MODELO OCULTO DE MARKOV PARA IMPUTAÇÃO DE GENÓTIPOS DE MARCADORES MOLECULARES

Elias Silva de MEDEIROS[1]
Roseli Aparecida LEANDRO[2]

■ RESUMO: Muitas são as características quantitativas que são, significativamente, influenciadas por fatores genéticos. O mapeamento genético tem sido utilizado para identificar regiões do genoma que contribuem de forma direta no desenvolvimento destas características. Os experimentos com marcadores moleculares podem fornecer em seu conjunto de dados, uma sequência de dados genotípicos sem mensuração. Estes dados não observados podem ser resultados de erros de genotipagem ou marcadores não informativos. Consequentemente, estas informações ausentes a respeito dos genótipos nos marcadores é um problema que pode acarretar dificuldades no estudo do mapeamento do genoma. Assim, este artigo tem por objetivo realizar a imputação dos genótipos dos marcadores, propondo na modelagem o modelo oculto de Markov (HMM) para inferir estes dados. Para aplicação desta metodologia foi considerado um conjunto de dados de marcadores moleculares microssatélites. A abordagem HMM para imputação dos genótipos ausentes nos marcadores apresentou resultados significativos por meio de simulações, uma vez que as medidas para avaliarem a acurácia da imputação neste estudo, evidenciaram um bom desempenho com relação a imputação destes genótipos. Adicionalmente, a metodologia proposta neste artigo torna-se uma alternativa à imputação nos marcadores, a qual pode ser empregada em marcadores SNPs (*Single Nucleotide Polymorphisms*) obtidos a partir de genotipagem por sequenciamento (GBS) para espécies que ainda não dispõem do genoma sequenciado, resultando em uma redução dos custos de genotipagem, em especial, para implementação da seleção genômica.

■ PALAVRAS-CHAVE: Genoma sequenciado; cadeias de Markov; erro de genotipagem.

[1]Universidade Federal da Grande Dourados - UFGD, Faculdade de Ciências Exatas e Tecnologia, CEP: 79804-970, Dourados, MS, Brasil. E-mail: *eliasmedeiros@ufgd.edu.br*
[2]Universidade de São Paulo - USP, Escola Superior de Agricultura Luiz de Queiroz, Departamento de Ciências Exatas - LCE, CEP: 13418-900, Piracicaba, SP, Brasil. E-mail: *raleandr@usp.br*

1 Introdução

Um dos principais objetivos dos estudos que envolvem o mapeamento genético, consiste na identificação de regiões do genoma que influenciam, por meios dos seus genes, as variações fenotípicas de uma certa característica de interesse. Com o advento de novas tecnologias, nos últimos anos diversos trabalhos têm sido conduzidos para analisarem a associação ampla do genoma (Genome-Wide Associations Studies - GWAS) com as características fenotípicas, sendo que agora o mapeamento genérico consiste na presença de milhares de marcadores distribuídos ao longo de todo o genoma (SAHANA et al., 2010). Entretanto, os conjuntos de dados utilizados no mapeamento genético, podem conter uma quantidade significativa de genótipos ausentes nos marcadores. Por exemplo, os dados de SNPs (*Single Nucleotide Polymorphisms*) obtidos a partir de genotipagem por sequenciamento (GBS), geralmente contêm uma grande porcentagem de dados perdidos (ELBASYONI et al., 2018). Estes dados ausentes ocorrem, principalmente, devido aos erros de genotipagem e de marcadores não informativos. Na prática, existem algumas alternativas para lidar com este tipo de problema, tais como, repetir a genotipagem em regiões com genótipos ausentes (as vezes inviável, devido ao alto custo operacional); remover os marcadores que possuem genótipos ausentes (implicam perdas de informações); e o mais aconselhado, imputar os dados ausentes (ROBERTS et al., 2007; CHUD et al., 2015; ROSHYARA et al., 2016).

Diferentes métodos de imputação de genótipos têm sido utilizados para realizarem inferências em dados genômicos com o intuito de obter bons resultados na seleção genômica (SARGOLZAEI; CHESNAIS; SCHENKEL, 2014; PEREIRA et al., 2017). Os experimentos que utilizam de marcadores SNPs via GBS na genotipagem proporcionam uma grande quantidade de marcadores, entretanto estes experimentos apresentam um número significativo de dados perdidos, chegando a ser de até 70% de observações sem registros (RUTKOSKI et al., 2013). Assim, a imputação de genótipos é de suma importância para realizar inferências com o intuito de melhorar a precisão da predição genômica (HAYES et al., 2011; ELBASYONI et al., 2018).

Contudo, as informações ausentes a respeito dos genótipos nos marcadores moleculares pode ser considerada como um problema comum em estudo de mapeamento genético e, por conseguinte, na identificação de regiões do genoma que contribuem à variação de uma ou mais características fenotípicas de uma espécie. Diante disso, para solucionar este problema se faz necessária à utilização de técnicas de imputação para inferir os dados desses genótipos (BROWNING; BROWNING, 2009; VENTURA et al., 2016; ELBASYONI et al., 2018).

Entretanto, ao trabalhar com conjunto de dados reais, a grande dificuldade consiste sobre a inferência da acurácia das imputações. Como solução, os métodos de imputação podem ser conduzidos no próprio conjunto amostral, fazendo com que seja possível mensurar a acurácia, bem como estimar a confiabilidade das imputações (PEREIRA et al., 2017). Na literatura são apresentados diversos programas computacionais, baseados em métodos estatísticos, para imputação

de genótipos dos marcadores, dentre estes destacam-se o *FastPHASE* (SERVIN; STEPHENS, 2007), o *IMPUTE2* (HOWIE; DONNELLY; MARCHINI, 2009) e o *Beagle* (BROWNING; BROWNING, 2009). Estes programas têm em comum o fato de que na sua metodologia utilizam-se dos modelos ocultos de Markov (HMM) para inferir os genótipos ausentes nos marcadores (WHALEN et al., 2018). Entretanto, estes programas computacionais não apresentam com detalhes como é realizada a inferência às imputações.

O principal objetivo desse trabalho é explorar as eficiências dos modelos HMM como método de imputação dos genótipos, detalhando como serão utilizadas as informações dos marcadores para construir as matrizes das probabilidades de transição e de emissão que constituem esta metodologia.

2 Material

O conjunto de dados utilizado neste trabalho foram obtidos de um experimento realizado por Sibov et al. (2004) e Sabadin et al. (2008). Em resumo, foi obtida uma população a partir do cruzamento entre duas linhagens endogâmicas $L-08-05F$ e $L-14-4B$, que possuem comportamento contrastante para produção de grãos de milho. Essas linhas são provenientes das populações tropicais $IG-1$ e $BR-106$, respectivamente.

A progênie F_1 foi obtida cruzando as duas linhagens endogâmicas, sendo que quatro plantas desta geração foram autofecundadas para gerar 400 plantas da população F_2 das quais foram obtidas 400 progênies $F_{2:3}$, que foram cruzadas entre si e semeadas em fileiras contendo 20 plantas para aumentar a quantidade de sementes necessárias para avaliações experimentais (SIBOV et al., 2004). O experimento foi conduzido na cidade de Piracicaba, estado de São Paulo, Brasil, durante as safras de 1999 e 2000. As 400 progênies foram divididas em quatro grupos com 100 progênies cada e cada grupo foi avaliado um delineamento látice 10×10, com duas repetições cada um. Para mais detalhes técnicos deste experimento consulta os trabalhos de Sibov et al. (2004) e Sabadin et al. (2008).

Para imputação dos genótipos dos marcadores foi utilizado um mapa de ligação composto por 117 lócus de marcadores microssatélites, os quais foram distribuídos em dez grupos de ligação. O mapa genético apresentou com comprimento de 1634,20 cM e distância média entre as marcas de 14 cM. Embora nos dias atuais não se utilize mais de marcadores microssatélites no mapeamento do genoma, a teoria desenvolvida para solucionar o problema de observações em falta pode ainda ser utilizada na atual era de marcadores SNPs, uma vez que estes marcadores apresentam uma quantidade significativa de dados ausentes, em especial quando a genotipagem é realizada via GBS.

3 Métodos

3.1 HMM para imputação

As análises de marcadores moleculares que contêm informações genotípicas do indivíduo são importantes para identificar associações de genes. Os grandes conjuntos de dados derivados desses marcadores pode apresentar uma quantidade significativa de genótipos ausentes. Para solucionar este problema de observações ausentes se faz necessária à utilização de técnicas de imputação para inferir os dados desses genótipos (HOWIE; MARCHINI; STEPHENS, 2011; PEREIRA et al., 2017). Neste trabalho é proposta uma técnica de imputação baseadas nos modelos HMM.

Na Figura 1 é apresentado um esquema do modelo HMM, o qual foi estruturado neste trabalho da seguinte forma: g_i representando um estado não observado da cadeia de Markov, o_i uma variável aleatória observável, sendo que o_i depende apenas de g_i. Os elementos a_{ij} e e_{ik} representam as probabilidades de transição e de emissão, respectivamente.

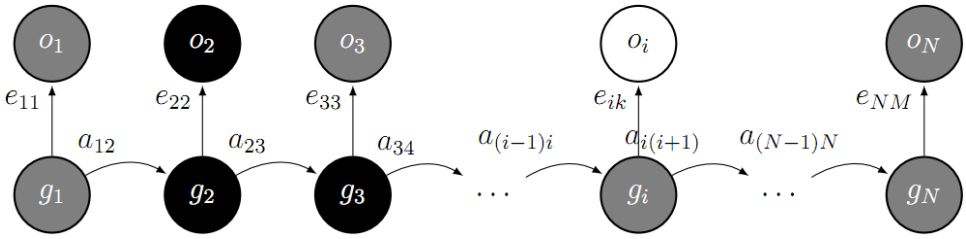


Figura 1 - Ilustração de uma cadeia de Markov oculta.

Um dos algoritmos mais tradicionais em um modelo HMM é o algoritmo *forward*. Este calcula a probabilidade de ocorrer toda a sequência de observações $\mathcal{O}$ dado o modelo, $P(\mathcal{O}|\boldsymbol{\theta})$ (ZHAO et al., 2017). O algoritmo consiste em definir uma sequência "ótima" de estados, dada uma sequência de observações $\mathcal{O}$ e um conjunto de parâmetros $\boldsymbol{\theta}$, $\gamma_t(i) = P(g_t = S_i|\mathcal{O}, \boldsymbol{\theta})$. A variável $\gamma_t(i)$ é definida como a probabilidade de iniciar o estado S_i na posição t, dado uma sequência observada $\mathcal{O}$ e o conjunto $\boldsymbol{\theta}$. Para determinar o estado mais provável de ocorrer na posição t, tem-se:

$$g_t = \underset{1 \leq i \leq N}{\operatorname{argmax}} \left[\gamma_t(i)\right], 1 \leq t \leq T. \tag{1}$$

A Equação 1 não assegura que a sequência de estados escolhida possa ser a "ideal". Por exemplo, pode ocorrer o fato de a probabilidade de transição seja igual a zero, resultando em uma sequência de estados com um estado inválido (MEDEIROS, 2014). Este problema consiste no fato de que a Equação 1 seleciona o estado mais provável para cada instante, sem levar em consideração a probabilidade de ocorrência de toda a sequência de estado. Assim, para determinar uma sequência "ideal" plausível, tem-se a necessidade de utilizar técnica de programação dinâmica.

Neste artigo, para resolução deste problema foi utilizado o algoritmo de *Viterbi*, o qual é computacionalmente eficiente para determinar a sequência mais provável de estados (CARRER; BRUZZONE, 2017). Neste estudo, a programação do algoritmo de *Viterbi* seguiu os seguintes passos:

a. Inicialização

$$\delta_1(i) = \pi_{g_i} \times e_{g_i}(o_i), 1 \leq i \leq N$$
$$\psi_1(i) = 0.$$

No passo (a), π_{g_i} e e_{g_i} são as probabilidades inicial e de emissão, respectivamente. A variável $\delta_t(i)$ representa a probabilidade máxima de uma única sequência de dentre todas as possíveis que terminam no estado S_i no tempo t, $\delta_t(i) = \max_{g_1, g_2, \cdots, g_{t-1}} P\left[g_1, g_2, \cdots, g_t = S_i, o_1, o_2, \cdots, o_t | \boldsymbol{\theta}\right]$. A segunda variável $\psi_t(i)$, tem por finalidade permitir acompanhar a melhor sequência final no estado S_i no tempo t, $\psi_t(i) = \operatorname{argmax}_{g_1, g_2, \cdots, g_{t-1}} P\left[g_1, g_2, \cdots, g_t = S_i, o_1, o_2, \cdots, o_t | \boldsymbol{\theta}\right]$.

b. Recursão

$$\delta_t(j) = \max_{1 \leq i \leq N} \left[\delta_{t-1}(i) \times a_{ij}\right] \times e_{g_j}(o_t)$$
$$\psi_t(j) = \operatorname*{argmax}_{1 \leq i \leq N} \left[\delta_{t-1}(i) \times a_{ij}\right].$$

No passo (b), tem-se que $2 \leq t \leq T$ e $1 \leq j \leq N$. O elemento a_{ij} está associado a probabilidade de transição. Para a imputação dos genótipos as probabilidades inicial, de transição e de emissão estão detalhadas na seção 3.2.

c. Terminação

$$P^*(\mathcal{O}|\boldsymbol{\theta}) = \max_{1 \leq i \leq N} \delta_T(i)$$
$$g_T^* = \operatorname*{argmax}_{1 \leq i \leq N} \left[\psi_T(i)\right].$$

d. Retrocedendo

$$G^* = \{g_1^*, g_2^*, \cdots, g_T^*\}$$

Neste trabalho, para determinar as estimativas dos parâmetros do modelo HMM, foi utilizado o algoritmo de *Viterbi Training* (VT), também conhecido como o algoritmo *K*-média, uma vez que este é computacionalmente menos intenso e mais estável em comparação com o algoritmo EM (CARRER; BRUZZONE, 2017; MEDEIROS, 2014).

3.2 HMM para imputação de genótipos

Para exemplificar a aplicação do HMM para imputação de genótipos nos marcadores, é apresentado a seguir uma estruturação de um modelo HMM em um população F_1.

- Considere um indivíduo originário de um retrocruzamento de duas linhagens puras, A e B, em que o pai F_1 foi cruzado novamente com A. Sendo assim, os possíveis valores genotípicos são, $\mathcal{G} = \{\mathcal{AA}, \mathcal{AB}\}$.

- O conjunto dos símbolos que são emitidos, sequência observável, é expresso por, $\mathcal{O} = \{\mathcal{A}, \mathcal{H}, \texttt{NA}\}$. Assim, $\mathcal{AA}$ emitirá o símbolo $\mathcal{A}$, $\mathcal{AB}$ emitirá H. Já NA representando um valor não observado.

- As probabilidades iniciais assumindo as regras de Mendel são $\pi(AA) = \pi(AB) = 1/2$. As probabilidades de transição ficam em função da fração de recombinação r, $a_{ij} = r$, para $i \neq j$. Naturalmente, $a_{ij} = 1 - r$, para $i = j$. Para determinar as expressões para as probabilidades de emissão, assume-se uma taxa de erro constante na genotipagem (ϵ), então $e_{g_i}(AA, A) = e_{g_i}(AB, H) = 1 - \epsilon$, e $e_{g_i}(AA, H) = e_{g_i}(AB, A) = \epsilon$. Tem-se ainda que, $e_{g_i}(AA, \texttt{NA}) = e_{g_i}(AB, \texttt{NA}) = 1$, pois $\texttt{NA} = \{A \text{ ou } H\}$ de modo que $e_{g_i}(AA, \texttt{NA}) = e_{g_i}(AA, A) + e_{g_i}(AA, H) = 1$.

Na Tabela 1 são apresentadas as expressões que foram utilizadas para o cálculo das probabilidade de transição no conjunto de dados em estudo, oriundos de uma população F_2.

Tabela 1 - Probabilidades de transição em uma população F_2

g	g'		
	AA	AB	BB
AA	$(1-r)^2$	$2r(1-r)$	r^2
AB	$r(1-r)$	$(1-r)^2 + r^2$	$r(1-r)$
BB	r^2	$2r(1-r)$	$(1-r)^2$

Em uma população F_2 os possíveis símbolos observados foram considerados $\mathcal{O}_d = \{A, H, B, \texttt{NA}\}$, com A, B e H correspondentes a dois homozigotos e um heterozigoto, respectivamente, NA corresponde a um valor completamente ausente e $\mathcal{G}_d = \{AA, AB, BB\}$ os possíveis valores genotípicos (Tabela 2).

De acordo com as regras de Mendel, as probabilidades iniciais foram $\pi(AA) = \pi(BB) = 1/4, \pi(AB) = 1/2$. Neste trabalho para o parâmetro relacionado as probabilidades de transição foi utilizado um valor inicial de $0,10$. Já para o parâmetro que compõe as probabilidades de emissão foi inserido um valor inicial de $0,01$. Para as probabilidades iniciais, foram utilizados valores inicias conforme regra de Mendel. A quantidade máxima de iterações para convergência do algoritmo VT foi de tamanho 100, uma vez que neste algoritmo um número acima de 60 já

Tabela 2 - As probabilidades de emissão em uma população F_2

g	$\mathcal{O}_d$			
	A	H	B	NA
AA	$1-\epsilon$	$\epsilon/2$	$\epsilon/2$	1
AB	$\epsilon/2$	$1-\epsilon$	$\epsilon/2$	1
BB	$\epsilon/2$	$\epsilon/2$	$1-\epsilon$	1

é considerada uma quantidade significativa de iterações (HUMBURG; BULGER; STONE, 2008). A convergência do algoritmo consistiu quando a diferença das estimativas dos parâmetros de transição e de emissão entre as iterações consecutivas foi menor do que 10^{-9}. Neste manuscrito, o algoritmo VT foi constituído dos seguintes passos:

a. Atribuiu valores iniciais para os parâmetros do modelo.

b. Obteve uma sequência de estados mais provável G por meio do algoritmo de *Viterbi*.

c. Calculou-se as probabilidades de transição (a_{ij}) e de emissão (e_{g_i}), dado o estado G.

d. Estimou-se os parâmetros do novo modelo usando as ocorrências estimadas dos estados de transição e de emissão e retornou-se ao passo (b).

3.3 Estudo de simulação

Após realizada a imputação no conjunto de dados utilizado neste estudo, utilizou-se deste mesmo conjunto para criar diferentes cenários com o intuito de mensurar a acurácia da imputação.

De forma aleatória, foi retirada do conjunto de dados reais, uma certa porcentagem de observações, correspondente aos valores de 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35% e 40%. Para cada percentual de dados "deixados de fora", foram realizadas 1000 replicações. Desta forma, ao todo foram construídos 9000 conjuntos de dados que possuíam dados ausentes, sendo que para cada conjunto foi aplicado o método de imputação utilizando a metodologia HMM.

A seguir é apresentada as métricas utilizadas para avaliar a acurácia da imputação.

3.4 Métodos para avaliar a acurácia

Para validação do método empregado para imputação dos genótipos dos marcadores moleculares serão utilizadas duas medidas descritas a seguir.

A raiz quadrada do erro quadrático médio normalizado - NRMSE (do inglês, *normalized root mean squared error*) foi calculada para determinar a acurácia da

imputação (MARTÍNEZ-CAÑADA et al., 2017). A NRMSE foi obtida de acordo com a seguinte expressão:

$$NRMSE = \sqrt{\frac{\frac{1}{Q}\sum_{q=1}^{Q}\left(g_q - \hat{g}_q\right)^2}{\max(g_q) - \min(g_q)}}. \quad (2)$$

Na equação (2), $q = 1, 2, \ldots, Q$ representa a quantidade de valores a serem imputados, g_q representa o valor real que foi ocultado da matriz completa dos genótipos dos marcadores e o seu respectivo valor imputado $\hat{g}_q$. Quanto menor for a $NRMSE$, melhor será para a validação na acurácia da imputação. Para uma avaliação cuidadosa da eficiência do algoritmo de imputação, além do cálculo da NRMSE para cada um dos 9000 conjuntos de dados, foi apresentada a taxa de concordância.

Os genótipos dos marcadores imputados foram comparados com os genótipos dos marcadores reais presentes no conjunto dos dados reais, e assim a proporção de genótipos que foram imputados corretamente ou erroneamente foi calculada. A taxa de concordância representou a proporção de genótipos que foram corretamente imputados. Adicionalmente, é comum, em estudos que realizam a acurácia da imputação, a utilização desta taxa (CHUD et al., 2015; VENTURA et al., 2016).

3.5 Análise computacional

Todas as análises estatísticas foram conduzidas no programa estatístico R (R CORE TEAM, 2018). Para imputação dos genótipos dos marcadores utilizou-se da biblioteca *HMM* (HIMMELMANN, 2015). Adicionalmente, este programa pode ser utilizado para construção do mapas genéticos utilizando a biblioteca *onemap* (MARGARIDO; SOUZA; GARCIA, 2007).

4 Resultados e discussões

4.1 Análise exploratória

O mapa genético foi composto de 117 marcadores microssatélites alocados em 10 cromossomos (Figura 2). Este mapa apresentou um comprimento total de $1634,20$ cM e distância média entre as marcas de 14 cM. O comprimento dos cromossomos variou de $89,10$ cM (cromossomo 10) a $242,80$ cM (cromossomo 1) e o número de marcas em cada cromossomo variou de 6 (cromossomo 10) a 18 (cromossomo 1). Nota-se ainda que, as marcas encontram-se distribuídas de forma aleatória por todo o genoma.

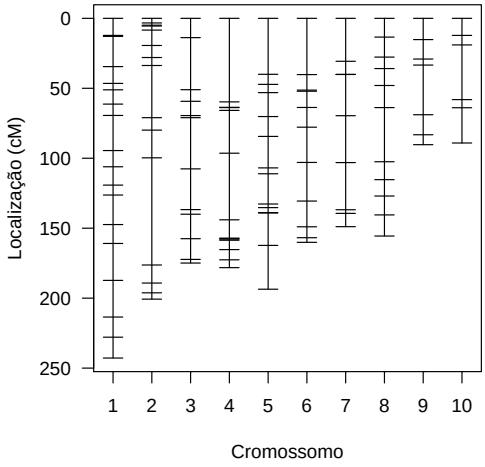
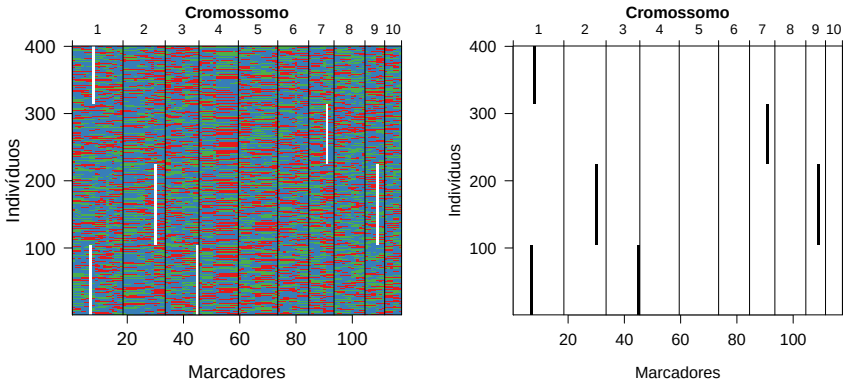


Figura 2 - Mapa genético contendo os 117 marcadores distribuídos ao longo dos 10 cromossomos.

4.2 Imputação dos genótipos

Na Figura 3(a) tem-se uma representação gráfica da matriz dos genótipos dos marcadores observados. O conjunto de dados é constituído de 400 indivíduos e 117 marcadores, ou seja, cada indivíduo foi observado 117 vezes, resultando numa matriz de ordem 400×117. Retornando a esta figura, os tons em branco, aproximadamente, 2% dos dados, representam a ausência de genótipos nos marcadores. Embora esta seja uma quantidade insignificante de dados ausentes, a título de aplicação foi considerado este conjunto de dados reais para realização de simulações, retirando uma certa parte dos dados para depois imputá-los e calcular a acurácia da imputação. Contudo, nos últimos anos o avanço das novas tecnologia tem possibilitado que no mapeamento seja utilizado milhares de marcadores SNPs, genotipados por sequenciamento de alto desempenho. Porém, este tipo de método de genotipagem apresentam, na maioria dos experimentos, uma quantidade significativa de dados em falta, podendo ser de até 70% ou mais de dados ausentes (RUTKOSKI et al., 2013; ELBASYONI et al., 2018). Logo, faz-se necessária a utilização de técnicas de imputação em conjunto de dados genômicos, com o intuito de melhorar a predição genômica em programas de melhoramento genético e reduzir os custos de genotipagem (SARGOLZAEI; CHESNAIS; SCHENKEL, 2014; CHUD et al., 2015).



(a) (b)

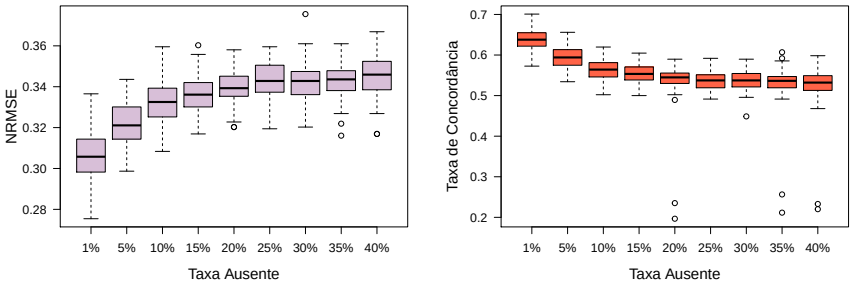
Figura 3 - Representação gráfica da matriz dos marcadores observados (a), sendo os genótipos AA, AB, BB exibidos nas cores vermelha, azul e verde, respectivamente; e uma grande mostrando quais genótipos estão ausentes (b).

4.3 Acurácia da imputação

No mapeamento genético a imputações em dados reais pode ocasionar grandes desafios para mensurar a acurácia da imputação. Assim, pode-se estar realizando simulações utilizando as informações do próprio conjunto amostral dos dados reais com o objetivo de apurar a qualidade da metodologia proposta para imputação. Neste artigo, após realizada a imputação na matriz dos dados reais dos genótipos foi realizado um estudo de simulação, no qual foram retiradas diferentes porcentagens de observações que variaram de 1% a 40%. Com base nas simulações foram calculadas as estatísticas NRMSE e a taxa de concordância nos diferentes cenários.

Os trabalhos que utilizaram da imputação de genótipos, em especial, em dados SNPs têm preferido a medida de acurácia baseada na correlação alélica (SARGOLZAEI; CHESNAIS; SCHENKEL, 2014). Entretanto, para obtenção de estimativas não viesadas, o cálculo desta correlação requer uma grande quantidade de indivíduos que formam o conjunto de dados a ser imputado, caso contrário os valores destas correlações podem apresentam altos valores de erros-padrões (VENTURA et al., 2016). Dessa forma, além do cálculo da taxa de concordância, foi utilizada a estatística NRMSE para aferir a acurácia.

Na Figura 4 tem-se os gráficos de box plot para as estatísticas NRMSE e da taxa de concordância. Cada gráfico box plot é referente ao cálculo de cada uma das estatísticas supracitadas, com base em mil replicações conforme descritas na seção 3.3. Notou-se que, quando inserido 1% de dados ausentes nos dados amostrados a estatística NRMSE (Figura 4(a)) apresentou uma média de 0,306 com um desvio



(a) (b)

Figura 4 - Box-plot da NRMSE (a) e a Taxa de Concordância (b) para os diferentes cenários de simulação.

padrão de 0,013. Quando considerada a taxa de 40% de observações ausentes tem-se que a NRMSE apresentou uma média de 0,345 com um desvio padrão de 0,010. Com relação a Taxa de Concordância (Figura 4(b)), tem-se que a medida que a taxa de observações ausentes aumentava a Taxa de Concordância diminuía cada vez mais. Estes resultados comprovam que os modelos HMM para imputação apresentam melhores acurácias quanto maior for a quantidade de dados observados.

Os programas computacionais que utilizam para imputação da metodologia HMM, apresentam restrições computacionais que estão relacionadas ao processo de amostragem intensivo (SARGOLZAEI; CHESNAIS; SCHENKEL, 2014). No trabalho destes autores supracitados foi proposto um novo método de imputação e comparado com o desempenho de dois *softwares*, *Beagle* e *Impute2*, sendo que ambos os programas são baseados em modelos HMM. Entretanto, não foi apresentado no trabalho destes autores detalhes do algoritmo que é utilizado nos dois programas e nem como foram construídas as matrizes de transição e de emissão. Vale salientar que ao utilizar um *software*, em especial, para imputação de genótipos, em geral os usurários não têm conhecimento dos detalhes que envolvem a construção do algoritmo, bem como não existe a possibilidade de estar inserindo novas abordagens na realização das imputações.

Considerações finais

Neste artigo, a imputação foi realizada utilizando os modelos HMM, apresentando detalhes da construção das matrizes de transição e de emissão, bem como o passo-a-passo do algoritmo de VT para determinar as melhor sequência de estados. Em todas as análises deste manuscrito foram construídas rotinas no programa R (R CORE TEAM, 2018), o qual possui um plataforma de acesso livre e

aberto, sendo possível aos demais pesquisadores estarem utilizando para imputação de genótipos, como também estarem inserindo novas estruturas no modelo HMM. Adicionalmente, os resultados indicam que o método proposto neste estudo para imputação tem alta precisão e que pode ser utilizado em conjunto de dados com milhares de marcadores que é comum, por exemplo, em espécies de gado.

Agradecimentos

Os autores agradecem aos revisores e editores pelos comentários e sugestões que ajudaram a melhorar a qualidade do trabalho.

MEDEIROS, E. S.; LEANDRO, R. A. hidden Markov model for imputation of molecular marker genotypes. *Rev. Bras. Biom.,* Lavras, v.37, n.1, p.107-120, 2019.

- *ABSTRACT: Several are the quantitative traits that are significantly influenced by genetic factors. Genetic mapping has been used to identify regions of the genome that contribute directly to the development of these characteristics. Molecular marker experiments can provide in their data set a sequence of unmeasured genotypic data. These missing data may be the result of genotyping errors or non-informative markers. Consequently, this missing information about the genotypes in the markers is a problem that can lead to difficulties in the study of genome mapping. Thus, this paper aims to impute the genotypes of the markers, proposing in the modeling the Hidden Markov Model (HMM) to infer this data. For the application of this methodology, a set of microsatellite molecular marker data was considered. The HMM approach to imputation of the missing genotypes in the markers presented significant results through simulations, since the measures to evaluate the imputation accuracy in this study, evidenced a good performance regarding the imputation of these genotypes. In addition, the methodology proposed in this article becomes an alternative to imputation in the markers, which can be employed in SNPs (t Single Nucleotide Polymorphisms) obtained from genotyping by sequencing (GBS) for species that do not yet have the sequenced genome, resulting in a reduction in genotyping costs, especially for the implementation of genomic selection.*

- *KEYWORDS: Genome sequenced; Markov chains; genotyping error.*

Referências

BROWNING, B. L.; BROWNING, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, v.84, n.2, p.210-223, 2009.

CARRER, L.; BRUZZONE, L. Automatic enhancement and detection of layering in radar sounder data based on a local scale Hidden Markov Model and the Viterbi

algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, Institute of Electrical and Electronics Engineers (IEEE), v.55, n.2, p.962-977, 2017.

CHUD, T. C. S. et al. Strategies for genotype imputation in composite beef cattle. *BMC Genetics*, v.16, n.1, 2015.

ELBASYONI, I. S. et al. A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. *Plant Science*, v.270, p.123-130, 2018.

HAYES, B. J. et al. Accuracy of genotype imputation in sheep breeds. *Animal Genetics*, v.43, n.1, p.72-80, 2011.

HIMMELMANN, L. *HMM: HMM - Hidden Markov Models*. [S.l.], 2015. R package version 1.0. Disponível em: ⟨https://CRAN.R-project.org/package=HMM⟩.

HOWIE, B.; MARCHINI, J.; STEPHENS, M. Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics*, v.1, n.6, p.457-470, 2011.

HOWIE, B. N.; DONNELLY, P.; MARCHINI, J. A flexible and accurate genotype imputation method for the next generation of Genome-Wide Association studies. *PLoS Genetics*, v.5, n.6, p.e1000529, 2009.

HUMBURG, P.; BULGER, D.; STONE, G. Parameter estimation for robust HMM analysis of ChIP-chip data. *BMC Bioinformatics*, v.9, n.1, p.343, 2008.

MARGARIDO, G. R. A.; SOUZA, A. P.; GARCIA, A. A. F. OneMap: software for genetic mapping in outcrossing species. *Hereditas*, v.144, n.3, p.78-79, 2007.

MARTÍNEZ-CAÑADA, P. et al. Genetic algorithm for optimization of models of the early stages in the visual system. *Neurocomputing*, v.250, p.101-108, aug 2017.

MEDEIROS, E. S. *Modelo oculto de Markov para imputação de genótipos de marcadores moleculares: Uma aplicação no mapeamento de QTL utilizando a abordagem bayesiana*. Tese (Doutorado) — Universidade de São Paulo, 2014.

PEREIRA, G. L. et al. Genotype imputation and accuracy evaluation in racing quarter horses genotyped using different commercial SNP panels. *Journal of Equine Veterinary Science*, v.58, p.89-96, 2017.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: ⟨https://www.R-project.org/⟩.

ROBERTS, A. et al. Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. *Bioinformatics*, v.23, n.13, p.i401-i407, 2007.

ROSHYARA, N. R. et al. Comparing performance of modern genotype imputation methods in different ethnicities. *Scientific Reports*, v.6, n.1, 2016.

RUTKOSKI, J. E. et al. Imputation of unordered markers and the impact on genomic selection accuracy. *G3: Genes, Genomes, Genetics*, v.3, n.3, p.427-439, 2013.

SABADIN, P. K. et al. QTL mapping for yield components in a tropical maize population using microsatellite markers. *Hereditas*, v.145, n.4, p.194-203, 2008.

SAHANA, G. et al. Genome-wide association mapping for female fertility traits in Danish and Swedish Holstein cattle. *Animal Genetics*, v.41, n.6, p.579-588, 2010.

SARGOLZAEI, M.; CHESNAIS, J. P.; SCHENKEL, F. S. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics*, v.15, n.1, p.478, 2014.

SERVIN, B.; STEPHENS, M. Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genetics*, v.3, n.7, p.e114, 2007.

SIBOV, S. T. et al. Molecular mapping in tropical maize (Zea mays L.) using microsatellite markers. 1. Map construction and localization of loci showing distorted segregation. *Hereditas*, v.139, n.2, p.96-106, 2004.

VENTURA, R. V. et al. Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. *Genetics Selection Evolution*, v.48, n.1, 2016.

WHALEN, A. et al. Assessment of the performance of hidden markov models for imputation in animal breeding. *Genetics Selection Evolution*, v.50, n.1, 2018.

ZHAO, J. et al. Software abnormal behavior detection based on Hidden Markov Model. In: *Innovative Mobile and Internet Services in Ubiquitous Computing*. p.929-940, 2017.

Recebido em 13.01.2018.

Aprovado após revisão em 06.11.2018.