_{ij}) d\theta_i = \frac{\alpha}{2} \; ; \; \int_{\theta^*}^{\infty} P_i(\theta_i = (\beta_{0i}, \beta_{1i}, \sigma_i^2) \mid y_{ij}) d\theta_i = \frac{\alpha}{2}.$$

2.3. Amostrador de Gibbs

Embora o amostrador de Gibbs seja um caso especial do algoritmo de Metropolis-Hastings (1970), sua implementação depende de algumas particularidades. O uso deste algoritmo requer o conhecimento das distribuições condicionais completas. Desta forma, nesse trabalho é necessário explicitar as distribuições $P_i(\beta_{0i} \mid Y, \beta_{1i}, \sigma_i^2)$, $P_i(\beta_{1i} \mid Y, \beta_{0i}, \sigma_i^2)$, $P_i(\sigma_i^2 \mid Y, \beta_{0i}, \beta_{1i})$.

As distribuições a posteriori condicionais, para cada genótipo $i = 1,2,3,4,5,$ são dadas por

i) Para β_{0i},

$$P_i(\beta_{0i} \mid Y, \beta_{1i}, \sigma_i^2) \propto \exp\left\{ -\frac{1}{2\sigma_{0i}^2}(\beta_{0i} - \mu_{0i})^2 - \frac{1}{2\sigma_i^2}\sum_{j=1}^{9}\left(y_{ij} - \left(\beta_{0i} + \beta_{1i} I_j\right)\right)^2 \right\}.$$

ii) Para β_{1i},

$$P_i(\beta_{1i} \mid Y, \beta_{0i}, \sigma_i^2) \propto \exp\left\{ -\frac{1}{2\sigma_{1i}^2}(\beta_{1i} - \mu_{1i})^2 - \frac{1}{2\sigma_i^2}\sum_{j=1}^{9}\left(y_{ij} - \left(\beta_{0i} + \beta_{1i} I_j\right)\right)^2 \right\}.$$

iii) Para σ_i^2,

$$P_i(\sigma_i^2 \mid Y, \beta_{0i}, \beta_{1i}) \propto \left(\sigma_i^2\right)^{-(\alpha+1)} \frac{1}{\left(\sqrt{2\pi}\sigma_i\right)^n} \exp\left\{ -\frac{1}{2\sigma_i^2}\sum_{j=1}^{9}\left(y_{ij} - (\beta_{0i} + \beta_{1i} I_j)\right)^2 - \frac{\beta}{\sigma_i^2} \right\}$$

.

Tomando como base o trabalho de COELHO-BARROS et al. (2008), foram considerados $\mu_{0i} = 0$, $\mu_{1i} = 0$, $\sigma_{0i}^2 = 1000000$, $\sigma_{1i}^2 = 1000000$, $\alpha_i = 0{,}01$ e $\beta_i = 0{,}01$, $i = 1,2,3,4,5$, nas distribuições a priori. Essa escolha de hiperparâmetros é motivada para que se tenham distribuições a priori não-informativas e tal que a convergência do algoritmo seja observada.

Apresentado os ingredientes para a implementação do algoritmo, o Amostrador de Gibbs, para cada genótipo, pode ser descrito da seguinte forma:

Algoritmo 1.

1. Defina os valores iniciais da cadeia $\theta_0 = (\beta_0^{(0)}, \beta_1^{(0)}, \sigma^{2(0)})$;
2. Para $r \geq 1$, faça $s = 0$;
   (a) Para $s \geq 0$
      i. Simule $\beta_0^{(s+1)}$ de $P(\beta_0 \mid Y, \beta_1, \sigma^2)$;
      ii. Simule $\beta_1^{(s+1)}$ de $P(\beta_1 \mid Y, \beta_0, \sigma^2)$;
      iii. Simule $\left(\frac{1}{\sigma^2}\right)^{(s+1)}$ de $P(\frac{1}{\sigma_r^2} \mid Y, \beta_0, \beta_1)$.
   (b) Se $s < s_{\max}$, faça $s \leftarrow s+1$ e retorne a ao passo (a); ($s_{máx}$ é o espaçamento entre as observações)
3. $\theta_r = \left(\beta_0^{(máx)}, \beta_1^{(máx)}, \left(\frac{1}{\sigma^2}\right)^{(máx)}\right)$;
4. $\theta_0 = \theta_r$;
5. Se $r < r_{máx}$, faça $r \leftarrow r+1$ e retorne ao passo (2); ($r_{máx}$ é o tamanho da amostra)
6. Estime θ por

$$\hat{\theta} = \bar{\theta}_{r_{máx}} = \frac{1}{r_{máx}} \sum_{r=1}^{r_{máx}} \theta_r \ .$$

Para simular das distribuições necessárias ao algoritmo, pode-se usar o método da transformação inversa (ROBERT E CASELLA, 2004). Entretanto, em diversos softwares estatísticos já existem distribuições de probabilidade com simuladores já implementados.

Neste trabalho, utilizou-se para obtenção das estimativas dos parâmetros de adaptabilidade e estabilidade o software WINBUGS (Bayesian inference using gibbs sampling for windows) (LUNN et al., 2000). Esse aplicativo é ferramenta extremamente útil na solução de problemas com modelagem bayesiana dada sua facilidade de manuseio e a grande gama de problemas que ele consegue resolver através do Amostrador de Gibbs. Este software foi desenvolvido pela unidade de Bioestatística da Medical Research Council da Universidade de Cambridge em um projeto iniciado em

1989 e está disponível para download em http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml. O uso deste software faz com que não seja necessário o conhecimento das distribuições condicionais completas a posteriori por parte do pesquisador, ele apenas requer a especificação da distribuição dos dados, prioris e dos valores iniciais da cadeia.

3. RESULTADOS E DISCUSSÃO

As estimativas pontuais dos parâmetros β_{0i}, β_{1i} e σ_i^2, obtidas através do *MQO* e seus respectivos intervalos de confiança para cada um dos cinco genótipos em estudo são apresentados na Tabela 1. Estas estimativas foram obtidas por meio do software GENES (CRUZ, 2006).

Tabela 1 – Parâmetros de adaptabilidade e estabilidade, estimados segundo o método dos mínimos quadrados ordinários (*MQO*), para cinco genótipos em nove ambientes

Genótipos	Parâmetros	Estimativas	Intervalo de Confiança (95%)
	β_0	5,57	[4.60;6.53]
1	β_1	1,28	[0.58;1.98]
	σ^2	1,50	[0,66;6,21]
	β_0	5,84	[5,30;6,39]
2	β_1	1.15	[0,76;1,54]
	σ^2	0,47	[0,21;1,95]
	β_0	5,74	[5,52;5,97]
3	β_1	1,08	[0,92;1,24]
	σ^2	0,08	[0,03;0,33]
	β_0	5,00	[4,47;5,53]
4	β_1	1,19	[0,81;1,58]
	σ^2	0,45	[0,20;1,86]
	β_0	4,48	[3,89;5,05]
5	β_1	0,30	[-0,13;0,718]
	σ^2	0,55	[0,24;2,34]

Para obtenção das estimativas dos parâmetros, baseado na metodologia bayesiana via amostrador de Gibbs CASELLA e GEORGE (1992), utilizou-se o sofware WINBUGS (LUNN et al. 2000). Foram geradas 10000 amostras, das quais as 1000 primeiras foram descartadas com a finalidade de eliminar o efeito dos valores iniciais usados no algoritmo de simulação. A Tabela 2 apresenta os parâmetros de adaptabilidade e estabilidade, estimados segundo esta metodologia.

Através das Tabelas 1 e 2 percebe-se que as estimativas pontuais encontradas pelas duas abordagens (*MQO* e inferência bayesiana) são iguais ou semelhantes, mostrando que estas metodologias são equivalentes, quando da estimação pontual de parâmetros de adaptabilidade e estabilidade.

Tabela 2 – Parâmetros de adaptabilidade e estabilidade, estimados segundo a metodologia bayesiana, para cinco genótipos em nove ambientes

Genótipos	Parâmetros	Estimativas	Intervalo de Confiança (95%)
	β_0	5,57	[4.60;6.53]
1	β_1	1,28	[0.58;1.98]
	σ^2	1,50	[0,66;6,21]
	β_0	5,84	[5,30;6,39]
2	β_1	1.15	[0,76;1,54]
	σ^2	0,47	[0,21;1,95]
	β_0	5,74	[5,52;5,97]
3	β_1	1,08	[0,92;1,24]
	σ^2	0,08	[0,03;0,33]
	β_0	5,00	[4,47;5,53]
4	β_1	1,19	[0,81;1,58]
	σ^2	0,45	[0,20;1,86]
	β_0	4,48	[3,89;5,05]
5	β_1	0,30	[-0,13;0,718]
	σ^2	0,55	[0,24;2,34]

Considerando as estimativas por intervalos, observa-se que os intervalos de credibilidade possuem amplitudes menores que os obtidos a partir do método tradicional, o qual requer a pressuposição de normalidade, mostrando-se mais precisos. Isso pode ser devido à existência de poucas observações na amostra (ambientes), fazendo com que a distribuição dos dados se afaste da normalidade, enquanto que na

metodologia bayesiana não existe a necessidade do uso de aproximações ou da pressuposição da normalidade.

A convergência do algoritmo foi verificada através de gráficos temporais das amostras geradas. Segundo MELO e EHLERS (2006), essa é uma das ferramentas utilizadas para verificar a convergência, onde são plotados os valores gerados da variável em cada iteração do algoritmo. A aleatoriedade deste gráfico indica a convergência da cadeia para a distribuição de interesse.

As Figuras 1, 2, 3, 4 e 5 apresentam os gráficos para verificação da convergência e a densidade de cada uma das variáveis em estudo para os genótipos 1, 2, 3, 4 e 5 respectivamente. Verifica-se que a convergência das cadeias utilizadas na estimação dos parâmetros foi alcançada (Figuras 1, 2, 3, 4 e 5). Portanto, pode-se inferir que as estimativas dos parâmetros de interesse são, verdadeiramente, as médias das distribuições a posteriori. A assimetria verificada nos gráficos das densidades a posteriori de $\frac{1}{\sigma_i^2}$, *i*= 1,2,3,4,5, pode ser percebida a partir dos gráficos temporais das amostras geradas, uma vez que a mesma possui alguns valores “altos”.

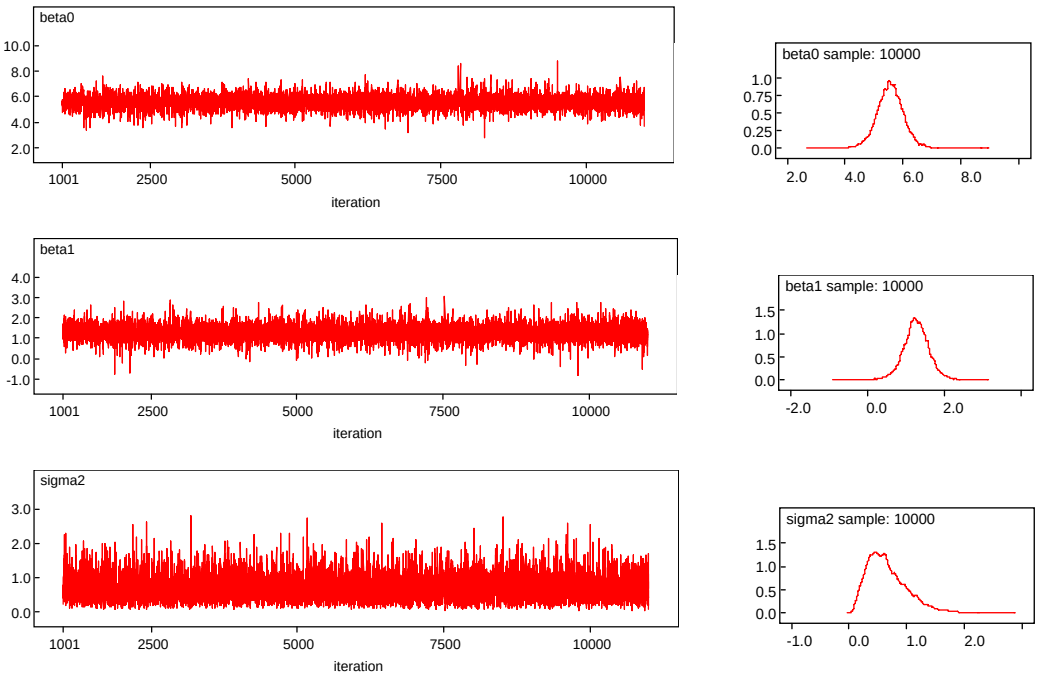

Figura 1- Trajetória das cadeias e aproximações das densidades a posterioris dos parâmetros β_{01}, β_{11} e $\frac{1}{\sigma_1^2}$.

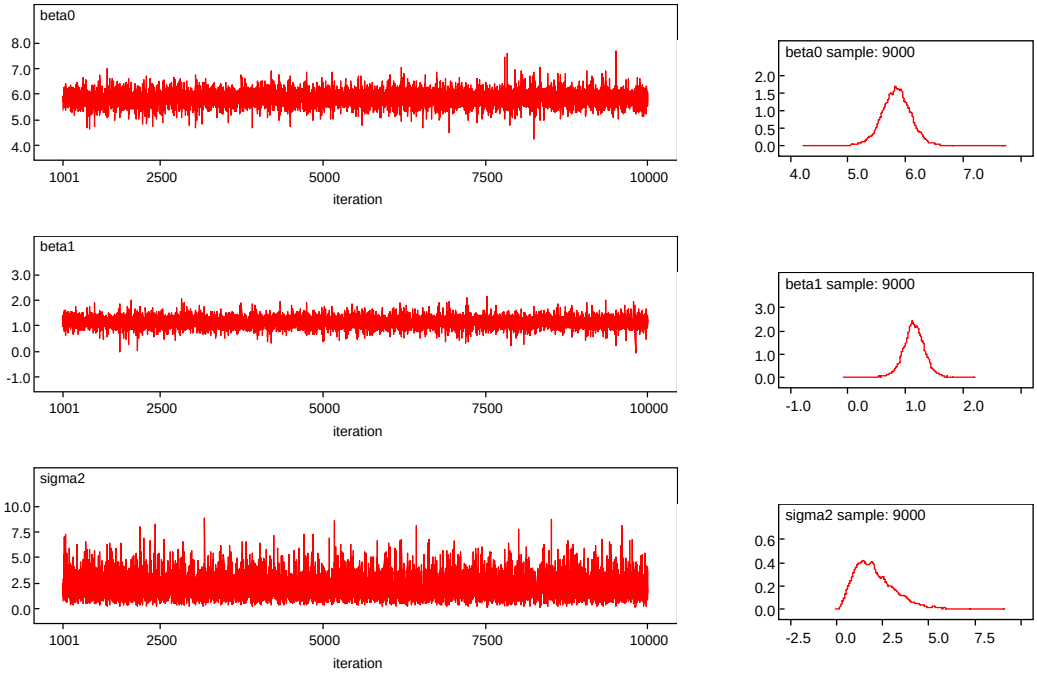

Figura 2 - Trajetória das cadeias e aproximações das densidades a posterioris dos parâmetros β_{02}, β_{12} e $\frac{1}{\sigma_2^2}$.

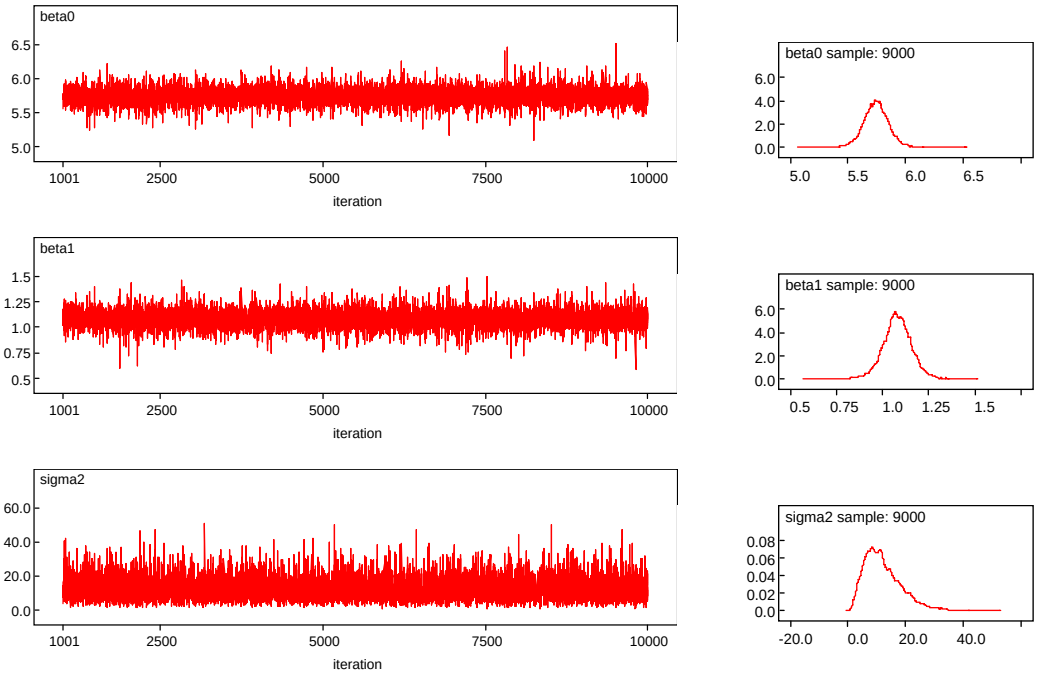

Figura 3 - Trajetória das cadeias e aproximações das densidades a posterioris dos parâmetros β_{03}, β_{13} e $\frac{1}{\sigma_3^2}$.

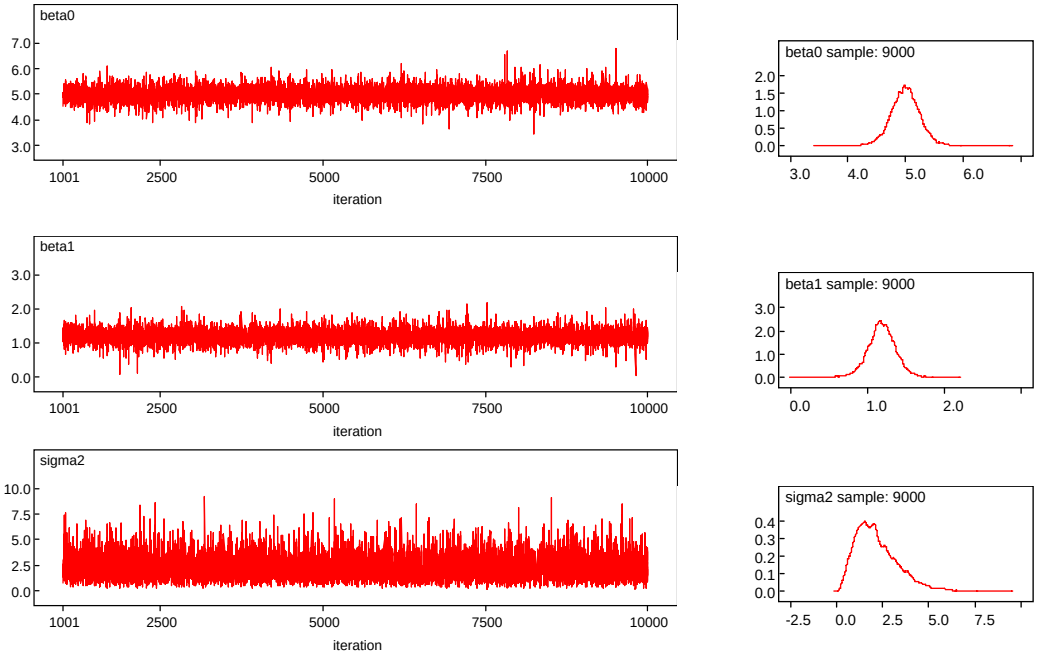

Figura 4 - Trajetória das cadeias e aproximações das densidades a posterioris dos parâmetros β_{04}, β_{14} e $\frac{1}{\sigma_4^2}$.

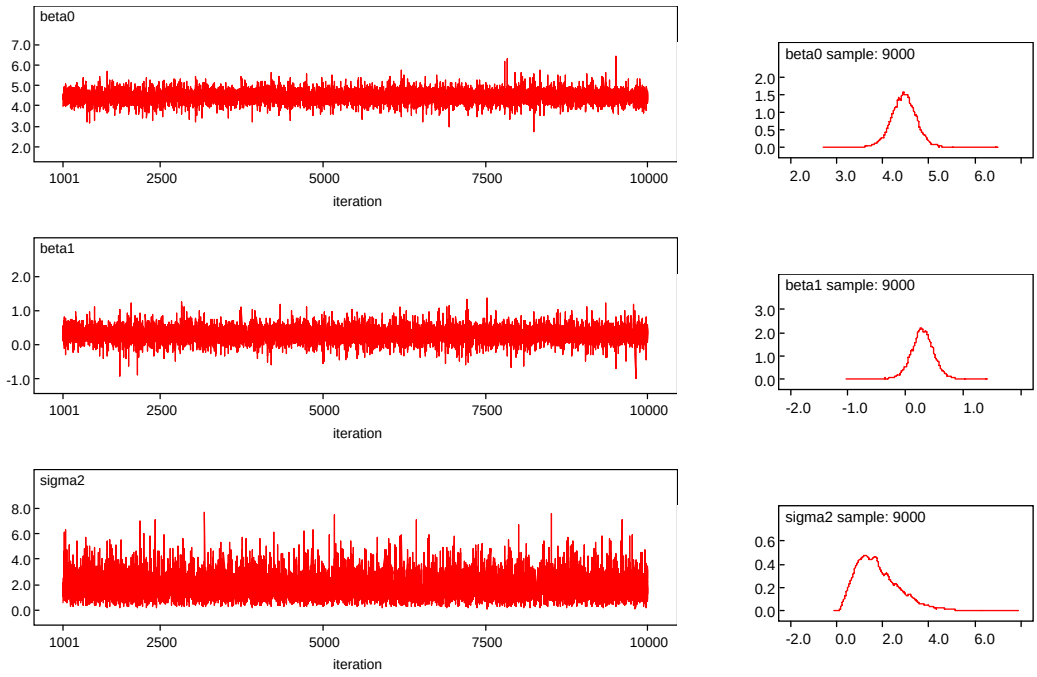

Figura 5 - Trajetória das cadeias e aproximações das densidades a posterioris dos parâmetros β_{05}, β_{15} e $\frac{1}{\sigma_5^2}$.

4. CONCLUSÕES

1. A estimação pontual dos parâmetros de adaptabilidade e estabilidade através das metodologias de mínimos quadrados ordinários e inferência bayesiana (através do amostrador de Gibbs) foram semelhantes;
2. Os intervalos de credibilidade, em geral obtiveram menores amplitudes, indicando uma maior precisão na estimação dos parâmetros;

REFERÊNCIAS BIBLIOGRÁFICAS

CASELLA, G.; GEORGE, E. Explaining the Gibbs Sampler. **The American Statistician**, v.46, p. 167-157, 1992.

COELHO-BARROS, E. A.; SIMÕES, P. A.; ACHCAR, J. A.; MARTINEZ, E. .Z; SHIMANO, A. C. Métodos de estimação em regressão linear múltipla: aplicação a dados clínicos. **Revista Colombiana de Estadística**, v. 31, p. 111-129, 2008.

CRUZ, C. D. **Programa GENES: BIOMETRIA**. Editora – UFV. Viçosa - MG, 2006. 382p.

CRUZ, C. D.; REGAZZI, A.J.; CARNEIRO, P. C. **Modelos biométricos aplicados ao melhoramento genético**. Editora UFV. Viçosa, 2004. 480p.

EHLERS, R. S. (2007) **Introdução à Inferência Bayesiana.** Disponível em http://leg.ufpr.br/~ ehlers/bayes. Acesso em: dezembro 2008.

FINLAY, K.W.; WILKINSON, G.N. The analysis of adaptation in a plant-breeding programme. **Australian Journal of Agricultural Research**, v. 14, p. 742-754, 1963.

HASTINGS, W. Monte Carlo Sampling Methods Using Markov Chains and TheirApplications. **Biometrika**, v.57, p.97-109, 1970.

LUNN, D.J.; THOMAS, A.; BEST, N.; SPIEGELHALTER, D. WINBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. **Statistics and Computing**, v.10, p.325-337, 2000.

MELO, L.; EHLERS, R. S. (2006) Seminário Winbugs**.** Disponível em http://est.ufpr.br/rt/winbugs.pdf. Acessado em: dezembro 2008.

MONTGOMERY, D. C.; PECK, E. A. **Introduction to Linear Regression Analysis.** Wiley and Sons, 1992, 544p.

O'HARA, R. B.; CANO, J. M.; OVASKAINEN, O.; TEPLITSKY, C.; ALHO J. S. Bayesian approaches in evolutionary quantitative genetics. **J. Evol. Biol.**, v. 21, p. 949–957, 2008.

PAULINO, C. D.; AMARAL TURKMAN, A.; MURTEIRA, B. **Estatística Bayesiana**. Fundação Calouste Gulbenkian, 2003, 446p.

REIS, R. L.; MUNIZ, J. A.; SILVA, F. F.; SÁFADI, T.; AQUINO, L. H. Inferência Bayesiana na análise genética de populações diplóides: estimação do coeficiente de endogamia e da taxa de fecundação cruzada. **Ciência Rural**, v. 38, p.1258-1265, 2008.

ROBERT, C.; CASELLA, G. **Monte Carlo Statistical Methods.** 2ed. Springer, 2004. 645p.

SILVA, J. W. **Análise Bayesiana de um modelo linear generalizado misto: Emprego no melhoramento de plantas**. Lavras, MG: UFLA, 2004. 76p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária). Universidade Federal de Lavras, Lavras.

CONSIDERAÇÕES FINAIS

Este trabalho abordou os principais métodos *MCMC*, apresentando os fundamentos teóricos dos algoritmos com base na teoria de cadeias de Markov. Além disso, buscou-se apresentar exemplos e aplicações destes algoritmos relacionados ao melhoramento genético.

No Capítulo 1, o algoritmo de Metropolis-Hastings foi utilizado para estimar a freqüência de recombinação entre pares de marcadores de uma população F_2 simulada. O algoritmo mostrou-se de fácil utilização e obteve resultados semelhantes aos obtidos pelos métodos gráfico e iterativo de Newton-Raphson. Além disso, mostrou-se que independente do valor inicial da cadeia, o algoritmo converge para a distribuição de interesse.

Foi mostrado também que uma pequena modificação no algoritmo de Metropolis-Hastings é capaz de transformá-lo em um algoritmo de otimização, chamado *simulated annealing.*

O *simulated anneliang* foi utilizado no estabelecimento da melhor ordem de ligação na construção de mapas genéticos. O algoritmo foi comparado com o método da delineação rápida em cadeia o qual está implementado no software GQMOL. Para grupos de ligação muito saturados, distância adjacente entre marcas de 2 cM, e com maior número de marcas, 51 marcas, o método baseado em simulação estocástica, *simulated annealing*, apresentou ordens com distância (*SARF*) iguais ou menores que o método delineação rápida em cadeia. Nos demais casos, ambos os métodos foram equivalentes, apresentando mesmos valores de *SARF*. O número de indivíduos na população não influencia o ordenamento e sim as estimativas das freqüências de recombinação (Capítulo 2).

No Capítulo 3, estimaram-se os parâmetros de adaptabilidade e estabilidade, utilizados para estudo da interação genótipos x ambientes, do modelo proposto por Finlay e Wilkinson (1963), fazendo uso da inferência bayesiana, evidenciando seus passos, e a utilização do amostrador de Gibbs através do software WINBUGS. Mostrou-se que o amostrador de Gibbs é eficiente para amostrar das distribuições marginais completas, possibilitando obter estimativas pontuais iguais ou semelhantes a obtidas pelo método dos mínimos quadrados ordinários. Além disso, observou-se que os intervalos de credibilidade em geral obtiveram menores amplitudes, indicando uma maior precisão na estimação dos parâmetros.

Em todas as aplicações os algoritmos apresentaram resultados satisfatórios, mostrando-se úteis nas diversas situações apresentadas. Desta forma demonstrou-se que tais métodos são uma ferramenta extremamente útil aos melhoristas.