Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”

Comparação de algoritmos usados na construção de mapas genéticos

Marcelo Mollinari

Dissertação apresentada para a obtenção do título de
Mestre em Agronomia. Área de concentração: Genética e
Melhoramento de Plantas

Piracicaba
2007

Marcelo Mollinari
Engenheiro Agrônomo

Comparação de algoritmos usados na construção de mapas genéticos

Orientador:

Prof. Dr. ANTONIO AUGUSTO FRANCO GARCIA

Dissertação apresentada para a obtenção do título de
Mestre em Agronomia. Área de concentração: Genética e
Melhoramento de Plantas

Piracicaba
2007

Dados Internacionais de Catalogação na Publicação (CIP)
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP

Mollinari, Marcelo

Comparação de algoritmos usados na construção de mapas genéticos / Marcelo
Mollinari - - Piracicaba, 2007.
73 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura Luiz de Queiroz, 2007.
Bibliografia.

1. Algoritmos 2. Cadeias de Markov 3. Mapeamento genético 4. Marcador Molecular
I. Título

CDD 575.12

“Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor”

*Aos meus pais Fani e Norberto,
por sempre me apoiarem e acreditarem
nos meus sonhos*

Agradecimentos

Ao professor Antonio Augusto Franco Garcia, pela orientação, ensinamentos, amizade e confiança em mim depositada desde a Iniciação Científica. Qualquer palavra aqui escrita seria pouco para expressar o quanto sou grato por esses 7 anos que convivemos juntos.

Ao professor Roland Vencovsky, pela co-orientação, atenção e pelas conversas repletas de ótimas idéias.

À minha irmã Flávia, ao meu cunhado Luis e à minha amada sobrinha Ana Clara, que chegou para encher de luz e amor minha vida.

Ao Gabriel e à Marta, pela leitura dos originais e excelentes críticas e sugestões.

Ao Carlos Macedônio, pelas idéias e ajuda com a parte computacional do trabalho.

Ao Departamento de Genética da ESALQ/USP, pela oportunidade.

Ao CNPq, pela bolsa de estudos concedida.

Ao Centro Nacional de Processamento de Alto Desempenho da UNICAMP, pela concessão de uso dos seus computadores.

Aos “irmãos” de orientação Gabriel, Marta, Millor, Priscilla e Rodrigo, por sempre darem ótimas idéias e sugestões, contribuindo muito, não só para o desenvolvimento deste trabalho, mas também para meu desenvolvimento científico e pessoal.

À Ana Schilling e à Carolina Mathias Moreira, pela valiosa ajuda com o $\text{\LaTeX}$, afinal, como diria a Ana, essa dissertação precisava de margens.

Aos amigos e colegas do curso, pela convivência e aprendizado.

À Clarissa, que muito me apoiou nessa reta final com paciência e compreensão nos momentos difíceis.

Aos meus amigos Alexandre Garcia e Fernando Piotto, pelas noites em claro acompanhadas de copos de café, livros e artigos.

A todos amigos, mesmo os não diretamente envolvidos com o trabalho, mas que sempre estiveram presentes, especialmente à Ana Clara Mathias, Carol Mathias, Carol Nalon, Cris Millan, Luciana Jacob, Paulo César.

SUMÁRIO

RESUMO	6
ABSTRACT	7
1 INTRODUÇÃO	8
2 REVISÃO DE LITERATURA	12
2.1 Mapas Genéticos	12
2.1.1 Marcadores Moleculares	12
2.1.2 Algumas aplicações dos mapas genéticos	15
2.1.3 Construção de Mapas Genéticos	16
2.2 Cadeias de Markov	17
2.2.1 A Cadeia de Markov Observável	17
2.2.2 A Cadeia de Markov Oculta	18
2.3 Critérios para Avaliação de Ordens	21
2.3.1 Verossimilhança Via Cadeias de Markov Ocultas (LHMC)	22
2.3.2 SARF, PARF, SALOD e outros	26
2.4 Algoritmos de Ordenação	28
2.4.1 <i>TRY</i>	28
2.4.2 <i>Seriation (SER)</i>	29
2.4.3 <i>Rapid Chain Delineation - (RCD)</i>	31
2.4.4 <i>Recombination Counting and Ordering - (RECORD)</i>	31
2.4.5 <i>Unidirectional Growth - (UG)</i>	33
2.4.6 Outros Algoritmos de Ordenação	35
2.5 Verificação de erros na ordenação	36
3 MATERIAL E MÉTODOS	37
3.1 Material	37
3.1.1 Simulação dos Genótipos da População F_2	37
3.1.2 Simulação dos Marcadores	39
3.1.3 Simulação dos Dados Perdidos	40
3.1.4 Simulação das Amostras Monte Carlo	42
3.2 Métodos	42
3.2.1 Obtenção das matrizes R e LOD	42

3.2.2 Algoritmos de Ordenação	43
3.2.2.1 Algoritmo TRY	44
3.2.2.2 Algoritmos SER, RCD, RECORD e UG	44
3.2.2.3 Funções de Mapeamento	45
3.2.3 Algoritmo RIPPLE e Critérios para a Avaliação de Ordens	45
3.2.4 Avaliação dos Algoritmos e Critérios	46
3.2.4.1 Coeficiente de Correlação de Spearman	46
3.2.4.2 Distância Euclidiana	48
3.2.5 Recursos Computacionais Utilizados	48
4 RESULTADOS E DISCUSSÃO	50
4.1 Estimativas da Ordem dos Marcadores	50
4.2 Estimativas das Distâncias entre os Marcadores	57
4.3 Considerações Gerais	61
5 CONCLUSÕES	65
REFERÊNCIAS	66

RESUMO

Comparação de algoritmos usados na construção de mapas genéticos

Mapas genéticos são arranjos lineares que indicam a ordem e distância entre locos nos cromossomos de uma determinada espécie. Recentemente, a grande disponibilidade de marcadores moleculares tem tornado estes mapas cada vez mais saturados, sendo necessários métodos eficientes para sua construção. 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. Tal ordenação é considerada um caso especial do clássico problema do caixeiro viajante (TSP), que consiste em escolher a melhor ordem entre todas as possíveis. Entretanto, a estratégia de busca exaustiva torna-se inviável quando o número de marcadores é grande. Nesses casos, para que esses mapas possam ser construídos uma alternativa viável é a utilização de algoritmos que forneçam soluções aproximadas. O objetivo desse trabalho foi avaliar a eficiência dos algoritmos *Try* (TRY), *Seriation* (SER), *Rapid Chain Delineation* (RCD), *Recombination Counting and Ordering* (RECORD) e *Unidirectional Growth* (UG), além dos critérios PARF (produto mínimo das frações de recombinação adjacentes), SARF (soma mínima das frações de recombinação adjacentes), SALOD (soma máxima dos LOD scores adjacentes) e LMHC (verossimilhança via cadeias de Markov ocultas), usados juntamente com o algoritmo de verificação de erros RIPPLE, para a construção de mapas genéticos. Para tanto, foi simulado um mapa de ligação de uma espécie vegetal hipotética, diplóide e monóica, contendo 21 marcadores com distância fixa entre eles de 3 centimorgans. Usando o método Monte Carlo, foram obtidas aleatoriamente 550 populações F_2 com 100 e 400 indivíduos, além de diferentes combinações de marcadores dominantes e codominantes. Foi ainda simulada perda de 10% e 20% dos dados. Os resultados mostraram que os algoritmos TRY e SER tiveram bons resultados em todas as situações simuladas, mesmo com presença de elevado número de dados perdidos e marcadores dominantes ligados em repulsão, podendo ser então recomendado em situações práticas. Os algoritmos RECORD e UG apresentaram bons resultados na ausência de marcadores dominantes ligados em repulsão, podendo então ser recomendados em situações com poucos marcadores dominantes. Dentre todos os algoritmos, o RCD foi o que se mostrou menos eficiente. O critério LMHC, aplicado com o algoritmo RIPPLE, foi o que apresentou melhores resultados quando se deseja fazer verificações de erros na ordenação.

Palavras-chave: Monte Carlo; Estimativas Multiponto; Cadeia de Markov Oculta; Marcador Molecular

ABSTRACT

Comparison of algorithms used in the construction of genetic linkage maps

Genetic linkage maps are linear arrangements showing the order and distance between loci in chromosomes of a particular species. Recently, the availability of molecular markers has made such maps more saturated and efficient methods are needed for their construction. One of the steps that deserves more attention in the construction of genetic linkage maps is the ordering of genetic markers within each linkage group. This ordering is considered a special case of the classic *traveling salesman problem* (TSP), which consists in choosing the best order among all possible ones. However, the strategy of exhaustive search becomes unfeasible when the number of markers is large. One possible alternative to construct such maps is to use algorithms that provide approximate solutions. Thus, the aim of this work was to evaluate the efficiency of algorithms *Try* (TRY), *Seriation* (SER), *Rapid Chain Delineation* (RCD), *Recombination Counting and Ordering* (RECORD) and *Unidirectional Growth* (UG), as well as the criteria PARF (product of adjacent recombination fractions), SARF (sum of adjacent recombination fractions), SALOD (sum of adjacent lod scores) and LMHC (likelihood via hidden Markov chains), used with the RIPPLE algorithm for error verification, in the construction of genetic linkage maps. For doing so, a linkage map of a hypothetical diploid and monoecious plant species was simulated, containing 21 markers with fixed distance of 3 centimorgans between them. Using Monte Carlo methods, 550 F_2 populations were randomly simulated with 100 and 400 individuals, together with different combinations of dominant and codominant markers. 10 % and 20 % of missing data was also included. Results showed that the algorithms TRY and SER gave good results in all situations, even with presence of a large number of missing data and dominant markers linked in repulsion phase. Thus, these can be recommended for analyzing real data. The algorithms RECORD and UG gave good results in the absence of dominant markers linked in repulsion phase and can be used in this case. Among all algorithms, RCD was the least efficient. The criterion LHMC, applied with the RIPPLE algorithm, showed the best results when the goal is to check ordering errors.

Keywords: Monte Carlo; Multipoint estimates; Hidden Markov Chain; Molecular Marker

1 INTRODUÇÃO

Mapas genéticos (ou de ligação) são arranjos lineares que indicam a ordem e distância entre locos (genes, marcadores genéticos, etc) nos cromossomos de uma determinada espécie. Tais mapas são de grande importância em estudos filogenéticos e no melhoramento de plantas. Por exemplo, estudos de sintenia (AHN; TANKSLEY, 1993; CHETELAT; MEGLIC; CISNEROS, 2000; KU et al., 2001) e clonagem de genes (LEE, 1995; MORGANTE; SALAMINI, 2003; SALVI; TUBEROSA, 2005) com base em mapas genéticos representam técnicas importantes que podem ser consideradas em programas de melhoramento das mais variadas espécies vegetais.

Embora o genoma de diversos organismos venham sendo seqüenciados, a construção criteriosa de mapas genéticos é de fundamental importância para o mapeamento de locos que controlam caracteres quantitativos (*QTL - Quantitative Trait Loci*), que possivelmente é a aplicação com maior potencial de uso no melhoramento genético. Isso é importante para o entendimento da arquitetura genética desses caracteres, os quais normalmente têm elevada importância econômica (TAN; FU, 2006). Dessa forma, o sucesso do mapeamento de QTLs depende diretamente da ordenação correta dos marcadores moleculares nesses mapas (WU et al., 2003), bem como nas estimativas das distâncias entre locos.

Os primeiros mapas genéticos foram baseados em poucos marcadores morfológicos. Por exemplo, o mapa do cromossomo X de *Drosophila melanogaster*, publicado por Sturtevant (1913), foi feito utilizando apenas este tipo de marcador. Todavia, o desenvolvimento de mapas genéticos baseados em marcadores morfológicos é demorado e sofre com o fato de os mesmos serem raros no genoma. Recentemente, os marcadores moleculares contornaram tais limitações, já que estes são muito mais freqüentes no genoma. Alguns dos marcadores usados são RFLP (*Restriction Fragment Length Polymorphisms* - BOTSTEIN et al., 1980), RAPD (*Random Amplified Polymorphic DNA* - WILLIAMS et al., 1990), AFLP (*Amplified Fragment Length Polymorphism* - ZABEAU; VOS, 1993) e SSR (*Simple Sequence Repeat* - TAUTZ, 1989). Outros tipos de marcadores, como SNPs (*Single Nucleotide Polymorphism* - SYVÄNEN 2001), DArTs (*Diversity Arrays Technology* - JACCOUD et al., 2001) e EST-SSRs (*SSRs containing expressed sequence tags* - CATO et al., 2000; CARDLE et al., 2001) têm se tornado cada vez mais freqüentes em estudos genéticos, possibilitando uma cobertura mais ampla do genoma.

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 (*TSP - Traveling Salesman Problem*) (DOERGE, 1996; LIU, 1998), que consiste em escolher a melhor ordem entre $m!/2$ ordens possíveis, sendo m o número de marcadores. Uma forma simples de resolver o problema seria através de busca exaustiva, com comparação de todas as ordens possíveis, seguindo algum critério. Entretanto, esta estratégia torna-se inviável quando o número de marcadores é grande. Por exemplo, para $m = 20$ existem aproximadamente $2,43 \times 10^{18}$ ordens possíveis, o que torna o cálculo inviável, mesmo para supercomputadores (DOERGE, 1996; MESTER et al., 2003; TAN e FU, 2006). Em problemas reais, m pode variar de dezenas até centenas (MESTER et al., 2003), principalmente em função da ampla disponibilidade de marcadores moleculares que existe atualmente. Outra abordagem possível para a resolução do TSP é através de aproximações, as quais não consideram as $m!/2$ ordens possíveis. Portanto, o uso de algoritmos para a obtenção de soluções aproximadas é a única abordagem prática existente atualmente para o mapeamento genético em larga escala (LIU, 1998).

Dada a importância de obter ordens corretas, diversos algoritmos usando abordagens aproximadas foram propostos, incluindo *Simulated Annealing* (SA) (THOMPSON 1984; WEEKS; LANGE, 1987), *Stepwise Likelihood* (LATHROP et al., 1984), *Branch and Bound* (BB) (LATHROP et al., 1985), *Try* (TRY) (LANDER et al., 1987), *Seriation* (SER) (BUETOW; CHAKRAVARTI, 1987a), *Rapid Chain Delineation* (RCD) (DOERGE, 1996), *Genetic and Evolutionary Algorithm* (GEA) (MESTER et al., 2003), *Recombination Counting and Ordering* (RECORD) (VAN OS et al., 2005) e *Unidirectional Growth* (UG) (TAN; FU, 2006). Alguns desses algoritmos foram implementados em programas computacionais muito usados pelos geneticistas, tais como LINKAGE (LATHROP et al., 1984), MAPMAKER/EXP (LANDER et al., 1987), JoinMap (STAM, 1993), Emap da suíte de programas QTL Cartographer (BASTEN; WEIR; ZENG, 2003) e OneMap (MARGARIDO; SOUZA; GARCIA, 2007).

Lander et al. (1987) propuseram também o algoritmo RIPPLE para verificação de sub-ordens alternativas de janelas de m' marcadores em um mapa previamente construído com m marcadores ($m' < m$), através de buscas locais exaustivas considerando as $m'!/2$ ordens possíveis. Por exemplo, dada uma ordem com $m = 20$ e uma janela com $m' = 5$, as posições de 1 a 5 são permutadas (todas ordens possíveis) e as demais posições permaneçam inalteradas, sendo então as novas ordens

avaliadas. Se alguma das ordens obtidas for melhor que a ordem anterior, considerando-se algum critério, essa nova ordem é mantida. Em seguida, as posições de 2 a 6 são também permutadas (todas as ordens possíveis) e novas verificações são feitas. Esse procedimento é repetido até que todo o mapa tenha sido percorrido e as ordens alternativas verificadas.

Nota-se então que é necessário escolher bons critérios para a avaliação e comparação destas sub-ordens obtidas com o algoritmo RIPPLE. Muitos critérios de avaliação foram desenvolvidos, dentre os quais tem-se: *i*) soma mínima das frações de recombinação adjacentes (SARF - *sum of adjacent recombination fractions*) (FALK, 1989); *ii*) produto mínimo das frações de recombinação adjacentes (PARF - *product of adjacent recombination fractions*) (WILSON, 1988); *iii*) soma máxima dos LOD scores adjacentes (SALOD - *sum of adjacent lod scores*) (WEEKS; LANGE, 1987); *iv*) SALOD modificado pelo número de meioses informativas (SALEQ - *SALOD-equivalent number of fully informative meioses*) (OLSON; BOEHNKE, 1990); *v*) SALOD modificado pelo conteúdo de informação de polimorfismo (SALPIC - *SALOD-polymorphism information content*) (OLSON; BOEHNKE, 1990); *vi*) verossimilhança via cadeias de Markov ocultas (LHMC - *likelihood via hidden Markov chains*) (LANDER; GREEN, 1987); *vii*) quadrados mínimos (LS - *least squares*) (WEEKS; LANGE, 1987) e *viii*) quadrados mínimos ponderados (WLS - *weighted least squares*) (STAM, 1993).

Em situações práticas, os geneticistas não podem verificar se a ordem estimada para os marcadores do mapa de ligação, para dados experimentais, 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, possibilitando a avaliação dos resultados obtidos. Uma estratégia bastante utilizada em simulações é conhecida como método *Monte Carlo*. A aplicação do método *Monte Carlo* consiste na simulação aleatória de várias populações e subsequente análise dos dados com a metodologia de interesse; cada simulação é considerada uma repetição do experimento (MANLY, 1997). Buetow e Chakravarti (1987b) utilizaram o método *Monte Carlo* para testar o algoritmo SER. Usando esta mesma estratégia, Olson e Boehnke (1990) compararam oito critérios para ordenação preliminar de um número moderado de locos ($m = 6$).

Wu et al. (2003) utilizaram o método *Monte Carlo* para verificar a eficiência dos critérios SARF, PARF, SALOD e LHMC, além do algoritmo SER. Contudo, os parâmetros usados nas simulações não se aproximaram de situações reais, uma vez que os mapas simulados continham apenas cinco

marcadores por grupo de ligação, e não foram considerados dados perdidos, que normalmente ocorrem em situações reais. Outra limitação deste trabalho foi a utilização da população experimental do tipo duplo-haplóide, o que impossibilita a obtenção de resultados conclusivos a respeito das dificuldades encontradas na construção de mapas com a presença de marcadores dominantes, que também são muito usados na prática.

Hackett e Broadfoot (2003) também utilizaram a estratégia *Monte Carlo* para comparar a eficiência dos critérios LHMC, WLS e SARF. Novamente, os parâmetros utilizados pouco aproximaram-se de situações reais, uma vez que foram simulados apenas 10 marcadores por grupo de ligação em uma população duplo-haplóide, onde ocorre a mesma limitação com respeito à ordenação de marcadores dominantes. Além disso, não foram avaliados algoritmos para a obtenção de soluções aproximadas. Claramente, isto representa uma limitação, já que tais algoritmos são as únicas abordagens práticas, até o presente momento, para o mapeamento genético quando muitos locos estão envolvidos (LIU, 1998). Uma vez que os mapas atuais vêm se tornando cada vez mais saturados e o número de locos que devem ser analisados simultaneamente vêm crescendo rapidamente (TAN; FU, 2006), as conclusões que estes trabalhos apresentam não são definitivas para os geneticistas, que no momento de construção de seus mapas precisam escolher a melhor combinação entre método e critério para analisar seus dados.

Nesse contexto, o objetivo do presente trabalho foi avaliar a eficiência dos algoritmos TRY, SER, RCD, RECORD e UG, além dos critérios PARF, SARF, SALOD e LMHC, usados juntamente com o algoritmo RIPPLE, na construção de mapas genéticos. Para tanto, foi utilizado o método *Monte Carlo* com simulações próximas às situações reais, ou seja, envolvendo grupos de ligação com vários locos, marcadores dominantes e codominantes, bem como dados perdidos.

2 REVISÃO DE LITERATURA

2.1 Mapas Genéticos

2.1.1 Marcadores Moleculares

Um marcador genético fornece informações sobre variações alélicas em um determinado loco gênico. Os primeiros marcadores utilizados em estudos genéticos eram morfológicos, ou seja, correspondiam ao produto de expressão de um ou mais genes, podendo ser influenciados pelo ambiente e/ou pela ação de outros genes. Com o advento da biologia molecular, diversos tipos de marcadores genéticos foram desenvolvidos. Tais marcadores moleculares, também chamados de genômicos ou genotípicos, detectam o polimorfismo diretamente ao nível do DNA, não sofrendo qualquer tipo de influência ambiental ou gênica (SOUZA, 2001). Este tipo de marcador vem sendo aplicado em muitos estudos biológicos, como mapeamento genético, reconstrução filogenética, diversidade genética, testes de paternidade e aplicações forenses (SCHLÖTTERER, 2004).

Teoricamente, qualquer fragmento de DNA pode ser usado como marcador molecular, desde que revele polimorfismo entre indivíduos, ou seja, que esteja presente em duas ou mais formas alélicas na população. Os marcadores genéticos podem ser classificados em dois grupos: *codominantes* ou *dominantes*. Os codominantes podem apresentar ambos os alelos de um loco do indivíduo ao mesmo tempo, permitindo distinguir um loco heterozigoto dos homozigotos. Alguns dos marcadores que se enquadram nesta classe são RFLP (*Restriction Fragment Length Polymorphism* - BOTSTEIN et al., 1980), SSR (*Simple Sequence Repeats* - TAUTZ, 1989), EST-SSRs (*Expressed Sequence Tags containing SSRs* - CATO et al., 2000; CARDLE et al., 2001) e SNPs (*Single Nucleotide Polymorphism* - SYVÄNEN 2001). Os marcadores dominantes são detectados pela presença ou ausência dos fragmentos amplificados. Os marcadores dominantes mais comuns são RAPDs (*Random Amplified Polymorphic DNAs* - WILLIAMS et al., 1990), AFLPs (*Amplified Fragment Length Polymorphisms* - ZABEAU; VOS, 1993) e DArT (JACCOUD et al., 2001), que vem ganhando espaço recentemente pelo seu custo relativamente baixo e ampla distribuição no genoma.

Os marcadores codominantes RFLPs permitem que a variação existente no DNA seja acessada devido a diferenças no tamanho dos fragmentos produzidos quando a molécula de DNA é tratada com determinadas enzimas. Estas enzimas clivam a molécula de DNA em pontos específicos chamados *sítios de restrição*. As diferenças nos tamanhos dos fragmentos produzidos ocorrem

devido à variação na distribuição dos sítios de restrição de uma dada enzima no DNA. Para a visualização destes fragmentos, é necessário o uso de sondas clonadas específicas que se hibridizam aos fragmentos, tornando possível a sua visualização (SOUZA, 2001; SCHLÖTTERER, 2004).

A partir da invenção da PCR (*Polymerase Chain Reaction* - SAIKI et al., 1985) ocorreu um grande avanço na obtenção de marcadores moleculares. A PCR permitiu a amplificação e análise de qualquer região genômica em muitos indivíduos sem a necessidade de clonagem ou isolamento de grandes quantidades de DNA genômico (SCHLÖTTERER, 2004). Os primeiros marcadores que fizeram uso da PCR foram os SSRs (*microsatélites*). Os SSRs correspondem a seqüências de DNA com poucos pares de bases (2-6) de comprimento repetidas em tandem, ou seja, uma após a outra. Como o número de repetições em tandem geralmente varia, mas as seqüências de bases adjacentes aos microsatélites podem ser únicas no genoma e conservadas entre diferentes indivíduos, é possível amplificar este loco em diferentes genótipos, mostrando polimorfismo no tamanho do fragmento amplificado (SOUZA, 2001). Os microsatélites são marcadores codominantes altamente polimórficos (geralmente apresentam diversos alelos para um único loco), abundantes e amplamente distribuídos nas regiões eucromáticas do genoma. Devido a essas características, o surgimento dos marcadores SSR teve grande impacto na construção de mapas genéticos (SCHLÖTTERER, 2004).

Os EST-SSRs são obtidos combinando-se a técnica de SSRs com informações de bancos de dados de seqüências expressas (ESTs - *Expressed Sequence Tags*). Para tanto, é feita a busca de SSRs nestes bancos de dados utilizando-se algoritmos específicos. Esta estratégia possui vantagens como a diminuição do tempo e dos custos para a obtenção dos SSRs. Além disso, a presença de SSRs em transcritos de genes conhecidos sugere que eles possam ter importante papel na expressão e função gênica (KATENTY et al., 2002).

Os marcadores dominantes RAPD baseiam-se na realização da PCR, onde seqüências de DNA genômico são amplificadas ao acaso, a partir de iniciadores (*primers*) de seqüência aleatória. Havendo dois sítios complementares a esses iniciadores, localizados em direções opostas, ocorre a amplificação da região. Os AFLPs combinam os alvos para estudo do polimorfismo gerados pelo RFLP e RAPD, ou seja, a distribuição aleatória de sítios de restrição entre genomas e a amplificação aleatória de fragmentos empregando-se oligonucleotídeos com seqüências arbitrárias (SOUZA, 2001). Os marcadores DArT são baseados em hibridizações de microarranjos, sendo portanto independentes de géis. Por esta razão, podem ser feitos em larga escala e de maneira muito mais rápida.

O princípio da técnica é a construção de painéis de diversidade e a hibridização destes com representações genômicas, detectando a presença ou a ausência de fragmentos específicos no genoma (JACCOUD et al., 2001).

Os SNPs são uma classe de marcadores que acessam diferenças em um único nucleotídeo entre seqüências de DNA, os quais vêm sendo cada vez mais utilizados. Um SNP poderia, em princípio, envolver quatro variações diferentes de nucleotídeos, mas na prática o que se observa é que geralmente eles são bi-alélicos. Embora os SNPs sejam marcadores bi-alélicos, e portanto menos informativos que os marcadores SSRs (por loco), eles abrem a possibilidade para a construção de mapas genéticos altamente saturados, uma vez que estes são a forma mais pontual para detecção de polimorfismo no genoma (EDWARDS et al., 2007).

Figura 1 – Mudança da importância relativa dos diferentes marcadores moleculares ao longo do tempo. O eixo horizontal indica o tempo e o eixo vertical corresponde ao uso dos marcadores moleculares em termos relativos (Modificado de SCHLÖTTERER, 2004)

A Figura 1 mostra a importância relativa de diferentes tipos de marcadores genéticos desde o surgimento dos RFLPs. A construção de mapas genéticos e o mapeamento de QTLs sofreram grandes avanços com o surgimento dos marcadores moleculares, e também com as alterações no tipo do marcador usado. Atualmente, o número de marcadores disponíveis vem se tornando cada vez maior, sendo necessários algoritmos eficientes para a construção dos mapas genéticos, que estão

se tornando cada vez mais saturados.

2.1.2 Algumas aplicações dos mapas genéticos

Os mapas de ligação são de fundamental importância em diversos tipos de estudos genéticos, uma vez que fornecem “pontos de referência” ao longo do genoma. Por exemplo, através da comparação de mapas genéticos é possível estudar processos evolutivos do genoma, verificando-se sintenia entre espécies, rearranjos, duplicações e deleções de segmentos cromossômicos. A comparação de mapas genéticos permite também que espécies modelo forneçam diretrizes a respeito dos genomas e da evolução de diversas espécies cultivadas (PATERSON et al., 2000).

Ahn e Tanksley (1993) construíram mapas genéticos de arroz e milho baseados em locos ortólogos detectados através de um grupo comum de clones de cDNAs. Nesse trabalho, foram verificados grupos de ligação conservados representando dois terços dos genomas das duas espécies. Em alguns casos, cromossomos inteiros ou braços cromossômicos eram praticamente idênticos quanto à ordem e ao conteúdo dos genes. Além disso, cerca de 72% dos genes duplicados durante a poliploidização ainda estão presentes em duplicata no genoma do milho.

Estudos comparando mapas genéticos foram feitos em diversas outras espécies, tais como sorgo e milho (WHITKUS; DOEBLEY; LEE, 1992), arroz e cevada (KILIAN et al., 1995), sorgo e cana-de-açúcar (DUFOR et al., 1997), cana-de-açúcar, sorgo e milho (GUIMARAES; SILLS; SOBRAL, 1997), arroz e *Arabidopsis* (DEVOS et al., 1999), tomate cultivado e tomate selvagem (CHETELAT et al., 2000), batata e *Arabidopsis* (GEBHARDT et al., 2003), sorgo, cana-de-açúcar, milho, arroz, milheto, capim-buffel, trigo, cevada, aveia e centeio (BOWERS et al., 2003), algodão e *Arabidopsis* (RONG et al., 2005).

Uma aplicação muito importante dos mapas genéticos é relacionada ao estudo de locos que controlam caracteres quantitativos (QTL - *Quantitative Trait Loci*). Comparando a segregação dos marcadores com valores de médias fenotípicas de cada indivíduo da progênie é possível realizar o mapeamento de QTLs, ou seja, é possível estimar o número, a posição, os efeitos e as interações entre os QTLs, buscando-se associações entre fenótipo e genótipo. Como os mapas genéticos são a base para o mapeamento de QTLs, é de fundamental importância que estes sejam bem construídos. O mapeamento de QTLs pode levar ao melhor entendimento da base genética dos caracteres quantitativos, bem como da correlação entre estes caracteres, uma vez que é possível estimar se esta correlação é causada por ligação ou pleiotropia (FALCONER; MACKAY, 1996). Esta informação é

importante para definição de estratégias de melhoramento, pois a correlação causada pela ligação é transitória e a correlação causada pela pleiotropia permanecerá em gerações futuras. Outro exemplo refere-se à base genética da interação genótipo $\times$ ambiente. Usando resultados de experimentos, é possível saber se um QTL tem efeito estável ou instável em vários ambientes, bem como a classificação desses ambientes. Com estas informações, os melhoristas podem definir estratégias para capitalizar a interação e obter genótipos superiores via seleção assistida.

Outra técnica promissora é a clonagem de QTLs. Clonar um QTL significa identificar a sequência de DNA (codante ou não-codante) responsável por um QTL (SALVI; TUBEROSA, 2005). Um dos métodos de clonagem utilizado na prática é a clonagem posicional. Para tanto é necessário aumentar a precisão da localização dos mapas convencionais de QTLs através de mapeamento fino. Este tipo de mapeamento requer a saturação do mapa em locais específicos fazendo uso de linhagens quase-isogênicas (NIL - *Near Isogenic Lines*) (MORGANTE; SALAMINI, 2003; SALVI; TUBEROSA, 2005). Uma vez que os genes responsáveis pelas variações quantitativas estão disponíveis, tais informações podem ser transferidas a programas de melhoramento, que podem então realizar a seleção assistida por marcadores (MORGANTE; SALAMINI, 2003).

2.1.3 Construção de Mapas Genéticos

A construção de mapas de ligação é feita basicamente em duas etapas: *i*) agrupamento dos marcadores moleculares em grupos de ligação; *ii*) estimativa da ordem e da distância entre os marcadores dentro dos grupos de ligação. Para tanto, são necessários dados de marcadores para os indivíduos da população experimental analisada, como por exemplo, uma população F_2 . A partir desses dados, é possível obter estas estimativas através de metodologias estatísticas apropriadas. As estatísticas usadas para a primeira etapa são usualmente o teste de qui-quadrado (χ^2), o teste da razão de verossimilhança ou um limiar para a frequência de recombinação (DOERGE, 1996). Contudo, essa etapa normalmente é facilitada pelo fato de que vários tipos de marcadores (como os microssatélites e os RFLPs) possuem posição cromossômica predeterminada. Conseqüentemente, a obtenção da ordem dos marcadores dentro dos grupos de ligação normalmente é a etapa do mapeamento que merece mais atenção.

Escolher a melhor ordem dos marcadores em um grupo de ligação enquadra-se numa classe de problemas muito conhecido na matemática e nas ciências da computação como “classe dos problemas de tempo polinomial não determinístico de difícil resolução” (*NP Hard Problem*) (MESTER et

al., 2003). O já citado problema do caixeiro viajante (*TSP - Traveling Salesman Problem*) constitui uma sub-classe desses problemas. No TSP, é impossível obter uma solução exata quando o número de marcadores é grande. Portanto, são necessários algoritmos que fornecem soluções aproximadas, os quais são o foco de estudo deste trabalho.

Resolvido o problema da ordenação, para completar a construção do mapa são necessárias estimativas das distâncias entre os marcadores. Como os parâmetros estimados entre os marcadores são frações de recombinação, que não são aditivas, faz-se necessário converter esses valores em medidas de distância. Para tanto, são utilizadas funções de mapeamento (LIU, 1998). Muitas funções foram propostas, sendo que as mais usuais são as seguintes:

$$d_{ij} = -\frac{1}{2} \ln(1 - 2r_{ij}) \quad (\text{HALDANE, 1919}) \quad (1)$$

$$d_{ij} = r_{ij} \quad (\text{MORGAN, 1928}) \quad (2)$$

$$d_{ij} = \frac{1}{4} \ln \left[\frac{1 + 2r_{ij}}{1 - 2r_{ij}} \right] \quad (\text{KOSAMBI, 1944}) \quad (3)$$

sendo r_{ij} a fração de recombinação entre os locos i e j , e d_{ij} a distância (em Morgans) entre os locos i e j .

2.2 Cadeias de Markov

Cadeias de Markov são modelos matemáticos usados para estudar processos resultantes de seqüências de observações, exatamente como ocorre nos mapas genéticos. Dada a grande importância das cadeias de Markov no mapeamento, serão apresentadas a seguir duas situações: uma abordando o processo markoviano observável e outra abordando o processo markoviano oculto. Ambos foram apresentados por Rabiner (1989), cujas idéias foram aqui adaptadas. Posteriormente, será feita a expansão dos conceitos apresentados para o caso dos mapas genéticos, com base em idéias apresentadas por Lander e Green (1987) e Jiang e Zeng (1997).

2.2.1 A Cadeia de Markov Observável

Para introduzir o assunto, um exemplo de simples entendimento será apresentado. Considere que o clima de um determinado local pode apresentar um dos três estados: *chuvoso*, *nublado* ou *ensolarado*. Considere agora uma seqüência de quatro dias e os respectivos estados que o clima pode apresentar nesses dias, como por exemplo, *chuvoso*, *nublado*, *chuvoso* e *ensolarado*. O estado

chuvoso pode ser representado por E_1 , o estado nublado por E_2 e o estado ensolarado por E_3 ; os dias podem ser representados por $t = 1, 2, 3, 4$ e o número de estados por $N = 3$. Então, outra forma de representar a sequência de observações exemplificada seria $\{E_1, E_2, E_1, E_3\}$, correspondendo aos dias $t = 1, 2, 3, 4$. É de interesse saber a probabilidade de ocorrência desta sequência de observações.

O modelo de Markov assume que a probabilidade de ocorrência de uma determinada condição climática em um dia depende exclusivamente da condição climática do dia anterior. Essa é a chamada *independência condicional* ou *propriedade markoviana*. Assumindo-se a independência condicional, denomina-se *probabilidade de transição* a probabilidade de observar o estado E_j no tempo t , dado que no tempo $t - 1$ o estado observado foi E_i . Estas probabilidades podem ser arranjadas em uma matriz, cujo número de linhas e o número de colunas é igual ao número de estados N . Esta matriz é chamada de *matriz de transição*, e é denotada por

$$\mathbf{H} = [h_{ij}]_{N \times N} \quad (4)$$

sendo h_{ij} a probabilidade do estado atual ser E_j dado que o estado anterior foi E_i .

A Figura 2 mostra todas as possíveis transições entre os estados do exemplo e suas respectivas probabilidades h_{ij} . Então, a probabilidade da sequência de observações $\{E_1, E_2, E_1, E_3\}$ pode ser calculada simplesmente multiplicando-se as probabilidades de transição entre os estados da sequência. Assumindo-se o modelo de Markov e a matriz de transição $[h_{ij}]_{3 \times 3}$, a probabilidade de ocorrência da sequência $\{E_1, E_2, E_1, E_3\}$, dado que o primeiro dia foi E_1 , é o produto das *probabilidades de transição* entre os estados da sequência:

$$P[\{E_1, E_2, E_1, E_3\} | \text{Modelo}] = h_{12} \times h_{21} \times h_{13} \quad (5)$$

Nota-se que esta multiplicação pode ser feita graças à propriedade markoviana mencionada anteriormente.

2.2.2 A Cadeia de Markov Oculta

Os modelos de Markov ocultos assumem que os estados da cadeia estão ocultos atrás das observações. Em outras palavras, o que é observado não é necessariamente um estado, mas sim um símbolo que diz algo a respeito do verdadeiro estado. Para melhor compreensão deste modelo, será

Figura 2 – Cadeia de Markov com três estados representando condições climáticas. O estado chuvoso está representado por E_1 , o estado nublado por E_2 e o estado ensolarado por E_3 . As probabilidades de transição entre estes estados estão representadas por h_{ij} , sendo que i é o estado de “partida” e j é o estado de “chegada”

apresentado novamente um exemplo didático.

Considere que existem quatro urnas numeradas. Dentro de cada urna existe uma proporção diferente de quatro cores de bolas: vermelha, azul, amarela e verde. Imagine que o seguinte experimento seja realizado: escolhe-se uma urna aleatoriamente, retira-se uma bola e anota-se a cor. A bola é então reposta na urna de origem. Em seguida, escolhe-se outra urna, que pode inclusive ser a mesma, de acordo com uma probabilidade de transição, e repete-se o processo. Procede-se novamente dessa forma por 10 vezes. Neste caso, é de interesse saber qual é a probabilidade de observar uma determinada seqüência de cores, sem saber de quais urnas foram retiradas as bolas. Diz-se então que os estados (urnas) estão ocultos. Este experimento pode ser modelado usando a *cadeia de Markov oculta*.

Neste caso, os estados são denotados por $E = \{E_1, E_2, \dots, E_N\}$ (N estados), os símbolos das observações são denotados por $V = \{v_1, v_2, \dots, v_U\}$ (U símbolos), o estado atual é denotado por τ_t , o número de observações é denotado por Y e a seqüência de observações é denotada por $O = O_1, O_2, \dots, O_Y$. No exemplo, $E = \{E_1, E_2, E_3, E_4\}$ (urnas), $V = \{v_1, v_2, v_3, v_4\}$ (vermelho, azul, verde e amarelo) e $Y = 10$. Verificando-se a proporção das cores das bolas nas urnas, é possível *inferir* sobre a urna de origem de cada bola. Por exemplo, se uma urna tiver apenas bolas vermelhas, uma outra apenas bolas amarelas, uma outra apenas bolas azuis e finalmente a última apenas bolas verdes, é possível dizer, com certeza, de que urna foi retirada cada bola. Se mais de

uma cor ocorrer para cada urna, é impossível saber ao certo de que urna foi retirada a bola, mas é possível associar uma probabilidade a essa incerteza. A Figura 3 mostra um exemplo de uma cadeia de Markov oculta.

Figura 3 – Cadeia de Markov oculta com quatro estados (urnas) e quatro símbolos (cores). A proporção das cores das bolas nas urnas fornece informações sobre a origem das observações. Por exemplo, a probabilidade da bola vermelha ter sido retirada da urna 1 é maior do que a mesma bola ter sido retirada de qualquer outra urna

A função de probabilidade que liga os estados (urnas) às observações (cores) é chamada de *função de emissão*, e é definida por:

$$b_j(w) = P[v_w \text{ em } t | \tau_t = E_j], \quad 1 \leq j \leq N \text{ e } 1 \leq w \leq U. \quad (6)$$

De maneira geral, a distribuição de probabilidades dos símbolos das observações para um determinado estado E_j é

$$\mathbf{B} = \{b_j(w)\} \quad (7)$$

Esta distribuição fornece a probabilidade de observar um determinado símbolo dado um determinado estado. Por exemplo, verificando a proporção de cores nas urnas da Figura 3, a distribuição

de probabilidade B para $j = 1$ (urna 1) é:

$$\mathbf{B} = \{b_1(w)\} = \begin{cases} 17/20 & \text{se } w = 1 (v_1=\text{vermelho}) \\ 1/20 & \text{se } w = 2 (v_2=\text{azul}) \\ 1/20 & \text{se } w = 3 (v_3=\text{verde}) \\ 1/20 & \text{se } w = 4 (v_4=\text{amarelo}) \end{cases} \quad (8)$$

O raciocínio análogo pode ser feito para as demais urnas. Pode ainda ser definida a distribuição de probabilidade inicial, que indica a probabilidade do estado inicial τ_1 ser E_i :

$$\pi = \{\pi_i\}, \text{ sendo } \pi_i = P[\tau_1 = E_i], \quad 1 \leq i \leq N. \quad (9)$$

Por conveniência, a matriz de transição $\mathbf{H}$, a distribuição de probabilidade $\mathbf{B}$ e a distribuição de probabilidade inicial π podem ser representadas por λ (parâmetros do modelo):

$$\lambda = (\mathbf{H}, \mathbf{B}, \pi) \quad (10)$$

Finalmente, a probabilidade da seqüência de observações O dado o modelo λ , pode ser obtida através da eq. (11):

$$P(O|\lambda) = \sum_{\tau_1, \tau_2, \dots, \tau_T} \pi_{\tau_1} b_{\tau_1}(O_1) h_{\tau_1 \tau_2} b_{\tau_2}(O_2) \dots h_{\tau_{T-1} \tau_T} b_{\tau_T}(O_T) \quad (11)$$

Se este experimento for repetido n vezes, o produto das probabilidades $P(O|\lambda)$ para todo n é proporcional à verossimilhança do modelo λ , calculada via cadeia de Markov oculta. Portanto, com base nessas idéias, é possível obter estimativas de máxima verossimilhança para os parâmetros do modelo λ . Detalhes sobre como aplicar esses conceitos aos mapas genéticos serão apresentados a seguir.

2.3 Critérios para Avaliação de Ordens

Um dos pontos importantes na ordenação de marcadores é o critério que será usado para julgar se uma determinada ordem é melhor que outra. Através desses critérios é possível associar um valor que representa o quão boa é uma ordem de marcadores à luz dos dados. Tais critérios são

importantes, pois são usados em passos dos algoritmos de ordenação, ou ainda no momento da verificação de erros após a construção do mapa. Diversos critérios desenvolvidos serão apresentados a seguir.

2.3.1 Verossimilhança Via Cadeias de Markov Ocultas (LHMC)

Lander e Green (1987) usaram as cadeias de Markov ocultas para reconstrução de mapas genéticos e para o cálculo da sua verossimilhança. Jiang e Zeng (1997) também usaram as cadeias de Markov ocultas para desenvolver um algoritmo geral para analisar dados proveniente de marcadores dominantes e dados perdidos em populações derivadas de linhagens endogâmicas. Essas idéias serão apresentadas a seguir.

Considere uma população F_2 proveniente do cruzamento de duas linhagens P_1 e P_2 . Suponha que existam m marcadores posicionados em um mapa com ordem $M_1, \dots, M_m$. O *genótipo* do marcador $M_k, k = 1, \dots, m$, é denotado x_k e pode assumir valores 1, 0 e -1 se M_k for homozigoto como P_1 , heterozigoto ou homozigoto como P_2 , respectivamente. O *fenótipo* do marcador M_k , para o mesmo indivíduo, é denotado z_k . Quando o marcador é completamente observável, o fenótipo é igual ao genótipo, i.e., $z_k = x_k = \{1\}, = \{0\}$ ou $= \{-1\}$. Quando o marcador não é observado, como por exemplo quando há um dado perdido, o genótipo é desconhecido e $z_k = \{1, 0, -1\} = M$ (*missing*). Quando o marcador é parcialmente observável, como ocorre por exemplo com os marcadores dominantes, o fenótipo pode corresponder a mais de um genótipo. Arbitrariamente, se o fenótipo dominante representa o homozigoto do tipo P_1 ($z_k = \{1, 0\} = D$ para dominante ou $z_k \neq -1$), o fenótipo recessivo representa o homozigoto do tipo P_2 ($z_k = \{0, -1\} = R$ para recessivo ou $z_k \neq 1$).

A sequência $M_1, \dots, M_m$, pode ser vista como uma cadeia de Markov oculta. Neste caso, os estados ocultos são os genótipos dos marcadores e os símbolos observáveis são os fenótipos. Caso todos os marcadores sejam codominantes e não existam dados perdidos, a cadeia será observável. A matriz de transição entre os genótipos dos locos é função da fração de recombinação entre os locos e é dada por:

$$\mathbf{H}(r_k) = \begin{bmatrix} P(x_{k+1} = 1|x_k = 1) & P(x_{k+1} = 0|x_k = 1) & P(x_{k+1} = -1|x_k = 1) \\ P(x_{k+1} = 1|x_k = 0) & P(x_{k+1} = 0|x_k = 0) & P(x_{k+1} = -1|x_k = 0) \\ P(x_{k+1} = 1|x_k = -1) & P(x_{k+1} = 0|x_k = -1) & P(x_{k+1} = -1|x_k = -1) \end{bmatrix} \quad (12)$$

$$= \begin{bmatrix} (1-r_k)^2 & 2r_k(1-r_k) & r_k^2 \\ r_k(1-r_k) & (1-r_k)^2 + r_k^2 & r_k(1-r_k) \\ r_k^2 & 2r_k(1-r_k) & (1-r_k)^2 \end{bmatrix}, \quad (13)$$

sendo r_k a fração de recombinação entre os marcadores M_k e M_{k+1} .

As funções de emissão são específicas para cada tipo de genótipo x_k dos marcadores, ou seja,

1, 0 ou -1 . Por exemplo, para $x_k = 1$ (i.e., homozigoto dominante), a distribuição de probabilidades de emissão considerando-se um marcador dominante é

$$\mathbf{B} = \{b_1(z_k)\} = \begin{cases} 1 & \text{se } z_k = D \\ 0 & \text{se } z_k = -1 \end{cases} \quad (14)$$

e assim sucessivamente para os demais genótipos e tipos de marcador. A distribuição de probabilidades inicial (*a priori*) é função das proporções mendelianas para o tipo de população experimental considerada e é facilmente obtida. No caso de uma população F_2 esta distribuição é:

$$\pi = \{\pi_{x_k}\} = \begin{cases} 1/4 & \text{se } x_k = 1 \\ 1/2 & \text{se } x_k = 0 \\ 1/4 & \text{se } x_k = -1 \end{cases} \quad (15)$$

Jiang e Zeng (1997) apresentaram uma notação matricial para o cálculo da verossimilhança de um dado mapa genético (seqüência de locos ordenados e com suas distâncias estimadas). Na notação empregada, distribuição de probabilidades *a priori*, análoga à π , é representada pelo vetor $\mathbf{q}$. Para uma população F_2 ,

$$\mathbf{q}' = \left[\frac{1}{4}, \quad \frac{1}{2}, \quad \frac{1}{4} \right]. \quad (16)$$

As matrizes $\mathbf{I}_{z_j}$, sendo $z_j = M, D, R, 1, 0$ ou -1 (dependendo do conteúdo de informação do fenótipo do marcador M_j), são matrizes indicadoras dos fenótipos, que são usadas para pré-multiplicar a matriz de transição, eliminando dessa forma os elementos cujas transições têm proba-

bilidade de ocorrência zero:

$$\mathbf{I}_M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{I}_D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{I}_R = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{I}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{I}_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{I}_{-1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Por exemplo, quando o fenótipo do marcador é D , $\mathbf{I}_D$ elimina a última linha da matriz de transição na multiplicação $\mathbf{I}_D \mathbf{H}(r_k)$, uma vez que o fenótipo $z_k = D$ exclui a possibilidade do genótipo ser $x_k = -1$.

Para o cálculo da verossimilhança do modelo, sejam $z_{l1}, \dots, z_{lm}$ as observações dos fenótipos dos marcadores $M_1, \dots, M_m$ para o l -ésimo indivíduo em uma população F_2 . A verossimilhança das observações depende: *i*) do modelo assumido para os eventos de recombinação; *ii*) da ordem dos marcadores. Assumindo-se ausência de interferência e que os marcadores estão no mesmo grupo de ligação, com ordem $M_1, \dots, M_m$, a verossimilhança é dada por

$$L = \prod_{l=1}^n \mathbf{q}' \mathbf{I}_{z_{l1}} \mathbf{H}(r_1) \mathbf{I}_{z_{l2}} \mathbf{H}(r_2) \mathbf{I}_{z_{l3}} \dots \mathbf{H}(r_{m-1}) \mathbf{I}_{z_{lm}} \mathbf{c} \quad (17)$$

sendo n o número de indivíduos e $\mathbf{c} = \{1\}_{3 \times 1}$. Esta verossimilhança pode ser usada como critério para a comparação de diferentes ordens (LANDER; GREEN, 1987).

Um ponto importante a ser ressaltado é que, dada uma ordem, é possível obter estimativas multiponto das frações de recombinação entre os marcadores, via cadeia de Markov oculta. Isso permite recuperar informação de todo o mapa de ligação, melhorando as estimativas, principalmente quando há marcadores dominantes e dados perdidos (LANDER; GREEN, 1987). Obtidas tais estimativas multiponto, a verossimilhança pode ser recalculada de forma ainda mais precisa, melhorando conseqüentemente o processo de construção dos mapas.

Seja uma determinada ordem $M_1, M_2, \dots, M_m$. As frações de recombinação podem ser estimadas através do algoritmo EM (DEMPSTER; LAIRD; RUBIN, 1977) num processo iterativo. A atualização da estimativa a cada iteração, dependerá da estimativa da fração de recombinação entre alguns ou todos intervalos, dependendo da informatividade dos fenótipos dos marcadores. Para a at-

ualização da fração de recombinação do k -ésimo intervalo, é necessário calcular $P(x_k, x_{k+1} | z_i, \dots, z_o)$ para cada indivíduo, sendo i e o as posições dos marcadores completamente informativos mais adjacentes à esquerda e à direita da posição k . $P(x_k, x_{k+1} | z_i, \dots, z_o)$ pode ser expresso na forma matricial como

$$\mathbf{A}_k = \frac{\left[(\mathbf{q} \circ \mathbf{p}_k^L) (\mathbf{p}_{k+1}^R)' \right] \circ \mathbf{H}(r_k)}{(\mathbf{q} \circ \mathbf{p}_k^L)' \mathbf{H}(r_k) \mathbf{p}_{k+1}^R} \quad (18)$$

em que $\circ$ denota o produto elemento a elemento das matrizes e $'$ denota transposição de vetores ou matrizes. $\mathbf{p}_k^R$ é um vetor que contém as probabilidades de observar uma sequência de marcadores com fenótipos $z_{k+1}, \dots, z_o$, à direita do marcador com genótipo x_k , e pode ser calculado através da equação

$$\mathbf{p}_k^R = \mathbf{H}(r_k) \mathbf{I}_{z_{k+1}} \mathbf{H}(r_{k+1}) \mathbf{I}_{z_{k+2}}, \dots, \mathbf{H}(r_{o-1}) \mathbf{I}_{z_o} \mathbf{c}. \quad (19)$$

Analogamente, $\mathbf{p}_k^L$ é um vetor que contém as probabilidades de observar uma sequência de marcadores com fenótipos $z_i, \dots, z_k$, à esquerda do marcador com genótipo x_k e pode ser calculado através da equação

$$\mathbf{p}_k^L = \mathbf{I}_{z_k} \mathbf{H}(r_{k-1}) \mathbf{I}_{z_{k-1}} \mathbf{H}(r_{k-2}) \mathbf{I}_{z_{k-2}}, \dots, \mathbf{H}(r_i) \mathbf{I}_{z_i} \mathbf{c}. \quad (20)$$

Para uma população F_2 , $\mathbf{A}_k$ é uma matriz 3×3 , pois, neste caso, existem apenas 3 classes genotípicas. Seja a_{jcd} um elemento da c -ésima linha e da d -ésima coluna de $\mathbf{A}_k$ para o indivíduo j . A frequência de recombinação para o k -ésimo intervalo pode ser atualizada através da equação

$$\tilde{r}_k = \frac{1}{2n} \sum_{j=1}^n \left[a_{j12} + 2a_{j13} + a_{j21} + \frac{r_k^2}{(1 - r_k)^2} 2a_{j22} + a_{j23} + 2a_{j31} + a_{j32} \right], \quad (21)$$

sendo r_k a estimativa da fração de recombinação obtida na iteração anterior. Dessa forma, completa-se a reconstrução do mapa, obtendo-se a verossimilhança e as estimativas *multiponto* das frações de recombinação via cadeias de Markov ocultas. Jiang e Zeng (1997) mostraram ainda como usar essas idéias para outros tipos de populações derivadas de linhagens endogâmicas, como retrocruzamentos, progênies com qualquer número de auto-fecundações e linhagens endogâmicas recombinantes.

2.3.2 SARF, PARF, SALOD e outros

Ao contrário da verossimilhança obtida via cadeia de Markov oculta, a qual leva em conta a informação de todos os marcadores, para o cálculo dos critérios SARF (*sum of adjacent recom-*

bination fractions - FALK, 1989), PARF (*product of adjacent recombination fractions* - WILSON, 1988) e SALOD (*sum of adjacent lod scores* - WEEKS; LANGE, 1987), são necessárias estatísticas obtidas através da análises de dois pontos. Considere m locos $M_1, M_2, \dots, M_m$ em um grupo de ligação. As seguintes matrizes são usadas para denotar as estatísticas obtidas através das análises de dois pontos (LIU, 1998):

$$\mathbf{R} = [\hat{r}_{M_i M_j}]_{m \times m} \quad (22)$$

$$\mathbf{LOD} = [\text{lod}_{M_i M_j}]_{m \times m} \quad (23)$$

sendo $\hat{r}_{M_i M_j}$ e $\text{lod}_{M_i M_j}$ a estimativa de máxima verossimilhança de dois pontos da fração de recombinação e o LOD (*logarithm of odds ratio* - MORTON, 1955), respectivamente, entre os locos M_i e M_j ; i e j são indicadores das posições dos locos para uma determinada ordem (i e j se alteram se a ordem dos locos for diferente). Dessa forma, as variáveis M_i e M_j nas equações 22 e 23 são indicadoras de posições no genoma e não indicadoras das posições dos marcadores (LIU, 1998). A seguir, são apresentados exemplos das matrizes $\mathbf{R}$ e $\mathbf{LOD}$ para 4 marcadores pertencentes a um mesmo grupo de ligação:

$$\mathbf{R} = \begin{bmatrix} 0,00 & 0,01 & 0,04 & 0,05 \\ 0,01 & 0,00 & 0,02 & 0,03 \\ 0,04 & 0,02 & 0,00 & 0,01 \\ 0,05 & 0,03 & 0,01 & 0,00 \end{bmatrix},$$

$$\mathbf{LOD} = \begin{bmatrix} 78,83 & 7,34 & 4,14 & 3,19 \\ 7,34 & 76,23 & 7,12 & 3,89 \\ 4,14 & 7,12 & 81,44 & 7,44 \\ 3,19 & 3,89 & 7,44 & 79,20 \end{bmatrix}$$

Nota-se que $\hat{r}_{M_i M_j} = \hat{r}_{M_j M_i}$ e $\text{lod}_{M_i M_j} = \text{lod}_{M_j M_i}$, ou seja, a matriz é simétrica. A diagonal principal pode ser desconsiderada, pois trata-se de estimativas entre o mesmo marcador. Portanto, para fins práticos, são calculados apenas os elementos localizados acima da diagonal principal.

A construção de um mapa de ligação pode ser vista como um processo de minimização das

frações de recombinação entre marcadores adjacentes (STURTEVANT, 1913). O índice SARF (*sum of adjacent recombination fractions* - FALK, 1989) fornece a soma das frações de recombinação adjacentes de uma dada ordem. Dessa forma, usando-se este critério, as ordens que apresentam menores SARFs entre os marcadores adjacentes são consideradas as mais prováveis de serem as ordens corretas. O SARF para uma dada ordem $1, 2, 3, \dots, m-1, m$ é dada pela fórmula (LIU, 1998):

$$SARF = \sum_{i=1}^{m-1} \hat{r}_{M_i M_{i+1}} \quad (24)$$

Analogamente ao SARF, o critério PARF (*product of adjacent recombination fractions* - WILSON, 1988) baseia-se no produto das frações de recombinação entre marcadores adjacentes. Usando-se este critério, as ordens que apresentam menores PARFs entre os marcadores adjacentes são consideradas as mais prováveis de serem as corretas. O PARF para uma dada ordem $1, 2, 3, \dots, m-1, m$ é dada por (LIU, 1998):

$$PARF = \prod_{i=1}^{m-1} \hat{r}_{M_i M_{i+1}} \quad (25)$$

O índice SALOD (*sum of adjacent lod scores* - WEEKS; LANGE, 1987) baseia-se na obtenção da soma dos LODs entre marcadores adjacentes. Usando-se este critério, as ordens que apresentam maiores SALODs entre os marcadores adjacentes são consideradas as mais prováveis de serem as corretas. O SALOD para uma determinada ordem é:

$$SALOD = \sum_{i=1}^{m-1} lod_{M_i M_{i+1}} \quad (26)$$

Segundo Olson e Boehnke (1990), o critério SALOD é sensível à informatividade dos marcadores e tende a agrupar os locos mais informativos. Para contornar este problema, Olson e Boehnke (1990) propuseram uma modificação no SALOD:

$$SALEQ = \sum_{i=1}^{m-1} \frac{lod_{M_i M_{i+1}}}{EN_{M_i M_{i+1}}} \quad (27)$$

sendo $EN_{M_i M_{i+1}}$ o número equivalente de meioses completamente informativas entre os locos M_i e M_{i+1} , que pode ser estimado por

$$EN_{M_i M_{i+1}} = \log_{10} 2 + \hat{r}_{M_i M_{i+1}} \log_{10} \hat{r}_{M_i M_{i+1}} + (1 - \hat{r}_{M_i M_{i+1}}) \log_{10} (1 - \hat{r}_{M_i M_{i+1}}), \quad (28)$$

Outra modificação proposta por Olson e Boehnke (1990) no SALOD originou o critério SALPIC (*SALOD-polymorphism information content*). O SALPIC leva em conta o conteúdo de informação de polimorfismo de cada marcador considerado e é dado pela fórmula (LIU, 1998)

$$SALPIC = \sum_{i=1}^{m-1} \frac{\log_{M_i M_{i+1}}}{PIC_{M_i} PIC_{M_{i+1}}}, \quad (29)$$

sendo PIC o conteúdo de informação de polimorfismo, definido por Botstein et al. (1980).

Weeks e Lange (1987) propuseram usar o método dos quadrados mínimos (LS - *least squares*) para obter estimativas de distâncias entre os marcadores de um dado mapa. Essas estimativas são obtidas de modo que a soma de quadrados dos desvios entre os valores observados (distâncias calculadas através das análises de dois pontos) e os estimados (distância real entre os locos) seja mínima. A ordem com o menor quadrado mínimo é considerada a melhor ordem. O método dos quadrados mínimos ponderados (WLS - *weighted least squares*) (STAM, 1993) usa o LOD dos marcadores em questão para ponderar os quadrados dos desvios, baseando-se no fato de que as estimativas das distâncias entre os marcadores têm diferentes níveis de acurácia.

2.4 Algoritmos de Ordenação

2.4.1 TRY

O algoritmo *TRY*, proposto por Lander et al. (1987) e implementado no programa computacional MAPMAKER/EXP, baseia-se na idéia de acrescentar marcadores sequencialmente em todos os intervalos de uma subordem inicial, com m' marcadores, testando-se qual é a posição mais provável de cada um deles. Para isso, o grupo de ligação já deve estar estabelecido. Esta subordem inicial pode ser obtida de maneira exaustiva, ou seja, comparando as $m'!/2$ ordens possíveis.

Primeiramente, para cada uma das $m'!/2$ ordens, é reconstruído o mapa e calcula-se a verossimilhança via cadeia de Markov oculta. O sub-mapa que apresentar a maior verossimilhança entre as $m'!/2$ calculadas é considerado o sub-mapa correto e irá receber os marcadores ainda não posicionados. Escolhe-se aleatoriamente um marcador ainda não mapeado. Posiciona-se este marcador no início, no fim e em todos os intervalos entre os marcadores, ou seja, nas $m' + 1$ posições possíveis, sendo m' o número de locos já ordenados. Para cada uma dessas posições o mapa é reconstruído, ou seja, são obtidas novas estimativas multiponto de frações de recombinação, e é calculada a verossimilhança da nova ordem. A ordem que apresentar a maior verossimilhança é considerada a ordem

correta e o marcador é então posicionado. A Figura 4 mostra esquematicamente este procedimento.

Figura 4 – Exemplo de construção de um mapa genético usando o algoritmo *TRY*. O marcador M_6 deve ser posicionado numa ordem pré-estabelecida com $m' = 5$. As setas indicam as posições em que o marcador M_6 será testado

Com base nessa nova ordem obtida, escolhe-se o próximo marcador mais informativo a ser posicionado e repete-se o passo anterior. Este procedimento é repetido até que o último marcador tenha sido posicionado, estimando-se assim o mapa genético com todos marcadores do grupo de ligação.

2.4.2 Seriation (SER)

O algoritmo *seriation* (SER) foi primeiramente proposto por Gelfand (1971) para recuperar a ordem temporal de sítios arqueológicos de acordo com os artefatos localizados nos mesmos. Para tanto, foi calculada uma matriz de similaridade usando os artefatos em comum encontrados nos sítios. No contexto de ordenação de locos em mapas genéticos, o algoritmo SER foi introduzido por Buetow e Chakravarti (1987a) a fim de obter o SARF mínimo de um conjunto de locos a serem ordenados dentro de um grupo de ligação (LIU, 1998). A partir das estimativas da fração de recombinação de dois pontos, os marcadores são ordenados.

Seja $M_i, i = 1, \dots, k, \dots, m$, o i -ésimo marcador do conjunto de m marcadores e seja $r_{M_i M_j}$ a fração de recombinação entre dois marcadores quaisquer, M_i e M_j .

- i.) Início: começa-se com cada um dos m marcadores $M_1, M_2, \dots, M_m$. Quando M_i é o marcador inicial, posiciona-se M_j à direita de M_i se a fração de recombinação é a menor entre as frações de recombinação entre M_i com todos os outros $m - 1$ marcadores. Depois deste passo, restam $m - 2$ marcadores para serem posicionados.
- ii.) $m - 2$ marcadores restantes: escolhe-se o marcador M_k dos marcadores não posicionados, tal que ele possua a menor fração de recombinação com M_i . Comparam-se as frações de

recombinação de M_k com os dois locos que estão atualmente externos no grupo de marcadores já posicionados, M_{esq} (marcador do lado esquerdo) e M_{dir} (marcador do lado direito). Se

$$\hat{r}_{M_k M_{dir}} > \hat{r}_{M_k M_{esq}} \quad (30)$$

posiciona-se M_k à esquerda do grupo de marcadores atualmente posicionados. Se

$$\hat{r}_{M_k M_{esq}} > \hat{r}_{M_k M_{dir}} \quad (31)$$

posiciona-se M_k à direita. Repete-se o procedimento acima até que todos os marcadores estejam posicionados.

- iii.) Resolvendo empates: precisam ser considerados apenas quando os marcadores serão colocados no mesmo lado, por exemplo

$$M_{esq}, \dots, M_{dir}, (M_j, M_k), \quad \text{e} \quad \hat{r}_{M_j M_{dir}} = \hat{r}_{M_k M_{dir}} \quad (32)$$

Os dois marcadores (M_j e M_k) devem ser ordenados com respeito ao marcador mais externo naquele lado no grupo de marcadores ordenados. Se esta abordagem falhar em resolver o empate, os dois próximos marcadores internos no grupo ordenado devem ser considerados.

- iv.) Determinação da ordem dos marcadores: para m marcadores, m ordens serão obtidas. Quando a matriz de frações de recombinação é monótona, ou seja, os valores das estimativas aumentam monotonamente à medida que se afastam da diagonal principal, uma única ordem será obtida. Se a matriz não for monótona, um índice de continuidade (CI) pode ser usado para medir a adequação à monotonicidade de cada ordem obtida. Para cada uma das m ordens, o CI da matriz de fração de recombinação é estimado usando

$$CI = \sum_{i < j} \frac{\hat{r}_{M_i M_j}}{(j - i)^2} \quad (33)$$

sendo i e j indicadores das posições no genoma e $\hat{r}_{M_i M_j}$ a estimativa da fração de recombinação entre os marcadores localizados nas duas posições no genoma (eq. 22). A melhor ordem é aquela que resulta numa matriz com o menor CI (LIU, 1998).

2.4.3 *Rapid Chain Delineation - (RCD)*

O algoritmo RCD (DOERGE, 1996) é um algoritmo que tem como vantagem o fato de ser computacionalmente simples. Como é apresentado a seguir, sem cálculos complexos é possível obter a ordem dos marcadores no mapa. O RCD também requer as estimativas de $\hat{r}$ e não necessita que os marcadores estejam previamente separados em grupos de ligação. Este algoritmo pode ser descrito nos seguintes passos:

- i.) Verifica-se qual é o par de marcadores (M_i, M_j) que possui a menor fração de recombinação. Esses marcadores iniciarão a cadeia.
- ii.) Verifica-se qual é o marcador não mapeado que apresenta a menor fração de recombinação (estatisticamente significativa) com um dos dois marcadores terminais. Posiciona-se este marcador ao lado daquele com o qual apresentou a menor fração de recombinação.
- iii.) Este procedimento é repetido até que todos os marcadores sejam adicionados à cadeia.
- iv.) Quando não houver mais nenhum marcador significativamente ligado ao grupo com base nas frações de recombinação, um novo grupo de ligação é criado a partir do conjunto de marcadores não mapeados e o processo é repetido. Com isso, completa-se a construção do mapa.

2.4.4 *Recombination Counting and Ordering - (RECORD)*

O algoritmo RECORD (VAN OS et al., 2005) usa como critério de otimização a contagem de eventos de recombinação (COUNT) para ordenar marcadores dentro de um grupo de ligação pré-estabelecido. A sequência que apresentar o menor COUNT dentre as sequências avaliadas é considerada a que tem maior probabilidade de estar correta. Em um retrocruzamento sem dados perdidos, este número pode ser facilmente obtido através da soma da contagem do número de recombinantes para os pares de locos de uma determinada sequência. Quando a informação é incompleta (por exemplo, devido a dados perdidos, dominância em uma população F_2 , ou mistura de classes), este valor pode ser calculado através do número esperado de eventos de recombinação (s), baseado em estimativas de máxima verossimilhança das frações de recombinação entre locos adjacentes. Por exemplo, o valor esperado de recombinações do fenótipo $AAB_$ numa população F_2 com o loco

marcador A codominante e o loco marcador B dominante, pode ser calculado através da eq. 34:

$$E(s|AAB_-) = 0 \times \frac{(1-r)^2}{1-r^2} + 1 \times \frac{2r(1-r)}{1-r^2} = \frac{2r}{1+r} \quad (34)$$

sendo s o número esperado de recombinações condicionado à observação do fenótipo AAB_- e r a estimativa de máxima verossimilhança da fração de recombinação entre os locos A e B .

Dessa forma, pode ser construída a matriz $\mathbf{S} = [s_{M_i M_j}]_{m \times m}$ (m = número de marcadores), representando o número de eventos de recombinação entre os pares de marcadores M_i e M_j . O cálculo do critério COUNT para uma dada seqüência de m marcadores é obtida através da eq. (35):

$$COUNT = \sum_{i=1}^{m-1} s_{M_i M_{i+1}} \quad (35)$$

O algoritmo de construção do mapa é constituído dos seguintes passos:

- i.) Escolhe-se aleatoriamente um par de marcadores.
- ii.) Escolhe-se aleatoriamente um terceiro marcador e posiciona-se este marcador no início, no fim e entre os dois marcadores já posicionados. A ordem que apresentar o menor COUNT é considerada a ordem correta.
- iii.) Do mesmo modo, adicionam-se um a um os marcadores ainda não mapeados nas $m + 1$ posições, sendo m o número de locos já ordenados. Este passo é similar ao algoritmo *TRY*, explicado na Figura 4. Assim que todos os marcadores forem adicionados ao grupo de ligação, forma-se uma seqüência.
- iv.) Nesta seqüência é feita uma busca por menores COUNTs da seguinte forma:
 - 1.) Percorre-se a seqüência construída obtendo-se novas ordens através da inversão de subseqüências de tamanho $m' = 2, \dots, m - 1$. Por exemplo, considerando-se a seqüência 1, 2, 3, 4, 5, para $m' = 2$, as novas seqüências obtidas seriam 2, 1, 3, 4, 5; 1, 3, 2, 4, 5; 1, 2, 4, 3, 5 e 1, 2, 3, 5, 4; para $m' = 3$, têm-se 3, 2, 1, 4, 5; 1, 4, 3, 2, 5 e 1, 2, 5, 4, 3; para $m' = 4$, têm-se 4, 3, 2, 1, 5 e 1, 5, 4, 3, 2.
 - 2.) Para cada uma das novas seqüências obtidas é calculado o COUNT. Se a nova seqüência tiver menor COUNT que a seqüência anterior, ela é mantida.

- v.) Como este algoritmo não garante o mínimo global e obviamente depende de qual par de marcadores iniciou o processo, ele é repetido 10 vezes e é escolhida a sequência que apresentar o menor COUNT.

2.4.5 *Unidirectional Growth* - (UG)

O algoritmo *Unidirectional Growth* (UG) (TAN; FU, 2006) foi proposto com o objetivo de ser um método rápido e com grande acurácia para fazer a ordenação de marcadores. A idéia básica é encontrar uma extremidade terminal do mapa e crescer a partir dela até que o mapa esteja completo. Este algoritmo também requer as estimativas de fração de recombinação de dois pontos.

No contexto do algoritmo, um mapa é definido como uma sequência ordenada de alguns ou de todos os locos de um conjunto de locos. Dado um conjunto de locos, o menor mapa tem apenas um loco e o maior mapa inclui todos os locos disponíveis. O maior mapa é chamado de mapa completo, enquanto que mapas menores são chamados de mapas parciais. O procedimento para estimar a ordem dos marcadores requer uma medida de proximidade entre cada par de marcadores, definido como d_{ij} , que mede a distância entre dois locos i e j e tem a propriedade de ser aditiva (i.e., $d_{ij} = d_{ik} + d_{kj}$, $i < k < j$). Logo, como as estimativas de dois pontos da fração de recombinação não têm essa propriedade, Tan e Fu (2006) sugerem a seguinte transformação:

$$d_{ij} = \hat{r}_{ij} + \frac{2}{n_{ij}} \sum_k \hat{r}_{ik} \hat{r}_{jk}, \quad (36)$$

sendo que a soma é feita em todos os k locos tais que $r_{ij} > r_{ik}$ e $r_{ij} > r_{jk}$; n_{ij} é o número de indivíduos para os locos considerados. O algoritmo UG baseia-se em dois teoremas, provados por (TAN; FU, 2006), que garantem propriedades aditivas para as distâncias.

Teorema 1. *Para um conjunto de m locos com distância d_{ij} entre os locos i e j , define-se*

$$T_{ij} = 2d_{ij} - (S_i + S_j), \quad (37)$$

em que

$$S_i = \sum_{k \neq i} d_{ik}. \quad (38)$$

Supondo que a ordem real do mapa é 1-2-3-...- m ; então

$$\min(T_{12}, T_{m-1m}, T_{1m}) = \min_{i < j}(T_{ij}) \quad (39)$$

se a distância for aditiva.

Teorema 2. Supondo que um mapa verdadeiro de m locos seja 1-2-3-...- m , então, para a distância aditiva d_{ij} ,

$$W_{12} = \min_{i=2}^m (W_{1i}), \quad (40)$$

em que

$$W_{1i} = (m - 1)d_{1i} - S_i. \quad (41)$$

O teorema 1 é importante para a determinação da extremidade terminal do mapa, a partir da qual ocorrerá o crescimento unidirecional; com base no teorema 2, é possível determinar a ordem em que os marcadores serão acrescentados ao mapa.

Os passos do algoritmo UG são, então, os seguintes:

- i.) Determinação do mapa terminal e direção de crescimento: Calcula-se T_{ij} , para todo $i < j$. O par de locos f e g que resultarem no menor valor de T serão tomados como o mapa terminal f - g , ou seja, os demais marcadores serão acrescentados a estes em uma única direção. Os marcadores f - g passam agora a ser designados como loco $m + 1$. A distância ente o loco $m + 1$ e um loco simples i é definida como

$$d_{im+1} = \begin{cases} \frac{1}{2} (d_{if} + d_{ig} - d_{fg}) & \text{se } (d_{if} + d_{ig}) > d_{fg} \\ 0 & \text{caso contrário} \end{cases}$$

Computa-se, com a nova distância

$$W_{im+1} = (m - 2)d_{im+1} - S_i. \quad (42)$$

O loco h que minimiza o valor de W_{im+1} é escolhido como o próximo loco a ser adicionado ao mapa terminal. O mapa parcial resultante (o loco $m + 2$), será f - g - h se $d_{fh} > d_{gh}$ ou h - f - g caso contrário. A primeira situação indica que todo crescimento subsequente será da

esquerda para direita e a segunda, da direita para esquerda. Toma-se a variável contadora $k = 2$ e procede-se com o passo seguinte.

- ii.) Atualização da distância: calcula-se a distância entre o loco composto $m + k$ e um loco simples i

$$d_{im+k} = \min(d_{im+k-1}, d_{ij}) \quad (43)$$

sendo j o loco simples que leva ao loco composto $m + k$ através de sua inclusão no loco composto $m + k - 1$

- iii.) Crescimento do mapa: calcula-se S_i para cada loco simples i com a distância atualizada e então

$$W_{im+k} = (m - k - 1)d_{im+k} - S_i. \quad (44)$$

O loco que fornecer o menor valor de W é então adicionado ao mapa parcial. O mapa parcial resultante é chamado de loco $m + k + 1$

- iv.) Repetem-se os passos *ii* e *iii* até que o mapa completo seja obtido.

2.4.6 Outros Algoritmos de Ordenação

Devido ao fato do TSP estar presente em diversas áreas do conhecimento, foram propostos muitos outros algoritmos para contornar a situação. Por exemplo, o anelamento simulado (*simulated annealing* - SA) (THOMPSON 1984; WEEKS; LANGE, 1987) é um método desenvolvido em analogia à termodinâmica de cristalização de um líquido: em altas temperaturas as moléculas de um líquido movem-se livremente. Quando ele é esfriado, esta mobilidade é perdida e os átomos tendem a se organizar em linhas a fim de formar um cristal puro. Este cristal é o estado de menor energia no sistema. A base essencial deste processo é que o resfriamento ocorra lentamente permitindo que os átomos posicionem-se corretamente. A este processo é dado o nome de anelamento, e é essencial que ele ocorra para que um estado de baixa energia seja alcançado (LIU, 1998). Estas idéias de anelamento podem ser simuladas (METROPOLIS, 1953) podendo ser usadas no contexto do mapeamento genético.

Outro algoritmo utilizado para solucionar o TSP é o *branch and bound* (BB) (LATHROP et al., 1985). O BB é um algoritmo que busca o espaço completo de soluções para um dado problema

a fim de encontrar a melhor solução. Contudo, a enumeração explícita de todas as soluções é inviável devido à natureza exponencial do TSP. Para contornar este problema, utilizam-se limites superiores e inferiores para a função a ser otimizada. Com esse limite imposto, um grande número de subconjuntos, que apresentam soluções muito improváveis, são descartados.

2.5 Verificação de erros na ordenação

Devido ao fato de os algoritmos de ordenação utilizarem um certo grau de aleatoriedade e métodos aproximados durante a construção do mapa, erros cometidos podem ser acumulados durante as etapas seguintes. Este tipo de erro é inerente aos algoritmos utilizados, que são conhecidos na literatura como *greedy algorithms* (CORMEN et al., 2003). Nestes algoritmos são feitas escolhas ótimas locais em cada estágio, a fim de se obter um ótimo global. Em outras palavras, as decisões tomadas em um estágio são feitas considerando que as decisões no estágio anterior estão corretas, não sendo possível alterá-las (CORMEN et al., 2003). Por exemplo, em cada passo do algoritmo TRY, assume-se que a ordem anterior é correta para que o próximo marcador seja posicionado.

O algoritmo RIPPLE, proposto por Lander et al. (1987), tem como objetivo fazer verificações de pequenos erros de ordenação, verificando subordens alternativas através de buscas locais exaustivas. Dado um mapa de ligação previamente construído, o algoritmo RIPPLE permuta m' marcadores vizinhos e compara os $m'/2$ mapas possíveis através de algum critério. Primeiramente, as posições $1, \dots, m'$ são permutadas, depois as posições $2, \dots, m' + 1$, e assim por diante, até que todo o mapa tenha sido percorrido e essas ordens alternativas verificadas. Se algum dos mapas resultantes apresentar um valor melhor para o critério adotado, ele é considerado correto e o mapa anterior é descartado.

3 MATERIAL E MÉTODOS

Nas simulações realizadas buscou-se representar situações próximas da realidade. O grupo de ligação simulado tem vários marcadores com diversos padrões de mistura, envolvendo marcadores codominantes e dominantes. Foram consideradas também populações F_2 com dois tamanhos e quantidades crescentes de dados perdidos. Em situações reais, estes fatores considerados são muito comuns, e certamente influenciam no mapa obtido. Embora diversos estudos tenham sido feitos para avaliar algoritmos de ordenação e critérios para comparação de ordens (OLSON; BOEHNKE, 1990; HACKETT; BROADFOOT, 2003; WU et al., 2003), nenhum deles contempla estas situações, que evidentemente não podem ser ignoradas.

3.1 Material

3.1.1 Simulação dos Genótipos da População F_2

Neste trabalho considerou-se um mapa genético contendo um grupo de ligação de uma espécie vegetal hipotética, diplóide e monóica, com 21 marcadores moleculares com distância fixa entre eles de 3 centimorgans. As distâncias e o número de marcadores foram escolhidos devido à crescente disponibilidade dos marcadores moleculares, o que possibilita a construção de mapas cada vez mais saturados, e portanto com menores distâncias entre locos. Imagina-se que este grupo de ligação representa núcleos saturados que serão ordenados dentro dos mapas reais com muitos marcadores. Vale lembrar que mapas com marcadores mais distantes (por exemplo, 10 cM em média), geralmente são mais fáceis de estimar, ou seja, os mapas obtidos a partir desses dados, devem se aproximar mais do mapa real, o que não necessariamente ocorre em situações com marcadores mais próximos. Dessa forma, as conclusões obtidas neste trabalho podem facilmente ser extrapoladas para situações reais. A Figura 5 apresenta uma representação gráfica do cromossomo simulado.

Com base no mapa de ligação obtido, foram simuladas populações experimentais do tipo F_2 , muito comuns na prática. Para tanto, considerou-se o cruzamento de dois genótipos homozigóticos com posterior autofecundação. Este tipo de população permite o estudo adequado do efeito dos marcadores dominantes na obtenção da estimativa da ordem e da distância dos mapas. Isso é possível pois a população F_2 apresenta três classes genotípicas (i.e., AA ; Aa ; aa), sendo que elas podem ser misturadas com a utilização dos marcadores dominantes (i.e., $A_$; aa). A Figura 6 mostra,

Figura 5 – Representação gráfica do cromossomo considerado neste trabalho mostrando os 21 marcadores com distância de 3 centimorgans. Este grupo de marcadores pode ser considerado um subgrupo dentro de um grupo de ligação maior

esquemáticamente, o processo de obtenção dessas populações.

Figura 6 – Esquema do cruzamento para a obtenção dos genótipos de uma população F_2 considerando-se os locos hipotéticos A e B (ligados ou não)

Nesta simulação, considerou-se a ausência dos dois tipos de interferência cromossômica: *i*) a ausência da interferência de posição da permuta, ou seja, a ocorrência de uma permuta não interferiu na ocorrência de outra permuta próxima; *ii*) a ausência de interferência de cromátide, o que implica que cada par de cromátides não irmãs foi equiprovavelmente envolvido no processo de permuta, independente de seu envolvimento em outras permutas (SPEED; MCPHECK; EVANS, 1992).

A simulação dos gametas para a formação da geração F_2 foi baseada no fato de os indivíduos da geração F_1 , vistos na Figura 6, possuírem todos seus locos em heterozigose. O primeiro passo para a obtenção de um indivíduo da geração F_2 foi a simulação da recombinação nos indivíduos da geração F_1 . Para tanto, assumiu-se que o número de permutas p_i é uma variável aleatória com distribuição de Poisson $p_i = Po(\lambda)$, com λ igual ao comprimento do cromossomo em Morgans

(BASTEN et al., 2003):

$$p_i = Po(\lambda), \text{ sendo} \quad (45)$$

$$\lambda = \sum_i d_{M_i M_{i+1}}, i = 1, \dots, 20 \quad (46)$$

As permutas foram posicionadas no cromossomo de acordo com uma distribuição uniforme. Desse modo, todos os pontos do cromossomo tiveram a mesma probabilidade de receber uma permuta. Com esse procedimento, foram gerados dois gametas provenientes dos pares de cromossomos homólogos após a recombinação. Um deles foi tomado ao acaso para simulação do indivíduo da geração seguinte. Para a obtenção de um indivíduo da população F_2 , este procedimento foi realizado duas vezes. Este processo foi repetido 100 e 400 vezes, formando uma população com 100 indivíduos e 400 indivíduos, respectivamente. Com isso, foram simulados dois arquivos de dados: um referente à população de tamanho 100 e outro referente à população de tamanho 400. Vale ressaltar que, no contexto do mapeamento genético, o conjunto de indivíduos originários de um cruzamento específico e que são genotipados e utilizados para construção do mapa de ligação, é chamado de população experimental. Embora este conjunto de indivíduos seja chamado de população, eles representam na verdade apenas uma amostra dos inúmeros indivíduos diferentes que podem ser originados naquele cruzamento específico. Dessa forma, as populações experimentais aqui consideradas são amostras através das quais é realizado um processo de inferência estatística, baseado em diferentes métodos. No caso, as estimativas de interesse são a ordem dos marcadores e a distância entre eles.

3.1.2 Simulação dos Marcadores

Quatro padrões, combinando diferentes tipos de marcadores, foram considerados: *i*) padrão $CCCCC$; *ii*) padrão $CD CDC$; *iii*) padrão $CDDDC$; e *iv*) padrão $CDD_R DC$, de modo que C , denota marcador codominante; D , marcador dominante em fase de associação e D_R , marcador dominante em fase de repulsão. Cada uma destas combinações foram posicionadas sequencialmente até atingir os 21 marcadores, como mostra a Figura 7.

A princípio, o padrão 1 é o mais informativo, uma vez que todos os marcadores são codominantes. Os padrões 2 e 3 contêm marcadores dominantes em diferentes proporções. O padrão 4 representa um caso com vários problemas na prática, pois contém marcadores dominantes em fase de repulsão entre marcadores dominantes vizinhos. Com esses padrões, foi possível abordar vários

Figura 7 – Mapas simulados a partir dos padrões de marcadores *CCCCC*, *CDCDC*, *CDDDC* e *CDDR_RDC*. *C*: marcador codominante; *D*: marcador dominante em fase de associação; *D_R*: marcador dominante em fase de repulsão

casos de mistura de tipos de marcador que ocorrem na prática através da simulação. Espera-se dessa forma que seja possível fazer recomendações práticas aos geneticistas com base nessas simulações.

Para a simulação dos diferentes tipos de marcadores procedeu-se da seguinte forma: os dados originários da simulação da população F_2 apresentam três classes genotípicas. Esse tipo de dado já refere-se ao marcador codominante, uma vez que as três classes são distinguíveis. Para a simulação dos marcadores dominantes em fase de associação, os dados dos marcadores codominantes foram repetidos e a classe de homozigotos referentes ao genitor 1 (P_1) foi misturada com a classe de heterozigotos, recebendo uma nova codificação. Para a simulação dos marcadores dominantes em repulsão, a classe de homozigotos referentes ao genitor 2 (P_2) foi misturada com a classe de heterozigotos, formando também uma nova classe.

3.1.3 Simulação dos Dados Perdidos

Um dos problemas bastante enfrentados por geneticistas na prática é a perda de dados. Existem diversas razões para a ocorrência dessa perda. Uma delas é a impossibilidade de leitura dos dados no gel, muitas vezes por falha nas reações de amplificação, o que dificulta a leitura das bandas. Tendo em vista este problema, foram simulados outros arquivos de dados retirando-se aleatoriamente 10% e 20% dos dados. Assim, para cada um dos arquivos simulados com os dados completos, foram produzidos outros dois com dados perdidos. Considerando-se todas estas situações, foram simulados 24 tipos de populações experimentais mostradas na Tabela 1.

Tabela 1 – Populações experimentais F_2 simuladas, combinando diferentes tamanhos populacionais, padrões de marcadores e dados perdidos

Tamanho da População	Padrão dos Marcadores	Dados Perdidos (%)
100	$CCCCC$	0
		10
		20
	$CD CDC$	0
		10
		20
	$CDDDC$	0
		10
		20
	CDD_RDC	0
		10
		20
400	$CCCCC$	0
		10
		20
	$CD CDC$	0
		10
		20
	$CDDDC$	0
		10
		20
	CDD_RDC	0
		10
		20

3.1.4 Simulação das Amostras Monte Carlo

Os métodos Monte Carlo geralmente envolvem a geração de amostras assumindo-se um determinado modelo, e a utilização das amostras obtidas para subsequente análise com a metodologia de interesse (MANLY, 2003). Nos tópicos anteriores, assumiu-se um modelo para a simulação das populações experimentais, o qual foi utilizado para a simulação das amostras Monte Carlo. Para a simulação das amostras Monte Carlo, cada uma das populações apresentadas na Tabela 1 foi simulada 550 vezes, gerando 13200 (550×24) populações. Como cada simulação envolve a realização de amostras com base em distribuições de probabilidade, espera-se, teoricamente, que cada amostra seja diferente das anteriores (assumindo independência). A avaliação dos algoritmos para ordenação dos locos pode então ser feita aplicando-os a cada amostra, que tecnicamente equivale a repetições. Todas as simulações deste trabalho foram realizadas no ambiente R, que é uma linguagem para análise de dados e construção de gráficos (IHAKA; GENTLEMAN, 1996).

3.2 Métodos

3.2.1 Obtenção das matrizes R e LOD

As matrizes **R** (eq. 22) e **LOD** (eq. 23) são fundamentais para emprego dos diversos algoritmos e critérios avaliados neste trabalho. Tais matrizes foram obtidas uma única vez para cada uma das 13200 populações analisadas, e tornadas disponíveis para os algoritmos e/ou critérios que necessitaram dessa informação. Para a obtenção da matriz **R** de frações de recombinação, foi usado o algoritmo EM como apresentado por Liu (1998). De forma sucinta, através do algoritmo EM é possível obter estimativas de máxima verossimilhança de modelos probabilísticos, nos casos em que este modelo depende de variáveis não observadas ou latentes. No caso das frações de recombinação, as variáveis latentes são proporções da mistura de classes genotípicas que não podem ser distinguidas (por exemplo: M_{-}). Este algoritmo é constituído por dois passos: *i*) passo E (*expectation*) que consiste no cálculo da esperança da verossimilhança incluindo as variáveis latentes como se fossem observadas; *ii*) passo M (*maximization*), que consiste no cálculo da estimativa de máxima verossimilhança dos parâmetros, maximizando a verossimilhança esperada encontrada no passo E. Assim, os parâmetros encontrados no passo M são usados no passo E até que ocorra convergência.

A matriz **LOD** (eq. 23), foi obtida através da seguinte equação (LIU, 1998):

$$lod_{M_i M_j} = \log_{10} \left[\frac{L(\hat{r}_{M_i M_j})}{L(r_{M_i M_j} = 1/2)} \right] \quad (47)$$

sendo $L(\hat{r}_{M_i M_j})$ a função de verossimilhança para a estimativa de ligação entre os marcadores M_i e M_j , $\hat{r}_{M_i M_j}$ a estimativa de máxima verossimilhança da fração de recombinação entre os locos M_i e M_j , e $L(r_{M_i M_j} = 1/2)$ a função de verossimilhança considerando ausência de ligação entre os locos M_i e M_j ($H_0 : r_{M_i M_j} = 1/2$)

Dessa forma, para cada uma das 13200 populações simuladas foram obtidas duas matrizes de dimensão 20×20 (número de intervalos entre os 21 marcadores simulados). As matrizes **R** e **LOD** foram calculadas usando um programa específico escrito em linguagem C (KERNIGHAN; RITCHIE, 1990).

3.2.2 Algoritmos de Ordenação

Os algoritmos de ordenação avaliados no presente trabalho foram: *try* (TRY) (LANDER et al., 1987), *seriation* (SER) (BUETOW; CHAKRAVARTI, 1987a), *rapid chain delineation* RCD (DOERGE, 1996), *recombination count and ordering* (RECORD) (VAN-OS et al., 2005) e *unidirectional growth* (UG) (TAN; FU, 2006). Estes algoritmos foram escolhidos por serem muito usados na prática, como é o caso dos algoritmos TRY, SER e RCD, ou por terem surgido recentemente na literatura, como é o caso dos algoritmos RECORD e UG, sobre os quais ainda não foram realizados maiores estudos, e propõem-se a realizar a ordenção dos casos complexos disponíveis atualmente. O algoritmo TRY é um dos mais utilizados na prática para a construção de mapas genéticos. Até o momento da elaboração deste trabalho, o programa MAPMAKER/EXP, que implementa o algoritmo TRY, foi citado 2860 vezes em publicações científicas (SCHOLAR, 2007). Isso mostra claramente a popularidade deste algoritmo e software para a construção de mapas.

Cada um dos algoritmos citados possui diferentes abordagens para a construção dos mapas genéticos. No algoritmo TRY, as frações de recombinação são obtidas através de estimativas multi-ponto de máxima verossimilhança via cadeia de Markov oculta. Essas estimativas são usadas para o cálculo da verossimilhança de uma dada ordem, que serve como critério para avaliação da mesma. Os demais algoritmos utilizam a matriz **R** de frações de recombinação, sendo conseqüentemente mais rápidos do ponto de vista computacional. O algoritmo SER baseia-se na monotonicidade dessa matriz para escolher a melhor ordem. O algoritmo RECORD fundamenta-se no número de eventos de recombinação, de modo a minimizá-lo. O RCD baseia-se na adição seqüencial de marcadores

levando em conta a matriz **R**. Finalmente, o UG tem como fundamento a adição seqüencial e unidirecional de marcadores baseando-se em propriedades aditivas das distâncias entre os marcadores.

Uma vez que estes algoritmos foram avaliados usando o método Monte Carlo, foi necessário que os mesmos pudessem ser executados automaticamente para as 13200 amostras Monte Carlo simuladas, como visto no item 3.1.4. Nem todos os programas computacionais disponíveis permitem que a construção do mapa seja feita de forma automática, e nem todos os algoritmos e critérios apresentados neste trabalho foram implementados em softwares, visto que são muito recentes na literatura. Dessa forma, optou-se por implementar todos os algoritmos avaliados neste trabalho utilizando o ambiente R (IHAKA; GENTLEMAN, 1996) e a linguagem de programação C (KERNIGHAN; RITCHIE, 1990)

A linguagem C é extremamente poderosa em termos de processamento computacional. O ambiente R permite que funções escritas em linguagem C sejam utilizadas dentro dos programas desenvolvidos nesse ambiente. Dessa forma, foi possível aliar as facilidades de análises estatísticas e gráficos do ambiente R com a velocidade de processamento computacional da linguagem C.

3.2.2.1 Algoritmo TRY

Como apresentado no item 2.4.1, foi necessária a obtenção de uma subordem inicial para a execução do algoritmo TRY. Esta subordem inicial foi obtida comparando-se a verossimilhança calculada via cadeia de Markov oculta das $m'!/2$ ordens. Claramente, quanto maior for m' , melhor será o funcionamento do algoritmo TRY. Todavia, este é um processo muito intensivo em termos computacionais, à medida que m' aumenta. Além do TSP, quanto maior o número de indivíduos, mais cálculos são necessários para a obtenção da verossimilhança, como pode ser visto na eq. 17. Então, optou-se por utilizar $m' = 5$, o que resulta em 60 sub-ordens. Para cada uma delas, o mapa foi reconstruído e foi calculada a verossimilhança via cadeia de Markov oculta, como mostrado no item 2.2. Devido a esta limitação e ao TSP, números superiores a esse inviabilizariam a análise baseada no método Monte Carlo, mesmo com a grande quantidade de recursos computacionais utilizados (ver item 3.2.5).

3.2.2.2 Algoritmos SER, RCD, RECORD e UG

Os algoritmos SER, RCD, RECORD e UG têm em comum o fato de utilizarem a fração de recombinação obtida através da análise das marcas duas a duas, ou seja, a matriz **R**, de modo

que este foi o único argumento necessário para a execução dos mesmos, exceto do RECORD, que também necessitou do número de indivíduos para o cálculo do número esperado de recombinações.

3.2.2.3 Funções de Mapeamento

Em todos os algoritmos, após a construção do mapa, as frações de recombinação obtidas entre os marcadores ordenados foram transformadas em distâncias (d_{ij}) usando a função de mapeamento de Haldane (1919), mostrada na eq. 1. Esta função foi escolhida, pois não leva em conta o fenômeno da interferência, concordando com as simulações realizadas no presente trabalho. Valores de frações de recombinação maiores que 0,5, embora sejam raros, podem ocorrer. Para que as distâncias não resultassem em valores infinitos, frações de recombinação maiores que 0,5 foram substituídos por 0,49999, como sugerido por Olson e Boehnke (1990) para evitar problemas numéricos.

3.2.3 Algoritmo RIPPLE e Critérios para a Avaliação de Ordens

Os critérios considerados no presente trabalho foram: 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), soma máxima dos LOD scores adjacentes (SALOD - *sum of adjacent lod scores*) (WEEKS; LANGE, 1987) e verossimilhança via cadeias de Markov ocultas (LHMC - *likelihood via hidden Markov chains*) (LANDER; GREEN, 1987). Estes critérios também foram escolhidos por serem muito usados na prática e estarem implementados em softwares. Todos eles foram avaliados juntamente com o algoritmo RIPPLE. Como visto no item 2.5, o algoritmo RIPPLE tem como finalidade procurar por ordens mais prováveis, em mapas previamente construídos, através da permutação de m' marcadores vizinhos.

Para o critério LHMC, foi usado janela com $m' = 4$. Números maiores inviabilizariam as análises Monte Carlo, pois a obtenção do LHMC envolve cálculos complexos e intensivos em termos computacionais. Já no caso dos critérios SARF, PARF e SALOD, foi usado $m' = 6$, pois estes critérios são obtidos através de cálculos simples e rápidos. Assim como os algoritmos, os critérios SARF, PARF e SALOD, e o algoritmo RIPPLE, foram todos implementados no ambiente R (IHAKA; GENTLEMAN, 1996); o critério LHMC foi implementado em linguagem C (KERNIGHAN; RITCHIE, 1990), por ser computacionalmente muito intensivo.

Para o cálculo da verossimilhança via cadeia de Markov oculta (LHMC), foram usados os

códigos-fonte em linguagem C (KERNIGHAN; RITCHIE, 1990) da função **est.map**. Esta função faz parte do pacote computacional livre **R/qtl** (BROMAN, 2007). Este pacote é destinado ao mapeamento de QTLs e foi desenvolvido para o ambiente R. Como no presente trabalho foram usadas apenas populações experimentais F_2 (Figura 6), os códigos originais foram adaptados para este caso específico.

Para obtenção dos critérios SARF, PARF e SALOD foram usadas as matrizes **R** e **LOD**, previamente calculadas. Para evitar problemas numéricos para valores de PARFs iguais a zero, valores de fração de recombinação $\hat{r}_{M_i M_j} = 0$ foram substituídos por $\hat{r}_{M_i M_j} = 0,00001$, como sugerido por (OLSON; BOEHNKE, 1990). Foram ainda utilizados logaritmos para transformar o produto das frações de recombinação em soma, diminuindo assim o tempo de processamento computacional no cálculo do PARF.

Cada uma das 13200 populações simuladas teve seu mapa construído usando os algoritmos TRY, SER, RCD, RECORD e UG. Em cada um desses mapas foi aplicado o algoritmo RIPPLE com os critérios LHMC, SARF, PARF e SALOD, resultando em 25 combinações apresentadas na Tabela 2. Para cada conjunto de 24 populações simuladas foram construídos mapas genéticos usando as 25 combinações de algoritmos e RIPPLE com critérios considerados neste trabalho. Dessa forma, para cada amostra Monte Carlo foram construídos 600 mapas, que representam 600 situações diferentes na prática. Com o uso da estratégia Monte Carlo, cada uma dessas situações foram repetidas 550 vezes. Portanto, ao final das análises foram obtidos 330.000 mapas genéticos.

3.2.4 Avaliação dos Algoritmos e Critérios

3.2.4.1 Coeficiente de Correlação de Spearman

No contexto do presente trabalho, para verificar se a ordem obtida aproxima-se da ordem real simulada, foi utilizado o valor absoluto do coeficiente de correlação ordinal, ou coeficiente de Spearman (1902). O coeficiente de correlação de Spearman é uma medida de correlação não-paramétrica, que leva em conta a ordem das observações. Foi utilizado o valor absoluto do coeficiente, pois ordens invertidas também são ordens corretas no contexto dos mapas genéticos. O coeficiente de correlação de Spearman é dado pela fórmula

$$\rho = 1 - \frac{6 \sum (i - i^*)^2}{m(m^2 - 1)}, i = 1, \dots, m, \text{ sendo} \quad (48)$$

Tabela 2 – Combinações de algoritmos e métodos usados para construção dos mapas simulados. Os mapas obtidos com os algoritmos TRY, SER, RCD, RECORD e UG foram reavaliados usando-se o algoritmo RIPPLE com todos os critérios abordados neste trabalho

Algoritmo	Critério
TRY	Sem RIPPLE
	LHMC + RIPPLE
	SARF + RIPPLE
	PARF + RIPPLE
	SALOD + RIPPLE
SER	Sem RIPPLE
	LHMC + RIPPLE
	SARF + RIPPLE
	PARF + RIPPLE
	SALOD + RIPPLE
RCD	Sem RIPPLE
	LHMC + RIPPLE
	SARF + RIPPLE
	PARF + RIPPLE
	SALOD + RIPPLE
RECORD	Sem RIPPLE
	LHMC + RIPPLE
	SARF + RIPPLE
	PARF + RIPPLE
	SALOD + RIPPLE
UG	Sem RIPPLE
	LHMC + RIPPLE
	SARF + RIPPLE
	PARF + RIPPLE
	SALOD + RIPPLE

ρ o coeficiente de correlação entre as ordens dos mapas estimado e real, i a posição (“rank”) do marcador M_i no mapa construído e i^* a posição (“rank”) do marcador M_i^* no mapa real.

Para cada um dos 330.000 mapas construídos a partir das amostras simuladas, foi calculado o coeficiente de correlação com o mapa real simulado. Logo, mapas que apresentaram ordens idênticas ao mapa real tiveram valor absoluto do coeficiente de correlação $|\rho| = 1$. Para visualização dos coeficientes, foram construídos boxplots com as 550 amostras Monte Carlo, correspondendo às 600 situações de mapeamento abordadas neste trabalho.

3.2.4.2 Distância Euclidiana

Um outro ponto a ser considerado na comparação dos algoritmos e critérios usados diz respeito às distâncias obtidas entre os marcadores. Assim, as distâncias estimadas entre os marcadores também foram comparadas com as distâncias reais. Isso foi feito com o cálculo da distância euclidiana média entre o vetor de distâncias do mapa construído e o vetor de distâncias do mapa real. Para tanto, foi usada a equação

$$D = \left[\frac{1}{m} \sum_{i=1}^{m-1} \left(\hat{d}_{M_i M_{i+1}} - d_{M_i M_{i+1}} \right)^2 \right]^{1/2}, \quad (49)$$

sendo D a distância euclidiana média entre as distâncias do mapa estimado e do mapa real, $\hat{d}_{M_i M_{i+1}}$ a distância estimada entre os marcadores M_i e M_{i+1} , $d_{M_i M_{i+1}}$ a distância real entre os marcadores M_i e M_{i+1} e m o número de marcadores.

O vetor de distâncias do mapa construído foi obtido levando em conta a ordem real, ou seja, $M_1, M_2, \dots, M_{20}, M_{21}$. Isso foi feito somando-se as distâncias entre os marcadores no mapa construído, uma vez que as distâncias são aditivas. Do mesmo modo que no item 3.2.4.1, para cada um dos 330.000 mapas construídos foi calculada a distância euclidiana média. Dessa forma, mapas que apresentaram distâncias idênticas ao mapa real tiveram distância euclidiana média $D = 0$.

3.2.5 Recursos Computacionais Utilizados

Todas as análises necessárias para a realização do presente trabalho foram executadas nos supercomputadores do Centro Nacional de Processamento de Alto Desempenho da Universidade Estadual de Campinas (CENAPAD - UNICAMP - <http://www.cenapad.unicamp.br/>) e nos computadores do setor de informática do Departamento de Genética da Escola Superior de Agricultura “Luiz de

Queiroz” da Universidade de São Paulo (LGN/ESALQ/USP).

No CENAPAD foram utilizadas 4 CPUs Intel Itanium2 do sistema SGI Altix 1350 com 274 GB de memória RAM. A capacidade de processamento teórica do sistema utilizado é de 420 GFLOPS (*giga floating point operations per second*), sendo este o único no cenário nacional de processamento de alto desempenho. No Departamento de Genética (ESALQ/USP) foram usados dois CPUs Intel Xeon de 1.6 GHz, cada um com dois núcleos de processamento e memória RAM de 1GB. Usando estes recursos computacionais, o tempo total de processamento para a realização das análises foi de aproximadamente 3 meses e 15 dias.

4 RESULTADOS E DISCUSSÃO

4.1 Estimativas da Ordem dos Marcadores

As distribuições dos coeficientes de correlação entre as ordens dos mapas estimados e a ordem do mapa real, com base no método Monte Carlo, são mostradas nas Figuras 8 e 9. Em cada figura são apresentados cinco gráficos correspondentes aos algoritmos avaliados no presente trabalho. Para cada uma das figuras, um código de cores indica combinações de padrões dos marcadores (*CCCCC*, *CDCDC*, *CDDDC* e *CDD_RDC*), quantidade de dados perdidos (0%, 10% e 20%) e critério usado com o algoritmo RIPPLE (Sem RIPPLE, LHMC+RIPPLE, SARF+RIPPLE, PARF+RIPPLE e SALOD+RIPPLE), totalizando 60 combinações. Cada uma das distribuições dos coeficientes de correlação das 550 amostras Monte Carlo foi representada em um boxplot.

Em ambas figuras, os boxplots foram ordenados primeiramente de acordo com os padrões dos marcadores, começando pelo padrão *CDD_RDC*, que é o que apresentou maior dificuldade de ordenação, conforme esperado, passando pelos padrões *CDDDC* e *CDCDC*, até o padrão *CCCCC*, que é o padrão que apresentou poucos problemas de ordenação, em função de apresentar apenas marcadores completamente informativos (codominantes). Dentro desses quatro grupos, os boxplots foram ordenados de acordo com a porcentagem de dados perdidos, indo de 20% até 0%. Finalmente, dentro desses últimos grupos, os boxplots foram ordenados de acordo com os critérios usados com o algoritmo RIPPLE, na seguinte ordem: Sem RIPPLE, SARF+RIPPLE, PARF+RIPPLE, SALOD+RIPPLE e LHMC+RIPPLE. Dessa forma, as regiões mais à esquerda dos gráficos representam casos mais difíceis de se obter bons mapas.

De modo geral, verifica-se que para as populações com 100 indivíduos, os algoritmos TRY e SER apresentaram valores de correlação mais elevados e concentrados em torno dos altos valores (menor amplitude inter-quartílica), mesmo nos casos mais complicados. Os algoritmos RCD, RECORD e UG não tiveram resultados tão bons, com correlações mais baixas e com maior amplitude inter-quartílica nos casos mais complicados. Nos casos mais simples, o algoritmo RECORD teve resultados muito semelhantes ao TRY e SER, e o UG e RCD mostraram desempenho inferior. Para as populações com 400 indivíduos, os algoritmos TRY, SER e RECORD foram bastante parecidos, apresentando os melhores resultados, com altos valores de correlação e pequena amplitude inter-quartílica. O algoritmo RCD foi claramente inferior, principalmente nos casos mais complica-

Figura 8 – Boxplots das distribuições dos coeficientes de correlação entre os mapas construídos e o mapa real, para população de $n = 100$ indivíduos. São apresentados cinco gráficos correspondentes aos algoritmos avaliados (*Try*, *seriation*, *Rapid Chain Delineation*, *Recombination Counting and Ordering* e *Unidirectional Growth*). No eixo horizontal, o diagrama de cores indica o padrão dos marcadores simulados, a quantidade de dados perdidos e o critério usado com o algoritmo RIPPLE

Figura 9 – Boxplots das distribuições dos coeficientes de correlação entre os mapas construídos e o mapa real, para população de $n = 400$ indivíduos. São apresentados cinco gráficos correspondentes aos algoritmos avaliados (*Try*, *seriation*, *Rapid Chain Delineation*, *recombination counting and ordering* e *Unidirectional Growth*). No eixo horizontal, o diagrama de cores indica o padrão dos marcadores simulados, a quantidade de dados perdidos e o critério usado com o algoritmo RIPPLE

dos, e o algoritmo UG foi superior ao RCD, mas ligeiramente inferior aos demais. Isso mostra que aumentos no tamanho populacional torna bem mais simples o problema de ordenar os marcadores.

Mais detalhes sobre os resultados da simulação são apresentados na Tabela 3, que mostra a porcentagem de amostras Monte Carlo cujos mapas apresentaram coeficientes de correlações superiores a 0,95 para os cinco algoritmos avaliados. Com base nesses resultados (Figuras 8 e 9, Tabela 3), nota-se que, como esperado, os casos mais difíceis para obtenção de bons mapas são aqueles que contém marcadores dominantes em fase de repulsão (padrão CDD_RDC), principalmente nas populações com 100 indivíduos. Isso ocorre porque para marcadores dominantes ligados em repulsão, a correta convergência no algoritmo EM para a estimativa de dois pontos da fração de recombinação, depende da presença de pelo menos uma dupla recombinação, ou seja, é necessário que ambos os pais tenham apresentado recombinação entre os dois marcadores em questão. Caso nenhum indivíduo se enquadre nessa situação, a estimativa de dois pontos converge para zero, mesmo que os marcadores tenham grande distância entre si. Claramente, esta é uma limitação desta abordagem que torna-se ainda maior quando o tamanho amostral é menor, visto que nesses casos as duplas recombinações tem maior probabilidade de não ser observadas simplesmente devido ao acaso. Possivelmente isso explique o fato dos algoritmos baseados exclusivamente em testes de dois pontos (RCD, RECORD e UG) apresentarem os piores resultados. Os métodos com abordagem mais geral (TRY e SER), do tipo multiponto, mostraram melhor desempenho. Para as populações com padrão CDD_RDC , é possível notar a tendência de diminuição nas correlações à medida que a porcentagem de dados perdidos aumenta, o que obviamente era esperado.

Com relação aos critérios, observou-se que SARF, PARF e SALOD, usados com o algoritmo RIPPLE, têm sempre piores resultados que o LHMC+RIPPLE, e sem a utilização do RIPPLE (exceto para o algoritmo RCD), com valores variando de 78,7% a 6,5%. Com a utilização de LHMC+RIPPLE e sem a utilização do RIPPLE, estes valores variaram de 96,7% a 24,7% (Tabela 3). Como o SARF, o PARF e o SALOD não apresentaram bom desempenho, a interpretação dos resultados será feita basicamente levando em conta os resultados do critério LHMC+RIPPLE, e também sem o RIPPLE. Com relação aos algoritmos, o TRY e SER apresentaram os melhores resultados, uma vez que as porcentagens de correlações acima de 0,95 variaram de 96,7% a 88,7% para o TRY e 93,3% a 70,9% para o SER, que podem ser considerados ótimos resultados. Os algoritmos RECORD e UG tiveram resultados intermediários, com porcentagens variando de 56,9% a 24,7% e de 64,0% a 42,5%, respectivamente. O algoritmo RCD apresentou os piores resultados,

Tabela 3 – Porcentagem de mapas que apresentaram coeficientes de correlação acima de 0,95 para os algoritmos *Try* (TRY), *Seriation* (SER), *Rapid Chain Delineation* (RCD), *Recombination Count and Ordering* (RECORD) e *Unidirectional Growth* (UG) e critérios *sum of adjacent recombination fractions* (SARF), *product of adjacent recombination fractions, sum of adjacent lod scores* (SALOD) e *likelihood via hidden Markov chains* (LHMC)

Algoritmos	Critérios	CCCCC			CDCDC			CDDDC			CDD _R DC		
		0%	10%	20%	0%	10%	20%	0%	10%	20%	0%	10%	20%
n = 100													
TRY	Sem RIPPLE	99,8	99,8	98,9	100,0	98,9	95,6	99,3	98,4	94,0	96,7	93,5	88,7
	SARF	99,8	99,8	99,1	100,0	98,9	95,8	99,3	98,5	94,5	78,7	67,8	60,0
	PARF	99,8	99,8	98,9	99,8	98,9	96,0	99,3	98,5	94,0	69,8	55,5	45,1
	SALOD	99,8	99,8	99,3	59,6	64,4	65,3	96,0	92,9	87,1	21,5	19,5	17,6
	LHMC	99,8	99,8	98,9	100,0	98,9	95,8	99,3	98,4	94,0	96,7	93,8	89,3
SER	Sem RIPPLE	100,0	100,0	99,8	100,0	99,8	99,5	99,8	99,8	99,8	93,3	84,4	72,9
	SARF	100,0	100,0	99,6	100,0	99,6	98,7	100,0	99,6	99,5	71,3	57,6	46,0
	PARF	100,0	100,0	99,6	100,0	99,5	97,6	99,8	99,3	98,0	53,8	42,2	31,5
	SALOD	100,0	100,0	99,5	62,4	68,4	69,5	97,3	92,2	89,8	22,0	17,6	14,7
	LHMC	100,0	100,0	99,8	100,0	99,8	99,5	99,8	99,8	99,5	93,3	84,2	70,9
RCD	Sem RIPPLE	99,8	99,3	96,5	98,2	92,4	83,8	99,5	93,8	85,5	4,7	2,0	0,2
	SARF	100,0	99,5	96,4	98,2	93,1	87,6	99,6	95,1	88,5	16,7	9,5	4,2
	PARF	99,8	99,5	96,2	98,5	94,0	85,8	99,6	95,1	87,6	14,0	6,5	2,9
	SALOD	100,0	99,5	96,4	59,5	61,8	59,3	96,4	87,5	81,3	1,8	2,4	0,9
	LHMC	99,8	99,3	96,5	98,4	92,2	84,5	99,5	94,0	85,6	5,3	2,2	0,4
RECORD	Sem RIPPLE	100,0	100,0	100,0	100,0	100,0	99,5	100,0	99,8	99,5	56,5	40,4	24,7
	SARF	100,0	100,0	100,0	100,0	100,0	99,8	100,0	99,8	99,5	55,6	38,7	26,2
	PARF	100,0	100,0	100,0	100,0	100,0	99,5	100,0	99,8	99,5	53,5	37,3	21,6
	SALOD	100,0	100,0	100,0	59,3	64,9	69,8	96,9	94,7	92,7	11,5	8,7	6,5
	LHMC	100,0	100,0	100,0	100,0	100,0	99,5	100,0	99,8	99,5	56,9	41,3	25,5
UG	Sem RIPPLE	99,6	98,9	97,6	98,4	97,6	96,2	98,2	98,0	95,6	64,0	52,9	43,1
	SARF	99,6	98,9	97,6	98,4	97,6	96,2	98,2	98,0	96,2	51,1	38,0	26,9
	PARF	99,6	98,9	97,6	98,4	97,6	96,0	98,2	98,0	96,4	39,5	28,4	20,7
	SALOD	99,6	98,9	97,6	58,2	64,4	66,4	95,3	92,9	88,9	17,8	11,5	11,6
	LHMC	99,6	98,9	97,6	98,4	97,6	96,2	98,2	98,0	95,6	63,1	52,5	42,5
n = 400													
TRY	Sem RIPPLE	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	99,5	99,3	99,8
	SARF	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	99,5	98,4	99,1
	PARF	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	98,4	96,9	97,1
	SALOD	100,0	100,0	100,0	52,7	54,0	56,2	100,0	100,0	99,1	3,8	5,5	9,3
	LHMC	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	99,5	99,3	99,8
SER	Sem RIPPLE	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	SARF	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	99,6	98,2	98,0
	PARF	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	96,2	92,5	90,0
	SALOD	100,0	100,0	100,0	52,5	55,3	58,0	100,0	100,0	99,1	8,5	11,1	15,3
	LHMC	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
RCD	Sem RIPPLE	100,0	100,0	100,0	100,0	100,0	99,8	100,0	100,0	99,8	44,0	29,6	18,0
	SARF	100,0	100,0	100,0	100,0	100,0	99,3	100,0	100,0	99,8	56,4	44,0	35,6
	PARF	100,0	100,0	100,0	100,0	100,0	99,8	100,0	100,0	99,6	51,8	39,3	29,8
	SALOD	100,0	100,0	100,0	52,7	54,0	56,2	100,0	99,8	97,8	8,2	6,5	6,0
	LHMC	100,0	100,0	100,0	100,0	100,0	99,8	100,0	100,0	99,8	42,4	30,4	18,9
RECORD	Sem RIPPLE	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	98,4	95,6	92,2
	SARF	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	98,4	95,6	92,4
	PARF	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	97,5	94,5	92,0
	SALOD	100,0	100,0	100,0	52,7	54,0	56,5	100,0	100,0	99,1	6,5	9,1	9,5
	LHMC	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	98,4	95,3	92,0
UG	Sem RIPPLE	100,0	100,0	99,8	99,8	100,0	99,8	100,0	99,8	99,6	98,7	96,7	93,8
	SARF	100,0	100,0	99,8	99,8	100,0	99,8	100,0	99,8	99,8	97,8	95,1	90,7
	PARF	100,0	100,0	99,8	99,8	100,0	99,8	100,0	99,8	99,6	92,2	88,9	82,4
	SALOD	100,0	100,0	99,8	52,4	54,4	56,5	100,0	99,8	98,7	13,6	13,6	13,8
	LHMC	100,0	100,0	99,8	99,8	100,0	99,8	100,0	99,8	99,6	98,7	96,4	93,8

com valores variando de 16,7% a 0,2% (considerando-se todos os critérios usados com o algoritmo RIPPLE). O algoritmo TRY, que usa a abordagem multiponto, normalmente fornece ordenações muito boas, mesmo sem o uso do RIPPLE. Nas simulações usando TRY e LHMC+RIPPLE, porcentagens de correlação acima de 0,95% foram 96,7%, 93,8% e 89,3% considerando-se 0%, 10% e 20% de dados perdidos. Sem o algoritmo RIPPLE, esses valores foram 96,7%, 93,5% e 88,7%. Com o algoritmo SER e critério LHMC+RIPPLE, as porcentagens foram 93,3%, 84,2 e 70,9% para 0%, 10% e 20% e sem o algoritmo RIPPLE esses valores foram 93,3%, 84,4% e 72,9%. Assim, o algoritmo TRY pode ser considerado ligeiramente superior.

Para populações com 400 indivíduos e padrão de marcadores CDD_RDC , o critério SALOD usado com o algoritmo RIPPLE apresentou resultados inferiores, com valores de porcentagem de correlação acima de 0,95 variando de 15,3% a 3,8%. Para os demais critérios aplicados com o algoritmo RIPPLE ou mesmo sem a utilização desse algoritmo, estes valores variaram de 100% a 29,6%. Dessa forma a interpretação dos resultados será feita excluindo-se o critério SALOD. Com relação aos algoritmos de ordenação, TRY, SER, RECORD e UG foram obtidos bons resultados, com valores de porcentagem de correlação acima de 0,95 variando de 99,8% a 96,9% para o TRY, de 100,0% a 90,0% para o SER, de 98,4% a 92,0% para o RECORD e 98,7% a 82,4% para o UG. O algoritmo RCD não apresentou resultados satisfatórios, mesmo quando combinado com o RIPPLE com vários critérios para verificação de erros, com valores variando de 56,4% a 18,0%. Isso mostra que aumentos no tamanho da amostra facilita a ordenação dos marcadores, sendo que praticamente qualquer método poderia ser usado.

Nos casos em que não existem marcadores dominantes em repulsão (padrões $CCCCC$, $CDCDC$ e $CDDDC$), todos os algoritmos apresentaram bons resultados (Figuras 8 e 9, Tabela 3). Para os dois tamanhos populacionais simulados ($n = 100$ e $n = 400$), o critério SALOD usado com o algoritmo RIPPLE no padrão $CDCDC$, apresentou resultados muito inferiores aos demais (variando de 69,8% a 52,4%, Tabela 3). Para os outros critérios, estes valores foram superiores, variando de 100% a 83,8%. Nas Figuras 8 e 9 nota-se também uma ligeira diminuição da média das correlações com o critério SALOD+RIPPLE para o padrão $CDDDC$. Dessa forma as comparações serão feitas desconsiderando-se o SALOD, que não podem ser recomendados em situações práticas. No que refere aos algoritmos de ordenação, para populações com $n = 100$ indivíduos, destacam-se o TRY (100,0% a 94,0%), SER (100,0% a 97,6%), RECORD (100% a 99,5%) e UG (99,6% a 95,6%). O algoritmo RCD apresentou uma pequena desvantagem em relação aos demais, com

valores de porcentagem de coeficientes de correlação acima de 0,95 variando de 100,0% a 83,8%. De modo geral, para os padrões *CCCCC*, *CDCDC* e *CDDDC*, o algoritmo RECORD apresentou apenas 2,8% das correlações abaixo de 0,95; SER, TRY, UG e RCD apresentaram 3,0%, 4,3%, 4,8% e 7,7%, respectivamente. Entretanto, como pode ser visto na Figura 8, estas correlações abaixo de 0,95, apresentam diferentes variabilidades para cada algoritmo, sendo que o algoritmo UG apresentou maior número de valores próximos de zero. Nas populações com $n = 400$ todos os algoritmos apresentaram ótimos resultados, de modo que as porcentagens variaram de 100% a 99,3% (desconsiderando-se novamente o SALOD).

Algumas razões podem ser apontadas para o desempenho ruim do SALOD na presença de mistura de tipos de marcadores com mesma fase de ligação (padrões *CDCDC* e *CDDDC*). Como apontado por Olson e Boehnke (1990), o critério SALOD é sensível a diferenças na quantidade de informação dos marcadores, não sendo portanto um critério aconselhável quando há mistura de marcadores dominantes e codominantes. Isso ocorre pois o SALOD ignora o fato de que um número diferente de meioses informativas pode contribuir para o valor do LOD, de modo que os marcadores mais informativos têm a tendência de apresentar valores mais elevados de LOD quando comparados com os menos informativos. Dessa forma, se houver mistura de marcadores, o critério SALOD favorece o agrupamento de marcadores mais informativos (OLSON; BOEHNKE, 1990). Uma alternativa proposta por esses autores é a utilização dos critérios SALEQ e SALPIC. O critério SALEQ pondera o LOD pelo número de meioses informativas, e o critério SALPIC leva em conta o conteúdo de informação de polimorfismo de cada marcador considerado. Estes critérios deveriam ser melhores investigados em trabalhos futuros.

Alguns problemas decorrentes do uso do SALOD podem ser visualizados graficamente, como nos exemplos das Figuras 10 e 11, que mostram alguns resultados da ordenação de marcadores nas populações com padrões *CDCDC* e *CDDDC* para $n = 100$ e $n = 400$ indivíduos, e com 20% de dados perdidos. Cada gráfico contém 550 linhas representando os mapas construídos a partir das amostras Monte Carlo e 21 colunas representando os marcadores. Comparando os gráficos dos mapas construídos apenas com os algoritmos de ordenação, com os mapas construídos usando SALOD+RIPPLE, verifica-se claramente que ocorrem problemas na ordenação, tanto nos mapas obtidos a partir das populações com $n = 100$, quanto nos mapas obtidos a partir das populações com $n = 400$ para os dois padrões de marcadores. Entretanto, verifica-se que estes problemas são menores no padrão *CDDDC*. Isso ocorre pois nos mapas com este padrão, a proporção

de marcadores codominantes é menor e, como o SALOD favorece o agrupamento de marcadores mais informativos, a tendência para a formação de grupos contendo predominantemente marcadores codominantes é reduzida neste caso.

4.2 Estimativas das Distâncias entre os Marcadores

Para que os algoritmos e critérios possam ser avaliados, é importante também que as estimativas das distâncias entre os locos sejam consideradas. Na Figura 12 são apresentados resultados das distâncias euclidianas médias, calculadas entre as distâncias reais e as distâncias estimadas, usando os algoritmos TRY, SER, RCD, RECORD e UG. A distância euclidiana média pode ser usada para comparar as distâncias entre os marcadores obtidas a partir dos mapas de ligação estimados, com as distâncias reais. Por exemplo, valor de 1 cM para distância euclidiana média indica que os mapas diferem 1cM entre si, em média, para cada intervalo.

Verifica-se que, para populações apenas com marcadores codominantes (padrão *CCCCC*), tanto com $n = 100$ como com $n = 400$, todos os algoritmos e critérios forneceram distâncias euclidianas médias próximas de zero, indicando que há muita proximidade entre os valores reais e os estimados. Neste caso, as médias das distâncias euclidianas variaram de 0,29 a 0,41 cM para $n = 100$ e de 0,14 a 0,18 cM para $n = 400$.

Para populações com padrões de marcadores *CDCDC* e *CDDDC* e $n = 100$, os algoritmos tiveram comportamento semelhante e mostraram bons resultados. De forma geral, as distâncias aumentaram com o aumento da quantidade de dados perdidos. Para o algoritmo TRY, as médias das distâncias euclidianas variaram de 0,39 a 0,71 cM; para o algoritmo SER, variaram de 0,40 a 0,64 cM; para o algoritmo RCD, variaram de 0,39 a 0,66 cM; para o algoritmo RECORD, variaram de 0,38 a 0,64 cM; e finalmente para o algoritmo UG, as médias das distâncias variaram de 0,40 a 0,72 cM. Neste caso, a média das distâncias para todos algoritmos de ordenação sem a utilização do algoritmo RIPPLE foi 0,51 cM. Usando o algoritmo RIPPLE com o critério SARF, a média das distâncias considerando todos algoritmos foi 0,50 cM; com o critério PARF, a média foi 0,51 cM; com o critério SALOD, a média foi 0,56 cM; e com o critério LHMC a média foi de 0,48 cM. Para as populações com $n = 400$ e mesmo padrão de marcadores, as distâncias foram bastante similares àquelas obtidas apenas com marcadores codominantes.

Nas populações com padrões de marcadores *CDD_RDC* e $n = 100$, novamente todos algoritmos apresentaram resultados muito semelhantes. As médias das distâncias variaram de 0,50 a 1,60 cM

Figura 10 – Representação gráfica das ordens dos marcadores obtidas com os algoritmos *Try* (TRY), *Seriation* (SER), *Rapid Chain Delineation* (RCD), *Recombination Count and Ordering* (RECORD) e *Unidirectional Growth* (UG) em 550 amostras Monte Carlo, padrão de marcadores *CDCDC* e 20% de dados perdidos. Quanto mais a ordem dos mapas se aproximar da escala apresentada à direita, melhor é a sua ordenação

Figura 11 – Representação gráfica das ordens dos marcadores obtidas com os algoritmos *Try* (TRY), *Seriation* (SER), *Rapid Chain Delineation* (RCD), *Recombination Count and Ordering* (RECORD) e *Unidirectional Growth* (UG) em 550 amostras Monte Carlo, padrão de marcadores *CDDDC* e 20% de dados perdidos. Quanto mais a ordem dos mapas se aproximar da escala apresentada à direita, melhor é a sua ordenação

Figura 12 – Distâncias euclidianas médias entre as distâncias reais e as distâncias estimadas para as amostras Monte Carlo usado os algoritmos *Try* (TRY), *Seriation* (SER), *Rapid Chain Delineation* (RCD), *Recombination Count and Ordering* (RECORD) e *Unidirectional Growth* (UG). As médias das distâncias euclidianas calculadas são representadas por pontos cheios e a amplitude entre o 5^o e o 95^o percentil (intervalo de confiança) é representada por uma linha

para o algoritmo TRY, de 0,70 a 1,81 cM para o SER, de 0,94 a 2,45 cM para o RCD, de 0,78 a 2,09 cM para o RECORD, e por fim, de 0,83 a 2,10 cM para o UG. De modo geral observa-se que os critérios LHMC e SALOD usados com o algoritmo RIPPLE, apresentam menores intervalos, com valores de limite inferior variando de 0,38 a 0,64 cM para o LHMC+RIPPLE e 0,51 a 0,63 cM para o SALOD+RIPPLE, e limites superiores variando de 1,29 a 2,96 cM para o LHMC+RIPPLE e 1,62 a 2,84 cM para o SALOD+RIPPLE. Para os critérios SARF e PARF, esses valores variaram de 0,73 a 1,35 (limite inferior) e de 2,79 a 3,93 (limite superior). No caso do algoritmo TRY, as distâncias obtidas sem a verificação do RIPPLE foram tão boas quanto aquelas obtidas utilizando-se o critério LHMC+RIPPLE, confirmando que o TRY é um algoritmo que mostra bons resultados mesmo sem o uso do RIPPLE, por ser baseado em análises multiponto.

Nas populações com $n = 400$ e padrões de marcadores CDD_RDC , as médias e as amplitudes das distâncias euclidianas foram menores que aquelas verificadas nas populações com $n = 100$, indicando melhores estimativas, conforme esperado. O comportamento dos algoritmos e critérios foi bastante similar. A média das distâncias usando o algoritmo TRY, variou de 0,25 a 0,84; usando o SER, variou de 0,31 a 0,89; usando o RCD, variou de 0,36 a 1,28; usando o RECORD, variou de 0,32 a 0,88; e finalmente com o UG, variou de 0,32 a 0,96. Observa-se ainda que os dados perdidos tiveram pouca influência na obtenção das estimativas das distâncias entre os marcadores nas populações com $n = 400$ indivíduos. Considerando-se o algoritmo TRY, verifica-se que os critérios PARF+RIPPLE e SARF+RIPPLE apresentaram resultados inferiores aos obtidos com o critério LHMC+RIPPLE, não ocorrendo sobreposição dos intervalos (0,18 a 0,39 para o LHMC+RIPPLE; 0,45 a 1,42 para o SARF; e 0,49 a 1,48 para o PARF).

4.3 Considerações Gerais

O objetivo desse trabalho foi avaliar a eficiência dos algoritmos TRY, SER, RCD, RECORD e UG, além dos critérios PARF, SARF, SALOD e LMHC, usados juntamente com o algoritmo RIPPLE, na construção de mapas genéticos. Usando simulações Monte Carlo, foram estudadas todas as combinações entre algoritmos e critérios, com tamanhos populacionais de 100 e 400 indivíduos e 0%, 10%, e 20% de dados perdidos.

Para amostras com $n = 100$ indivíduos, na presença de marcadores dominantes em repulsão, com base nos resultados obtidos, é possível recomendar de forma geral o uso dos algoritmos TRY e SER combinados com o critério LHMC+RIPPLE, visto que eles mostraram os melhores resultados.

Na presença apenas de marcadores dominantes em fase de associação, podem ser recomendados os algoritmos TRY, SER, RECORD e UG, que tiveram comportamento muito semelhantes com uma ligeira vantagem para o RECORD e para o SER, seguidos do TRY e UG. O RCD não apresentou bons resultados, não sendo portanto recomendado em situações práticas. Recomenda-se ainda o uso do algoritmo RIPPLE com o critério LHMC, que sempre teve bons resultados. Na presença apenas de marcadores codominantes, todos algoritmos e critérios estudados neste trabalho tiveram ótimos resultados.

Considerando $n = 400$ indivíduos, na presença de marcadores dominantes em repulsão, os algoritmos TRY, SER, RECORD e UG forneceram bons resultados, podendo ser recomendados. Contudo, os algoritmos TRY e SER tiveram resultados ligeiramente superiores. O RCD apresentou resultados inferiores aos demais algoritmos, sendo o único não recomendado neste caso. O critério SALOD para o RIPPLE foi o único que apresentou resultados ruins, não sendo recomendado nesta situação. Na presença apenas de marcadores dominantes em fase de associação, todos algoritmos e critérios estudados apresentaram bons resultados, exceto o critério SALOD para o RIPPLE. Finalmente, na presença apenas de marcadores codominantes, todos algoritmos e critérios tiveram bons resultados.

Com relação às distâncias estimadas, de modo geral, na presença de marcadores dominantes em repulsão e $n = 100$ indivíduos, todos algoritmos tiveram comportamento semelhante. Entretanto a combinação do algoritmo TRY com o critério LHMC para o RIPPLE apresentou resultados ligeiramente superiores, sendo portanto a mais recomendada. Para $n = 400$ não é recomendada a combinação do algoritmo TRY com os critérios SARF+RIPPLE ou PARF+RIPPLE. Nos casos em que marcadores dominantes em repulsão não estão presentes, todos os algoritmos e critérios apresentaram ótimas estimativas de distância para os dois tamanhos populacionais considerados.

Como destacado por Lander et. al. (1987), estimativas multipontos das frações de recombinação são preferíveis às estimativas baseadas em combinações de marcadores dois a dois (estimativas de dois pontos), uma vez que elas contornam a propagação de erros ocasionados devido à falta de informação entre algumas combinações de marcadores. Essa vantagem é particularmente importante quando existe mistura de marcadores dominantes e codominantes, pois a cadeia de Markov usada para obter estimativas multiponto, permite que marcadores mais informativos forneçam informações para os demais, em função de sua proximidade, compensando assim a falta de informação. Este fato é de extrema importância na saturação de mapas genéticos, visto que uma estratégia que

poderia ser usada na prática seria utilização de marcadores dominantes, que normalmente podem ser obtidos em grande número com relativa facilidade (por exemplo, AFLPs, DArTs), para saturar mapas que já contém marcadores codominantes, que formariam um núcleo a partir do qual seria compensada a falta de informação dos marcadores dominantes. Isso está em consonância com os resultados aqui apresentados, uma vez que o algoritmo TRY utiliza a verossimilhança obtida via estimativa multiponto com cadeia de Markov oculta como critério interno para avaliação de ordens, isso levou à obtenção de melhores resultados.

O principal problema enfrentado para utilizar métodos baseados em cadeias de Markov, no contexto do mapeamento genético, refere-se ao tempo de processamento computacional, visto que atualmente muitos marcadores estão disponíveis e o tempo computacional necessário para fazer as análises pode ser proibitivo. Este fato motiva o surgimento de algoritmos como o SER, RCD, RECORD e UG. Os resultados aqui obtidos mostram que a melhor alternativa ao uso do TRY seria o uso do algoritmo SER, que também mostrou bons resultados e não é computacionalmente intensivo. O SER utiliza como critério de seleção da melhor ordem a monotonicidade da matriz de fração de recombinação, construída a partir de m ordens obtidas num passo anterior do algoritmo. Embora a estimativa das frações de recombinação não sejam obtidas usando a informação de todos os marcadores ao mesmo tempo, como ocorre na cadeia de Markov oculta, a avaliação das ordens nesse algoritmo é baseada na ordem de todos os marcadores ao mesmo tempo, uma vez que o índice de continuidade (CI), usado para verificar a monotonicidade, leva em conta a matriz como um todo, e não apenas a informação de cada distância de dois pontos separadamente. Isso possivelmente fez com que tal algoritmo também mostrasse bom desempenho mesmo em situações complexas, ao contrário dos demais. Nota-se, portanto, que os algoritmos TRY e SER tem em comum o fato de avaliar ordens ótimas globais, ou seja, levando em conta todos marcadores para verificar se a ordem estabelecida é melhor ou pior que uma ordem alternativa. Isso pode explicar as vantagens que esses algoritmos mostraram sobre os demais nas simulações aqui apresentadas.

Com relação ao tempo de processamento, verificou-se que quanto maior o número de indivíduos e menor a quantidade de informação, maior o tempo necessário para a reconstrução do mapa e estimação da fração de recombinação via cadeia de Markov oculta. Isso ocorre pois quanto maior o número de indivíduos, maior o número de fatores no produtório usado para o cálculo da verossimilhança; quanto menor for a quantidade de informação fornecida por estes indivíduos, maior o número de estados ocultos que devem ser inferidos. Assim sendo, o tempo de processamento computacional

depende muito do tamanho amostral e da informatividade dos marcadores. Com os recursos computacionais usados no LGN/ESALQ/USP, foram necessários 6 minutos para a construção de um mapa a partir de uma população com 100 indivíduos, sem dados perdidos e contendo apenas marcadores codominantes, usando o algoritmo TRY e o RIPPLE com o critério LHMC. Entretanto, para uma população com 400 indivíduos, com 20% de dados perdidos e contendo marcadores dominantes em repulsão, o tempo necessário foi de aproximadamente 8 horas (cerca de 80 vezes maior). Fica claro que o geneticista precisa equacionar o problema antes de iniciar as análises para que os resultados sejam produzidos em um tempo viável.

Os critérios SARF, PARF e SALOD, quando aplicados com o algoritmo RIPPLE, usam a informação dos marcadores dois a dois e não mostraram bons resultados em alguns casos estudados. Por sua vez, o critério LHMC (que é multiponto), aplicado com o algoritmo RIPPLE, mesmo com uma janela limitada de 4 marcadores, mostrou uma tendência de fornecer melhores estimativas de distância em várias situações. No presente trabalho, janelas superiores a 4 marcadores usando a cadeia de Markov oculta inviabilizariam as análises Monte Carlo, mas em situações práticas em que apenas um mapa é construído, utilizando computadores modernos, esta janela poderia ser de tamanho por volta de $m' = 10$, o que certamente traria melhorias para as estimativas de ordem e de distância. Caso o número de marcadores por grupo de ligação seja muito elevado (da ordem de centenas ou milhares), uma alternativa seria portanto ordenar os marcadores com o algoritmo SER, aplicando em seguida o algoritmo RIPPLE com LHMC com uma janela de marcadores m' de acordo com a disponibilidade computacional.

Em situações práticas não se sabe “a priori” a ordem dos marcadores, não sendo possível então saber se haverá marcadores dominantes ligados em repulsão, que motivam o uso dos algoritmos TRY ou SER. Mesmo assim, em casos que existam poucos marcadores dominantes, há pouca chance de formação de “clusters” desse tipo de marcador e, conseqüentemente, há pouca chance deles estarem ligados em fase de repulsão. Nesse caso, a escolha do algoritmo poderia ser feita entre TRY, SER, RECORD e UG, recomendando-se usar o RIPPLE com o critério LHMC. Caso existam grandes proporções de marcadores dominantes, a probabilidade desses marcadores estarem agrupados em “clusters” aumenta, e conseqüentemente, a probabilidade dos mesmos estarem ligados em fase de repulsão é mais elevada. Dessa forma, os algoritmos TRY e SER seriam os mais indicados associados ao RIPPLE com o critério LHMC.

5 CONCLUSÕES

- i.) Os algoritmos TRY e SER foram os mais eficientes para a ordenação dos marcadores e estimação das distâncias entre os mesmos quando há presença de marcadores dominantes ligados em repulsão.
- ii.) Quando não há marcadores dominantes em repulsão os algoritmos com melhor desempenho foram TRY, SER, RECORD e UG.
- iii.) Com relação ao algoritmo RIPPLE, recomenda-se que o mesmo seja realizado usando o critério LHMC.

REFERÊNCIAS

AHN, S.; TANKSLEY, S.D. Comparative linkage maps of the rice and maize genomes.

Proceedings of the National Academy of Sciences of the United States, Washington, v. 90, p. 7980-7984, 1993.

BASTEN, C.J.; WEIR, B. S.; ZENG, Z. B. **QTL Cartographer**: Version 1.17. Department of Statistics, North Carolina State University, North Carolina: Raleigh, 2003.

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**, Baltimore, v. 32, p. 314-331, 1980.

BOWERS, J.E.; ABBEY, C.; ANDERSON, S.; CHANG, C.; DRAYE, X.; HOPPE, A. H.; JESSUP, R.; LEMKE, C.; LENNINGTON, J.; LI, Z.; LIN, Y.; LIU, S.; LUO, L.; MARLER, B.S.; MING, R.; MITCHELL, S.E.; QIANG, D.; REISCHMANN, K.; SCHULZE, S.R.; SKINNER, D.N.; WANG, Y.W.; KRESOVICH, S.; SCHERTZ, K.F.; PATERSON A.H. A High-Density Genetic Recombination Map of Sequence-Tagged Sites for Sorghum, as a Framework for Comparative Structural and Evolutionary Genomics of Tropical Grains and Grasses **Genetics**, Bethesda, v. 165, p. 367-386, 2003.

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

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

CATO, S.A.; GARDNER, R.C.; KENT, J.; RICHARDSON, T.E. A rapid PCR-based method for genetically mapping ESTs. **Theoretical and Applied Genetics**, New York, v. 102, p. 296-306,

2001.

CARDLE, L.; RAMSAY, L.; MILBOURNE, D.; MACAULAY, M.; MARSHALL, D.; WAUGH, R. Computational and experimental characterization of physically clustered simple sequence repeats in plants. **Genetics**, Bethesda, v. 156, p. 847-854, 2000.

CHETELAT, R.T.; MEGLIC, V.; CISNEROS, T. A. Genetic map of tomato based on BC_1 *Lycopersicon esculentum* x *Solanum lycopersicoides* reveals overall synteny but suppressed recombination between these homeologous genomes. **Genetics**, Bethesda, v. 154, p. 857-867, 2000.

CORMEN, T.H.; LEISERSON, C.E.; RIVEST, R.L.; STEIN, C. **Introduction to Algorithms**, 2.ed., Cambridge: MIT Press, 2003. 1048 p.

DEMPSTER, A.P., LAIRD, N.M., RUBIN, D.B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society**, London, v. 39, p. 1-38, 1977.

DEVOS, K.M.; BEALES, J.; NAGAMURA, Y.; SASAKI, T. Arabidopsis-Rice: Will colinearity allow gene prediction across the eudicot-monocot divide? **Genome Research**, Woodbury, v. 9, p. 825-829, 1999.

DOERGE, R.W. Constructing genetic maps by rapid chain delineation. **Journal of Quantitative Trait Loci**, Washington, v. 2, 1996. Disponível em:
<<http://wheat.pw.usda.gov/jag/papers96/paper696/doerge2.htm>>. Acesso em: 01 nov. 2007.

DUFOUR, P.; DEU, M.; GRIVET, L.; D'HONT, A.; PAULET, F.; GLASZMANN, J.C. Construction of a composite sorghum genome map and comparison with sugarcane, a related complex polyploid. **Theoretical and Applied Genetics**, New York, v. 94, p. 409-418, 1997.

EDWARDS, D.; FORSTER, J.W.; CHAGNÉ, D.; BATLEY, J. What are SNPs? In: ORAGUZIE, N.C.; RIKKERINK, E.H.A.; GARDINER, S.E.; SILVA, H.N.D. (Eds.) **Association Mapping in Plants**, New York: Springer, 2007. chap. 3, p. 41-52.

FALCONER, D.S.; MACKAY, T.F.C. **Introduction to Quantitative Genetics**. 4.ed. Essex: Longman Group Limited, 1996. 480 p.

FALK, C.T. A simple scheme for preliminary ordering of multiple loci: application 10 45 CF families. In: ELSTON, R.C.; SPENCE, M.A.; RODGE, S.E.; MCCLUE, J.W. (Eds.) **Multipoint mapping and linkage based upon affected pedigree members**, New York: Alan R Liss, 1989. chap. 6, p. 17-22.

GEBHARDT, C.; WALKEMEIER, B.; HENSELEWSKI, H.; BARAKAT, A.; DELSENY, M.; STÜBER, K. Comparative mapping between potato (*Solanum tuberosum*) and *Arabidopsis thaliana* reveals structurally conserved domains and ancient duplications in the potato genome. **The Plant Journal**, Oxford, v. 34, p. 529-541, 2003.

GELFAND, A. E. Rapid seriation methods with archaeological application. In: HODSON, F.R.; KENDALL, D.G.; TAUTU, P. (Eds.) **Mathematics in the Archaeological and Historical Sciences**, Edinburgh: Edinburgh University Press, 1971. chap. 6, p. 186-201.

SCHOLAR GOOGLE. Disponível em: <<http://scholar.google.com>>. Acesso em: 13 nov. 2007.

GUIMARAES, C.T.; SILLS, G.R.; SOBRAL, B.W.S. Comparative mapping of *Andropogoneae*: *Saccharum* L. (sugarcane) and its relation to sorghum and maize. **Proceedings of the National Academy of Sciences of the United States**, Washington, v. 94, p. 14261-14266, 1997.

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**, London, v. 90, p. 33-38, 2003.

IHAKA, R.; GENTLEMAN R. R: A Language for Data Analysis and Graphics. **Journal of Computational and Graphical Statistics**, Alexandria, v.5, p. 299-314, 1996.

JACCOUD, D.; PENG, P.; FEINSTEIN, D.; KILIAN, A. Diversity Arrays: a solid state technology for sequence information independent genotyping. **Nucleic Acids Research**, London, v. 29, e25, 2001.

JIANG, C.; ZENG, Z. B. Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. **Genetica**, Gravenhage, v. 101, p. 47-58, 1997.

KATENTY, R. V.; LA ROTA, M.; MATHEUS, D. E.; SORRELLS, M. E. Dataminig for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum, and wheat. **Plant Molecular Biology**, Dordrecht, v. 48, p.501-510, 2002.

KEYGENE N.V. (United States). Marc Zabeau.; Pieter Vos **Selective restriction fragment amplification: a general method for DNA fingerprinting US 7026115** dez. 1996.

KERNIGHAN, B.; RITCHIE, D. C. **C: a linguagem de programação padrão ANSI**, 15.ed. São Paulo:Editora Campus. 1989. 289 p.

KILIAN, A.; KUDRNA, D.A.; KLEINHOF, A.; YANO, M.; KURATA, N.; STEFFENSON, B.; SASAKI, T. Rice-barley synteny and its application to saturation mapping of the barley Rpg1 region. **Nucleic Acids Research**, London, v. 23, p. 2729-2733, 1995.

KU, H.; LIU, J.; DOGANLAR, S.; TANKSLEY, S.D. Exploitation of *Arabidopsis*-tomato synteny to construct a high-resolution map of the ovate-containing region in tomato chromosome 2. **Genome**, Ottawa, v. 44, p. 470-475, 2001.

LANDER, E.S.; GREEN, P. Construction of multilocus genetic linkage maps in humans. **Proceedings of the National Academy of Sciences of the United States**, Washington, v. 84, p. 2363-2367, 1987.

LANDER, E.S.; GREEN, P.; ABRAHANSON, J.; BARLOW, A.; DALY, M.J.; LINCON S.E.; NEWBURG, L. MAPMAKER: An interactive computing package for constructing primary genetic linkages of experimental and natural populations. **Genomics**, Orlando, v. 1, p. 174-181, 1987.

LATHROP, G. M.; LALOUEL, J. M.; JULIER, C.; OTT, J. Strategies for multilocus linkage analysis in human. **Proceedings of the National Academy of Sciences of the United States**, Washington, v. 81, p. 3446-3446, 1984.

LATHROP, G. M.; LALOUEL, J. M.; JULIER, C.; OTT, J. Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. **American Journal of Human Genetics**, Baltimore, v. 37, p. 482-498, 1985.

LEE, M. DNA markers and plant breeding programs. **Advances in Agronomy**, New York, v. 55, p.265-344, 1995.

LIU, B.H. **Statistical Genomics: Linkage, Mapping, and QTL Analysis**. 2.ed. Boca Raton: CRC Press, 1996. 648 p.

MANLY, B. **Randomization, Bootstrap and Monte Carlo Methods in Biology**, 2.ed. London: Chapman & Hall/CRC, 1997. 424 p.

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

MESTER, D.; RONIN, Y.; MINKOV, D.; NEVO, E.; KOROL, A. Constructing Large-Scale Genetic Maps Using an Evolutionary Strategy Algorithm **Genetics**, Bethesda, v. 165, p. 2269-2282, 2003.

METROPOLIS, N.; ROSENBLUTH, A.W.; ROSENBLUTH, M.N.; TELLER, A.H.; TELLER, E. Equations of State Calculations by Fast Computing Machines. **Journal of Chemical Physics**, Lancaster, v. 21, p. 1087-1092, 1953.

MORGANTE, M.; SALAMINI, F. From plant genomics to breeding practice. **Current Opinion in Biotechnology**, Philadelphia, v. 14, p. 214-219, 2003.

MORTON, N.E. Sequential tests for the detection of linkage. **American Journal of Human Genetics**, Baltimore, 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**, Baltimore, v. 47, p. 470-482, 1990.

PATERSON, A.H.; BOWERS, J.E.; BUROW, M.D.; DRAYE, X.; ELSIK, C.G.; JIANG, C.X.; KATSAR, C.S.; LAN, T.H.G.; LIN, Y.R.; MING, R.; WRIGHT, R.J. Comparative Genomics of Plant Chromosomes **The Plant Cell**, Baltimore, v. 12, p. 1523-1540, 2000.

RABINER, R.L. A tutorial on hidden Markov models and selected applications in speech recognition. **Proceedings of the Institute of Electrical and Electronics Engineers**, London, v. 77, p. 257-286, 1989.

RONG, J.; BOWERS, J.E.; SCHULZE, S.R.; WAGHMARE, V.N. ; ROGERS C.J.; PIERCE, G.J.; ZHANG, H.; ESTILL, J.C.; PATERSON, A.H. Comparative genomics of Gossypium and Arabidopsis: Unraveling the consequences of both ancient and recent polyploidy. **Genome Research**, Woodbury, v. 15, p. 1198-1210, 2005.

SAIKI, R.K.; SCHARF S.; FALOONA, F.; MULLIS, K.B.; HORN, G.T.; ERLICH, H.A.; ARNHEIM, N. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. **Science**, Washington, v. 230, p. 1350-1354, 1985.

SALVI, S.; TUBEROSA, R. To clone or not to clone plant QTLs: present and future challenges. **Trends in Plant Science**, Kidlington, v. 10, n. 6, p. 1360-1385, 2005.

SCHLÖTTERER, C. The evolution of molecular markers-just a matter of fashion? **Nature**, London, v. 5, p. 63-69, 2004.

SOUZA, A.P. Biologia molecular aplicada ao melhoramento. In: NASS, L.L.; VALOIS, A.C.C.; MELLO, I.S.; VALADARES-INGLIS, M.C. (Eds.) **Recusos genéticos e melhoramento de plantas**. 1.ed. Rondonópolis: Fundação MT, 2001. cap. 29, p. 939-965.

SPEARMAN, C. The proof and measurement of association between two things. **American Journal of Psychology**, Ithaca, v.15, p.72-101, 1904.

SPEED, T.P.; MCPREECK, M.S.; EVANS, S.N. Robustness of the no-interference model for ordering genetic markers. **Proceedings of the National Academy of Sciences of the United States**, Washington, v. 89, p. 3103-3106, 1992.

STAM, P. Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. **The Plant Journal**, Oxford, v. 3, p. 739-744, 1993.

STURTEVANT, A. H. The linear arrangement of six X-linked factors in *Drosophila*, as shown by their mode of association. **Journal of Experimental Zoology**, Philadelphia, v. 14, p. 43-59, 1913.

SYVÄNEM A.C. Accessing genetic variation: genotyping single nucleotide polymorphisms. **Nature Reviews Genetics**, New York, v. 2, p: 930-942, 2001.

TAN, Y.; FU, Y. A novel method for estimating linkage maps. **Genetics**, Bethesda, v. 173, p. 2383-2390, 2006.

TAUTZ, D. Hypervariability of simple sequences as a general source for polymorphic DNA markers. **Nucleic Acids Research**, London, v. 17, p. 6463-6471, 1989.

THOMPSON, E. A. Information Gain in Joint Linkage Analysis **IMA Journal of Mathematics Applied in Medicine and Biology**, Oxford, v. 1, p. 31-49, 1984.

VAN OS, H.; STAM, P.; VISSER, R. G. F.; VAN ECK, H. J. RECORD: a novel method for ordering loci on a genetic linkage map. **Theoretical and Applied Genetics**, New York, v. 112, p. 30-40, 2005.

WEEKS, D.; LANGE, K. Preliminary ranking procedure for multilocus ordering. **Genomics**, Orlando, v.1, 236-242, 1987.

WHITKUS, R.; DOEBLEY, J.; LEE, M. Comparative genome mapping of sorghum and maize. **Genetics**, Bethesda, v. 132, p. 119-130, 1992.

WILLIAMS, J.G.K.; KUBELIK, A.R.; LIVAK, K.J.; RAFALSKI, J.A.; SCOTT, V. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. **Nucleic Acids Research**, London, v. 18, p. 6531-6535, 1990.

WILSON, S.R. A major simplification in the preliminary ordering of linked loci. **Genetic Epidemiology**, New York, 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**, New York, v. 107, p. 568-573, 2003.