

Comparação dos algoritmos delineação rápida em cadeia e seriação, para a construção de mapas genéticos

Marcelo Mollinari⁽¹⁾, Gabriel Rodrigues Alves Margarido⁽¹⁾ e Antonio Augusto Franco Garcia⁽¹⁾

⁽¹⁾Escola Superior de Agricultura Luiz de Queiroz, Departamento de Genética, Caixa Postal 83, CEP 13418-900 Piracicaba, SP. E-mail: mmollina@carpa.ciagri.usp.br, gramarga@carpa.ciagri.usp.br, aafgarci@carpa.ciagri.usp.br

Resumo – O objetivo deste trabalho foi avaliar a eficiência, na construção de mapas genéticos, dos algoritmos seriação e delineação rápida em cadeia, além dos critérios para avaliação de ordens: produto mínimo das frações de recombinação adjacentes, soma mínima das frações de recombinação adjacentes e soma máxima dos LOD Scores adjacentes, quando usados com o algoritmo de verificação de erros “ripple”. Foi simulado um mapa com 24 marcadores, posicionados aleatoriamente a distâncias variadas, com média 10 cM. Por meio do método Monte Carlo, foram obtidas 1.000 populações de retrocruzamento e 1.000 populações F₂, com 200 indivíduos cada, e diferentes combinações de marcadores dominantes e co-dominantes (100% co-dominantes, 100% dominantes e mistura com 50% co-dominantes e 50% dominantes). Foi, também, simulada a perda de 25, 50 e 75% dos dados. Observou-se que os dois algoritmos avaliados tiveram desempenho semelhante e foram sensíveis à presença de dados perdidos e à presença de marcadores dominantes; esta última dificultou a obtenção de estimativas com boa acurácia, tanto da ordem quanto da distância. Além disso, observou-se que o algoritmo “ripple” geralmente aumenta o número de ordens corretas e pode ser combinado com os critérios soma mínima das frações de recombinação adjacentes e produto mínimo das frações de recombinação adjacentes.

Termos para indexação: algoritmo ripple, dados perdidos, marcadores dominantes e co-dominantes, método Monte Carlo, QTL.

Comparison of algorithms rapid chain delineation and seriation, for the construction of genetic linkage maps

Abstract – The objective of this work was to evaluate the efficiency for the construction of genetic linkage maps of the algorithms seriation and rapid chain delineation, as well as the criteria: product of adjacent recombination fractions, sum of adjacent recombination fractions, and sum of adjacent LOD Scores, used with the ripple algorithm. A genetic linkage map was simulated containing 24 markers with random distances between them, with an average of 10 cM. Using the Monte Carlo method, 1,000 backcross populations and 1,000 F₂ populations were simulated. The populations comprised 200 individuals each, as well as different combinations of dominant and codominant markers (100% codominant, 100% dominant and mixture containing 50% codominant and 50% dominant). It was also simulated 25, 50 e 75% of missing data. It was observed that both algorithms presented similar performance, and were sensitive to the presence of dominant markers, which makes it difficult to get estimates with good accuracy for both order and distance. Moreover, the algorithm ripple, when applied with the criteria sum of adjacent recombination fractions and product of adjacent recombination fractions, increased the number of correct orders.

Index terms: ripple algorithm, missing data, dominant and codominant markers, Monte Carlo method, QTL.

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. Pela elevada importância econômica dos QTL, seu

mapeamento é usualmente considerado como uma etapa inicial, antes da realização da seleção assistida por marcadores.

A recente grande disponibilidade de marcadores moleculares tornou possível o mapeamento genético em larga escala. Alguns dos marcadores usados no mapeamento são “restriction fragment length polymorphism” (RFLP) (Botstein et al., 1980), “randomly

amplified polymorphic DNA” (RAPD) (Williams et al., 1990), “amplified fragment length polymorphism” (AFLP) (Zabeau & Vos, 2005) e “simple sequence repeats” (SSR) (Tautz, 1989). Outros tipos de marcadores, como single nucleotide polymorphism (SNP) (Syvänen, 2001), diversity arrays technology (DartTs) (Jaccoud et al., 2001) e expressed sequence-tag-derived SSR (EST-SSRs) (Cato et al., 2001), têm se tornado cada vez mais frequentes em estudos genéticos.

Uma das etapas que merece mais atenção na construção de mapas de ligação é a ordenação dos marcadores genéticos dentro de cada grupo de ligação. A ordenação de marcadores em mapas é considerada um caso especial do clássico problema do caixeiro viajante (“traveling salesman problem” – TSP) (Liu, 1998), que consiste em escolher a melhor ordem entre $m!/2$ ordens possíveis, em que m é o número de marcadores. Em problemas reais, m pode variar de dezenas até centenas (Mester et al., 2003), o que torna este cálculo inviável. Diversos algoritmos foram propostos para realizar a ordenação em grupos com grande número de marcadores, entre eles: seriação (“seriation” – SER) (Buetow & Chakravarti, 1987) e delineação rápida em cadeia (“rapid chain delineation” – RCD) (Doerge, 1996), os quais não consideram todas as ordens possíveis.

Lander et al. (1987) propuseram o algoritmo “ripple”, para verificação de subordens alternativas de janelas de m' marcadores, em um mapa previamente construído com m marcadores ($m' < m$), por meio de buscas locais exaustivas que considerassem as $m!/2$ ordens possíveis. No entanto, é necessário combinar o “ripple” com a escolha de algum critério para avaliação dessas subordens. Vários critérios podem ser utilizados: soma mínima das frações de recombinação adjacentes (“sum of adjacent recombination fractions” – SARF) (Falk, 1989); produto mínimo das frações de recombinação adjacentes (“product of adjacent recombination fractions” – PARF) (Wilson, 1988); soma máxima dos LOD Scores adjacentes (“sum of adjacent LOD Scores” – SALOD) (Weeks & Lange, 1987).

Em situações práticas, os geneticistas não podem verificar se a ordem dos marcadores do mapa de ligação, obtido com o uso dessas diferentes abordagens, é efetivamente a ordem real, uma vez que ela não é previamente conhecida. Nesse contexto, torna-se importante o uso da simulação computacional, pois ela permite o conhecimento prévio da ordem verdadeira, o que possibilita avaliar os resultados obtidos. Um método

bastante utilizado para a análise de dados simulados é o método Monte Carlo, que consiste na simulação aleatória de várias populações e subsequente análise com a metodologia de interesse. Hackett & Broadfoot (2003) e Wu et al. (2003) utilizaram esse método para verificar a eficiência de diversos algoritmos e critérios de ordenação. Contudo, os mapas simulados nesses trabalhos continham apenas cinco e dez marcadores por grupo de ligação, respectivamente, e não foi feito um estudo detalhado de populações do tipo F_2 , que são muito comuns no mapeamento genético. Ainda, Wu et al. (2003) não consideraram a presença de dados perdidos, que geralmente ocorrem em situações reais.

Este trabalho teve por objetivo avaliar a eficiência dos algoritmos “rapid chain delineation” e “seriation” e, também, dos critérios soma mínima das frações de recombinação adjacentes, produto mínimo das frações de recombinação adjacentes e soma máxima dos LOD Scores adjacentes, associados ao uso do “ripple”, na construção de mapas genéticos.

Material e Métodos

Foi simulado um mapa genético de uma espécie vegetal hipotética, diplóide e monóica, com um cromossomo com 24 marcadores, pelo método proposto por Basten et al. (2005). Os marcadores foram posicionados a distâncias que variavam aleatoriamente, com média de 10 cM entre marcadores (desvio-padrão de 6 cM). Para que a simulação representasse situações encontradas na prática, esses parâmetros foram escolhidos após consulta a mapas de diversas culturas de importância econômica, disponíveis na literatura como: o milho (Castiglioni et al., 1999; Davis et al., 1999; Krakowsky et al., 2004), arroz (Price et al., 2000; Dong et al., 2003; Gu et al., 2004), soja (Chung et al., 2003) e trigo (Huang & Röder, 2004).

Com base no mapa de ligação obtido, foram simulados dois tipos de população: F_2 e retrocruzamento (RC). A simulação dos gametas para a formação das gerações F_2 e RC foi baseada no fato de os indivíduos da geração F_1 possuírem todos os seus locos em heterozigose. Primeiramente, foi feita a simulação da recombinação, tendo-se assumido o número de permutas como variável aleatória com distribuição de Poisson, com parâmetro igual ao comprimento do cromossomo em Morgans (Basten et al., 2005). As permutas foram posicionadas em seus respectivos cromossomos, de acordo com uma distribuição uniforme. Com esse procedimento, foram

gerados dois gametas provenientes dos pares de cromossomos homólogos após a recombinação, e um deles foi tomado ao acaso para simulação do indivíduo da geração seguinte. Para a obtenção de um indivíduo da população RC, o processo de simulação do gameta foi realizado apenas uma vez; no caso de um indivíduo da população F₂, o processo foi realizado duas vezes. Dessa forma, foram simuladas populações experimentais dos tipos RC e F₂ com 200 indivíduos cada.

Foram simulados outros arquivos de dados, tendo-se retirado de modo aleatório 25, 50 e 75% dos dados previamente simulados. Dessa forma, foram simuladas três situações com os dados perdidos e uma situação com os dados completos, para os dois tipos de populações experimentais. Nas populações do tipo F₂, foram simulados três tipos de combinação de marcadores: 100% de marcadores co-dominantes (CC); mistura com 50% de marcadores co-dominantes e 50% de marcadores dominantes (CD); e 100% de marcadores dominantes (DD). A escolha do genitor dominante foi feita ao acaso. Isso implica que os marcadores dominantes apareceram ligados tanto em repulsão, quanto em associação. Foram, então, gerados 16 tipos de populações experimentais no total.

Para a simulação das amostras Monte Carlo, o procedimento descrito anteriormente foi repetido 1.000 vezes para o mapa simulado. Como cada simulação envolve a realização de amostras com base em distribuições de probabilidade, espera-se, teoricamente, que cada amostra seja diferente das anteriores. A avaliação dos métodos de análise pode então ser feita pela aplicação dos métodos a cada amostra, o que tecnicamente equivale a repetições (Manly, 1997).

O método RCD (Doerge, 1996) consiste numa maneira simples para a ordenação de marcadores moleculares dentro dos grupos de ligação. Para que esse método fosse aplicado, foram calculadas, inicialmente, as estimativas de máxima verossimilhança das frações de recombinação, entre todos os marcadores. Isso foi feito por meio do teste de dois pontos (Liu, 1998). Para a ordenação do grupo de ligação simulado, verificou-se qual foi o par de marcadores que possuía a menor fração de recombinação. Verificou-se, então, o marcador ainda não mapeado que apresentou a menor fração de recombinação com os dois marcadores terminais, ao posicioná-los ao lado desse marcador. Esse processo foi repetido até que todos os marcadores fossem adicionados à cadeia (Doerge, 1996).

A ordenação dos marcadores, por meio do algoritmo serialização (Buetow & Chakravarti, 1987), também foi baseada nas frações de recombinação, calculadas ao se considerar todos os marcadores dois a dois. A construção do mapa iniciou-se com cada um dos m marcadores do grupo de ligação simulado. Com M_i como marcador inicial, posicionou-se M_j à direita de M_i, se a fração de recombinação fosse a menor das frações de recombinação entre M_i e todos os outros m-1 marcadores restantes. Após esse procedimento, escolheu-se o marcador M_k, entre aqueles ainda não posicionados, que apresentasse a menor fração de recombinação com M_i. Então, comparou-se a fração de recombinação de M_k com os dois marcadores externos no grupo de marcadores posicionados: M_{esq} (marcador do lado esquerdo) e M_{dir} (marcador do lado direito). Se a fração de recombinação entre os marcadores M_k e M_{dir} fosse maior que a fração de recombinação entre M_k e M_{esq}, posicionava-se M_{esq} à esquerda do grupo, caso contrário, posicionava-se M_{esq} à direita. Este procedimento foi repetido até que todos os marcadores estivessem posicionados (Liu, 1998).

Para m marcadores, m mapas foram obtidos. Se a matriz de frações de recombinação fosse monótona, ou seja, se os valores das estimativas aumentassem à medida que se afastavam da diagonal principal, uma única ordem era obtida. Se a matriz não era monótona, calculava-se um índice de continuidade (CI) para medir a adequação à monotonicidade. Para cada uma das m ordens, o CI da matriz de fração de recombinação foi estimado por meio da fórmula:

$$CI = \sum_{i < j} \frac{\hat{r}_{ij}}{(j-i)^2},$$

em que: i e j são indicadores das posições no genoma; e \hat{r}_{ij} é a estimativa da fração de recombinação, entre os marcadores localizados nas duas posições do genoma. A melhor ordem foi considerada como aquela que implicou numa matriz com menor CI (Liu, 1998).

Pelo fato de os algoritmos de ordenação utilizarem certo grau de aleatoriedade durante a construção do mapa, erros cometidos podem ser acumulados durante as etapas seguintes. O algoritmo “ripple” foi aplicado com o objetivo de fazer verificações de pequenos erros de ordenação, pelo uso de critérios calculados para subordens alternativas, via buscas locais exaustivas. Dado um mapa de ligação previamente construído, o algoritmo “ripple” permutou m´ marcadores vizinhos e

comparou os $m'/2$ mapas resultantes, por meio de algum critério. No caso do presente trabalho, foi usado $m' = 5$. Primeiramente, as posições 1,..., m' são permutadas, depois as posições 2,..., $m'+1$, e assim por diante até que o mapa todo tenha sido percorrido e essas ordens verificadas. Se algum dos mapas resultantes apresentasse um valor melhor para o critério adotado, ele era considerado correto e o mapa anterior descartado.

No presente estudo, foram avaliados os critérios SARF, PARF e SALOD. O índice SARF fornece a soma das frações de recombinação adjacentes de uma dada ordem. Com este critério, as ordens que apresentaram menores valores de SARF entre os marcadores adjacentes foram consideradas as mais prováveis de serem as corretas. O SARF para uma determinada ordem 1,2,3,...,m-1, m, é obtido pela fórmula (Liu, 1998):

$$\text{SARF} = \sum_{i=1}^{m-1} \hat{r}_{M_i M_{i+1}},$$

em que M_i e M_{i+1} são os marcadores nas posições i e $i+1$ e $\hat{r}_{M_i M_{i+1}}$ é a fração de recombinação entre esses marcadores.

O critério PARF baseia-se no produto das frações de recombinação entre marcadores adjacentes. Com este critério, as ordens que apresentaram menores valores de PARF, entre os marcadores adjacentes, foram consideradas as mais prováveis de serem as corretas. O PARF, para uma determinada ordem 1,2,3,...,m-1, m, é obtido pela fórmula (Liu, 1998):

$$\text{PARF} = \prod_{i=1}^{m-1} \hat{r}_{M_i M_{i+1}}$$

O índice SALOD baseia-se na obtenção da máxima soma dos LOD Scores (Morton, 1955) entre marcadores adjacentes. Com este critério, as ordens que apresentaram maiores valores de SALOD, entre os marcadores adjacentes, foram consideradas as mais prováveis de serem as corretas. O SALOD para uma determinada ordem é:

$$\text{SALOD} = \sum_{i=1}^{m-1} \text{lod}_{M_i M_{i+1}},$$

em que $\text{lod}_{M_i M_{i+1}}$ é o LOD Score entre os marcadores M_i e M_{i+1} .

Para a avaliação da eficiência dos algoritmos e dos métodos de ordenação, foram utilizados a correlação ordinal (Spearman, 1904) e o número de ordens corretas, entre as 1.000 amostras Monte Carlo. Como, todavia,

marcadores que se encontram mais próximos têm menor probabilidade de serem ordenados corretamente, do que aqueles que se encontram mais distantes, a ordem de marcadores próximos teve menor relevância no momento da avaliação de um algoritmo ou método. Portanto, considerou-se que marcadores que distavam menos que 5 cM ocupavam a mesma posição, ou seja, possuíam o mesmo "rank". O coeficiente de correlação ordinal foi obtido pela fórmula:

$$\rho = 1 - \frac{6 \sum (i-i^*)^2}{m(m^2-1)}, i=1, \dots, m,$$

em que ρ é o coeficiente de correlação entre a ordem do mapa estimado e a ordem do mapa real; i é a posição ("rank") do marcador M_i no mapa construído; e i^* é a posição ("rank") do marcador M_{i^*} no mapa simulado. Mapas que apresentaram ordens idênticas ao mapa real tiveram $|\rho| = 1$. Todas as análises foram feitas no ambiente R, que é uma linguagem para análise de dados e construção de gráficos (R Development Core Team, 2008), e em programas específicos escritos em linguagem C (Kernighan & Ritchie, 1989).

Resultados e Discussão

Com base nas correlações entre os mapas real e estimado (Tabela 1), pôde-se notar que, para ambos os algoritmos de ordenação, sem a utilização do "ripple" e na ausência de dados perdidos, as médias das correlações foram muito próximas de 1, com variações de 0,99 a 0,94 para SER, e de 1 a 0,97 para RCD. Em ambas as situações, os desvios-padrão das correlações foram pequenos e próximos de zero, o que indica um bom desempenho dos dois algoritmos, nesse caso. Para essa mesma situação, o algoritmo RCD apresentou uma grande quantidade de ordenações corretas (Tabela 2), tendo chegado a 989 mapas corretos a partir das populações F_2 , apenas com marcadores co-dominantes (CC), e a 978 mapas corretos a partir das populações RC. O algoritmo SER teve um resultado menos expressivo, com valores de 105 mapas corretos para as populações F_2 , apenas com marcadores co-dominantes, e 144 para as populações RC.

Na presença de marcadores dominantes (casos CD e DD) na população F_2 , sem dados perdidos, o número de ordens corretas para o algoritmo RCD foi de 369 e 30, respectivamente para CD e DD. Para o algoritmo SER, esses números foram 9 e 1. Todavia, é importante notar que, nesses casos, mesmo que muitas ordens

obtidas não tenham sido idênticas à ordem real simulada, as correlações foram bastante altas e variaram de 0,99 a 0,97 para o RCD, e de 0,97 a 0,94 para o SER, o que indica boa eficiência dos algoritmos. Para o número de ordens corretas, uma única alteração na ordem dos marcadores foi computada como uma ordem incorreta, mesmo que esses marcadores estivessem muito próximos; no entanto, sabe-se que mesmo bons mapas podem apresentar essas inversões, pois os intervalos de confiança das posições desses marcadores normalmente se sobrepõem e, a depender da amostra, podem ocorrer pequenas variações na ordem (Margarido et al., 2005).

O mapa real apresentou comprimento de 240,1 cM. Na ausência de dados perdidos, a média dos comprimentos dos mapas obtidos com o RCD na população F_2 , apenas com marcadores co-dominantes,

foi de 231,4 cM; com mistura de marcadores (CD), o valor foi de 225,5 cM; apenas com uso de marcadores dominantes (DD), foi de 206,1 cM; e, para população RC, o valor foi de 230,9 cM. Com o uso do algoritmo SER, esses valores foram de 261,5 cM para F_2 , apenas com marcadores co-dominantes; 279,9 cM para mistura de marcadores (CD); 281,5 cM para marcadores dominantes apenas (DD), e 256 cM para RC.

Verificou-se que, nos casos em que marcadores dominantes estavam presentes, foram obtidas as piores estimativas, mesmo na ausência de dados perdidos, uma vez que o comprimento médio estimado se afastou do comprimento real, e os desvios-padrão foram sempre maiores. Esses resultados indicam que as estimativas com base em combinações de marcadores dois a dois (estimativas de dois pontos), usadas no presente trabalho,

Tabela 1. Comprimento médio e média das correlações de “rank” obtidas pelos algoritmos delimitação rápida em cadeia (RCD) e seriação (SER), sem o uso do algoritmo “ripple”, nas amostras Monte Carlo simuladas para 0, 25, 50 e 75% de dados perdidos, nos delineamentos genéticos F_2 e retrocruzamento (RC)⁽¹⁾.

Situações avaliadas	0%				25%				50%				75%			
	F_2			RC												
	CC	CD	DD	CC												
Algoritmo RCD																
Comprimento médio dos mapas (cM)	231,4	225,5	206,1	230,9	233,5	227,9	194,1	234,1	242,3	222,1	137,3	249,4	213,1	140,9	63,0	199,9
Desvio-padrão do comprimento (cM)	7,9	24,6	38,5	3,7	12,3	30,2	42,1	5,2	33,7	47,0	51,0	31,0	64,6	53,5	39,5	60,2
Média das correlações de “rank”	1,00	0,99	0,97	1,00	1,00	0,96	0,87	1,00	0,98	0,76	0,62	0,96	0,64	0,44	0,33	0,60
Desvio-padrão das correlações	0,00	0,02	0,06	0,00	0,00	0,09	0,18	0,01	0,07	0,24	0,26	0,09	0,26	0,25	0,21	0,26
Algoritmo SER																
Comprimento médio dos mapas (cM)	261,1	279,9	281,5	256,0	302,7	314,3	297,9	302,7	346,6	338,1	272,2	351,5	264,2	219,9	166,4	336,5
Desvio-padrão do comprimento (cM)	21,3	32,1	44,1	10,6	29,8	44,0	59,1	12,1	51,5	72,1	92,6	48,2	166,2	141,2	107,5	120,2
Média das correlações de “rank”	0,99	0,97	0,94	0,99	0,96	0,91	0,83	0,97	0,88	0,76	0,63	0,86	0,57	0,45	0,37	0,54
Desvio-padrão das correlações	0,03	0,11	0,17	0,06	0,13	0,21	0,27	0,11	0,17	0,30	0,30	0,13	0,29	0,25	0,23	0,28

⁽¹⁾CC: 100% de marcadores co-dominantes; CD: mistura com 50% de marcadores co-dominantes e 50% de marcadores dominantes; e DD: 100% de marcadores dominantes.

Tabela 2. Número de ordens corretas obtidas pelos algoritmos delimitação rápida em cadeia (RCD) e seriação (SER), sem o uso do algoritmo “ripple”, nas amostras Monte Carlo simuladas para 0, 25, 50 e 75% de dados perdidos, nos delineamentos genéticos F_2 e retrocruzamento (RC)⁽¹⁾.

Situações avaliadas	0%				25%				50%				75%			
	F_2			RC	F_2			RC	F_2			RC	F_2			RC
	CC	CD	DD	CC	CC	CD	DD	CC	CC	CD	DD	CC	CC	CD	DD	CC
Algoritmo RCD																
RCD sem RIPPLE	989	369	30	978	808	98	1	655	264	3	0	103	0	0	0	0
RCD+RIPPLE+SARF	999	513	96	999	957	211	6	876	522	9	0	219	2	0	0	0
RCD+RIPPLE+PARF	999	479	77	998	961	175	3	880	548	3	0	247	0	0	0	0
RCD+RIPPLE+SALOD	999	216	10	999	952	131	1	874	480	10	0	190	1	0	0	0
Algoritmo SER																
SER sem RIPPLE	105	9	1	144	2	0	0	0	0	0	0	0	0	0	0	0
SER+RIPPLE+SARF	877	427	84	896	722	149	3	640	305	9	0	119	0	0	0	0
SER+RIPPLE+PARF	876	374	58	902	691	129	3	616	292	5	0	105	0	0	0	0
SER+RIPPLE+SALOD	876	186	3	897	681	94	1	621	241	6	0	98	0	0	0	0

⁽¹⁾CC: 100% de marcadores co-dominantes; CD: mistura com 50% de marcadores co-dominantes e 50% de marcadores dominantes; e DD: 100% de marcadores dominantes.

apresentam bons resultados nos casos em que existem apenas marcadores co-dominantes. À medida que o número de marcadores dominantes aumenta, essas estimativas tendem a fornecer valores com menor acurácia. Outra forma de abordar este problema seria por meio de estimativas multipontos (Lander & Green, 1987), que devem contornar a propagação de erros, ocasionados pela falta de informação entre algumas combinações de marcadores.

Pôde-se observar que, à medida que o número de dados perdidos aumentava, as médias das correlações diminuíam e seus desvios-padrão aumentavam. Para as populações F_2 , verificou-se que as que apresentavam mais dados perdidos foram as mais afetadas. Com 25% de dados perdidos, todas as situações ainda apresentavam correlações bastante próximas de 1, exceto nos casos em que existiam apenas marcadores dominantes nas populações F_2 , que apresentaram correlação igual a 0,87 para o RCD e 0,83 para o SER. Para essa quantidade de dados perdidos, os comprimentos dos mapas tiveram média de 222,4 cM, com o RCD, e 304,4 cM, com o SER. O número de ordens corretas foi 808, para F_2 com marcadores CC, e 665, para RC; ao passo que, para F_2 com marcadores CD e DD, esses valores foram de 98 e 1, respectivamente. Para o algoritmo SER foram obtidas apenas duas ordens corretas.

Com 50% dos dados perdidos, as duas situações que apresentavam apenas marcadores co-dominantes (F_2 com marcadores CC, e RC) tiveram melhores resultados que aquelas que apresentavam marcadores dominantes (F_2 com marcadores CD e DD). No primeiro caso (CC e RC), as médias das correlações foram 0,98 e 0,96, para o RCD, e 0,88 e 0,86, para o SER. No segundo caso (CD e DD), as médias das correlações foram de 0,76 (CD) e 0,62 (DD), para o RCD, e 0,73 (CD) e 0,63 (DD), para o SER. O número de ordens corretas foi de 264, para F_2 com marcadores CC, e 103, para RC; ao passo que, para F_2 com marcadores CD e DD, esses valores foram de 3 e 0, respectivamente. Para o algoritmo SER não foram obtidas ordens corretas.

Com 75% dos dados perdidos, nenhum dos algoritmos mostrou-se satisfatório, uma vez que as correlações obtidas foram baixas, com médias de 0,48 para o SER, e 0,50 para o RCD.

Pode-se notar a importância dos marcadores co-dominantes na obtenção de boas estimativas da ordem dos marcadores, nos dois algoritmos estudados, mesmo na presença de dados perdidos. Com a utilização desse

tipo de marcador, é possível observar ambos os alelos de um determinado loco, ao mesmo tempo, o que permite distinguir um indivíduo homocigoto de um heterocigoto. Isso evita misturas de classes genotípicas e leva a melhores estimativas de fração de recombinação e, conseqüentemente, a melhores estimativas dos mapas. Com estimativas multiponto, Jiang & Zeng (1997) mostraram que marcadores dominantes podem fornecer bons mapas, desde que associados a marcadores co-dominantes. A abordagem multiponto considera a informação de todos os marcadores ao mesmo tempo, para a obtenção das estimativas de fração de recombinação. Com isso, a falta de informação dos marcadores dominantes pode ser compensada pela informação dos marcadores co-dominantes, em função de sua proximidade. Parece claro, portanto, que estimativas multiponto devem ser consideradas em trabalhos futuros.

As correlações obtidas com os critérios PARF, SARF, e SALOD, quando aplicados juntamente com o algoritmo “ripple”, nos mapas pré-estabelecidos pelos algoritmos RCD e SER, foram bastante similares àquelas obtidas sem o uso do “ripple”, em todas as situações. Nesses casos, nos mapas pré-estabelecidos com o algoritmo RCD, as correlações variaram de 1 a 0,33 para o SARF, de 1 a 0,34 para o PARF, e de 1 a 0,32 para o SALOD. Nos mapas obtidos com algoritmo SER, esses valores foram de 0,99 a 0,36 para os três critérios. Entretanto, deve-se ressaltar que o “ripple” tem como objetivo a correção de erros locais, e é efetivo em pequenas regiões, uma vez que o número de marcadores usado é limitado, em razão do “problema do caixeiro viajante”. Nos casos em que as correlações obtidas sem o uso do “ripple” foram altas, esse algoritmo teve papel importante no aumento do número de ordens corretas, embora as correlações não tenham sofrido grandes alterações. Mapas cujos erros de ordenação são pequenos, geralmente, possuem correlações altas e, nesses casos, o algoritmo “ripple” é efetivo para a correção desses pequenos erros locais.

O número médio de ordens corretas, em todos os casos e sem o uso do “ripple”, com o RCD, foi de 268,62. Com o uso do “ripple” e do critério SARF, esse número foi de 338,06; com o PARF foi de 335,6; e com o SALOD foi de 303,94. Para o SER, esses valores foram 16,31 ordens corretas, sem o “ripple”; com o uso do “ripple” e do SARF, esse número foi de 264,47, com o PARF foi de 253,19 e com o SALOD foi de 231,50, o que evidencia que o “ripple” trouxe ganhos significativos,

quando usado nos mapas pré-estabelecidos pelos dois algoritmos.

No presente trabalho, foram utilizados $m' = 5$ marcadores para a execução do “ripple”, o que resultou em 60 ordens possíveis ($5!/2$), para cada posição de início. Uma alternativa para a melhoria das ordenações pode ser a de aumentar o número de marcadores utilizados nesse algoritmo, desde que o tempo de processamento seja computacionalmente viável. Por exemplo, $m' = 10$ acarretaria em 1.814.400 ordens possíveis ($10!/2$). Como neste trabalho foram utilizadas 1.000 amostras Monte Carlo, esse número de ordens aumentaria sensivelmente o tempo necessário para se realizarem as análises e, por isso não foi considerado. A única exceção na melhoria das ordens ocorreu com o critério SALOD, para a população F_2 sem dados perdidos e com mistura de marcadores (CD), a qual passou de 369 ordens corretas sem o uso de “ripple”, para 216 com o uso do “ripple”. Como apontado por Olson & Boehnke (1990), o critério SALOD é sensível a diferenças na quantidade de informação dos marcadores e, portanto, não é um critério aconselhável, quando há mistura de marcadores dominantes e co-dominantes. Os resultados aqui apresentados corroboram os de Olson & Boehnke (1990).

Para populações RC e F_2 com tamanho amostral $n = 200$, caso haja até 25% de dados perdidos, ambos os algoritmos podem ser empregados satisfatoriamente, dada a alta correlação dos mapas estimados com o mapa real, mesmo na presença de marcadores dominantes. No caso de 50% ou mais dados perdidos, ambos os algoritmos podem ser recomendados indistintamente, caso só haja marcadores co-dominantes. Caso marcadores dominantes sejam incluídos nessa situação, estudos com métodos que considerem estimativas multiponto devem ser investigados.

Conclusões

1. Os algoritmos delimitação rápida em cadeia e seriação apresentam resultados semelhantes, quando se consideram as correlações entre o mapa real e o estimado; ambos são sensíveis à presença de dados perdidos.

2. A presença de marcadores dominantes dificulta a obtenção de estimativas com boa acurácia, tanto da ordem como da distância, nos dois algoritmos estudados.

3. O algoritmo “ripple”, quando aplicado concomitantemente aos critérios soma mínima das

frações de recombinação adjacentes, produto mínimo das frações de recombinação adjacentes e soma máxima dos LOD Scores adjacentes, aumenta o número de ordens corretas pré-estabelecidas pelos algoritmos delimitação rápida em cadeia e seriação.

4. O critério soma máxima dos LOD Scores adjacentes não é recomendado, quando há misturas de marcadores dominantes e co-dominantes.

Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico, pela concessão de bolsas.

Referências

- BASTEN, C.J.; WEIR, B.S.; ZENG, Z.B. **QTL cartographer**: version 1.17: a reference manual and tutorial for QTL mapping. Raleigh: North Carolina State University, 2005. 190p.
- BOTSTEIN, D.; WHITE, R.L.; SKOLNICK, M.; DAVIS, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. **American Journal of Human Genetics**, v.32, p.314-331, 1980.
- BUETOW, K.H.; CHAKRAVARTI, A. Multipoint gene mapping using seriation. I. General methods. **American Journal of Human Genetics**, v.41, p.180-188, 1987.
- CASTIGLIONI, P.; AJMONE-MARSAN, P.; WIJK, R. van; MOTTO, M. AFLP markers in a molecular linkage map of maize: codominant scoring and linkage group distribution. **Theoretical and Applied Genetics**, v.99, p.425-431, 1999.
- CATO, S.A.; GARDNER, R.C.; KENT, J.; RICHARDSON, T.E. A rapid PCR-based method for genetically mapping ESTs. **Theoretical and Applied Genetics**, v.102, p.296-306, 2001.
- CHUNG, J.; BABKA, H.L.; GRAEF, G.L.; STASWICK, P.E.; LEE, D.J.; CREGAN, P.B.; SHOEMAKER, R.C.; SPECHT, J.E. The seed protein, oil, and yield QTL on soybean linkage group I. **Crop Science**, v.43, p.1053-1067, 2003.
- DAVIS, G.L.; McMULLEN, M.D.; BAYSDORFER, C.; MUSKET, T.; GRANT, D.; STAEBELL, M.; XU, G.; POLACCO, M.; KOSTER, L.; MELIA-HANCOCK, S.; HOCHINS, K.; CHAO, S.; COE JÚNIOR, E.H. A maize map standard with sequenced core markers, grass genome reference points and 932 expressed sequence tagged sites (ESTs) in a 1736-locus map. **Genetics**, v.152, p.1137-1172, 1999.
- DOERGE, R.W. Constructing genetic maps by rapid chain delimitation. **Journal of Quantitative Trait Loci**, v.2, p.121-132, 1996.
- DONG, Y.; TSUZUKI, E.; KAMIUNTEN, H.; TERAOKA, H.; LIN, D. Mapping of QTL for embryo size in rice. **Crop Science**, v.43, p.1068-1071, 2003.
- FALK, C.T. A simple scheme for preliminary ordering of multiple loci: application to 45 CF families. In: ELSTON, R.C.; SPENCE,

- M.A.; RODGE, S.E.; MCCLUE, J.W. (Ed.). **Multipoint mapping and linkage based upon affected pedigree members**. New York: Liss, 1989. p.17-22.
- GU, X.; KIANIAN, S.F.; FOLEY, M.E. Multiple loci and epistases control genetic variation for seed dormancy in weedy rice (*Oryza sativa*). **Genetics**, v.166, p.1503-1516, 2004.
- HACKETT, C.A.; BROADFOOT, L.B. Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. **Heredity**, v.90, p.33-38, 2003.
- HUANG, X.; RÖDER, M.S. Molecular mapping of powdery mildew resistance genes in wheat: a review. **Euphytica**, v.137, p.203-223, 2004.
- JACCOUD, D.; PENG, P.; FEINSTEIN, D.; KILIAN, A. Diversity arrays: a solid state technology for sequence information independent genotyping. **Nucleic Acids Research**, v.29, p.e25, 2001.
- JIANG, C.; ZENG, Z.B. Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. **Genetica**, v.101, p.47-58, 1997.
- KERNIGHAN, B.W.; RITCHIE, D.M. **C: a linguagem de programação: padrão ANSI**. 15.ed. São Paulo: Campus, 1989. 289p.
- KRAKOWSKY, M.D.; LEE, M.; WOODMAN-CLIKEMAN, W.L.; LONG, M.J.; SHAROPOVA, N. QTL mapping of resistance to stalk tunneling by the European corn borer in RILs of maize population B73 X De811. **Crop Science**, v.44, p.274-282, 2004.
- LANDER, E.S.; GREEN, P. Construction of multilocus genetic linkage maps in humans. **Proceedings of the National Academy of Sciences of the United States**, v.84, p.2363-2367, 1987.
- LANDER, E.S.; GREEN, P.; ABRAHAMSON, J.; BARLOW, A.; DALY, M.J.; LINCOLN, S.E.; NEWBURG, L. MAPMAKER: An interactive computing package for constructing primary genetic linkage map of experimental and natural populations. **Genomics**, v.1, p.174-181, 1987.
- LIU, B.H. **Statistical genomics: linkage, mapping, and QTL analysis**. 2nd ed. Boca Raton: CRC Press, 1998. 648p.
- MANLY, B.F.J. **Randomization, bootstrap and Monte Carlo methods in biology**. 2nd ed. London: Chapman & Hall, 1997. 399p.
- MARGARIDO, G.R.A.; MOLLINARI, M.; GARCIA, A.A.F. Uso do método de reamostragem bootstrap para validação da ordem de marcadores em mapas genéticos. In: CONGRESSO BRASILEIRO DE GENÉTICA, 51., 2005, Águas de Lindóia. **Anais**. Águas de Lindóia: Sociedade Brasileira de Genética, 2005. CD-ROM.
- MESTER, D.; RONIN, Y.; MINKOV, D.; NEVO, E.; KOROL, A. Constructing large-scale genetic maps using an evolutionary strategy algorithm. **Genetics**, v.165, p.2269-2282, 2003.
- MORTON, N.E. Sequential tests for the detection of linkage. **American Journal of Human Genetics**, v.7, p.277-318, 1955.
- OLSON, J.M.; BOEHNKE, M. Monte Carlo comparison of preliminary methods for ordering multiple genetic loci. **American Journal of Human Genetics**, v.47, p.470-482, 1990.
- PRICE, A.H.; STEELE, K.A.; MOORE, B.J.; BARRACLOUGH, P.B.; CLARK, L.J. A combined RFLP and AFLP linkage map of upland rice (*Oryza sativa* L.) used to identify QTL for root-penetration ability. **Theoretical and Applied Genetics**, v.100, p.49-56, 2000.
- R DEVELOPMENT CORE TEAM. **The R project for statistical computing**. Disponível em: <http://www.R-project.org>. Acesso em: 15 abr. 2008.
- SPEARMAN, C. The proof and measurement of association between two things. **American Journal of Psychology**, v.15, p.72-101, 1904.
- SYVÄNEN, A.C. Accessing genetic variation: genotyping single nucleotide polymorphisms. **Nature Reviews Genetics**, v.2, p.930-942, 2001.
- TAUTZ, D. Hypervariability of simple sequences as a general source for polymorphic DNA markers. **Nucleic Acids Research**, v.17, p.6463-6471, 1989.
- WEEKS, D.; LANGE, K. Preliminary ranking procedures for multilocus ordering. **Genomics**, v.1, p.236-242, 1987.
- WILLIAMS, J.G.; KUBELIK, A.R.; LIVAK, K.J.; RAFALSKI, J.A.; TINGEY, S.V. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. **Nucleic Acids Research**, v.18, p.6531-6535, 1990.
- WILSON, S.R. A major simplification in the preliminary ordering of linked loci. **Genetic Epidemiology**, v.5, p.75-80, 1988.
- WU, J.; JENKINS, J.; ZHU, J.; McCARTY, J.; WATSON, C. Monte Carlo simulations on marker grouping and ordering. **Theoretical and Applied Genetics**, v.107, p.568-573, 2003.
- ZABEAU, M.; VOS, P. **Selective restriction fragment amplification: a general method for DNA fingerprinting**. EP 0534858. 24 Sept. 1993, 27 Apr. 2005.

Recebido em 2 de dezembro de 2007 e aprovado em 16 de abril de 2008