EFICIÊNCIA DOS MÉTODOS DE OTIMIZAÇÃO SIMULATED ANNEALING, DELINEAÇÃO RÁPIDA EM CADEIA E RAMOS E CONEXÕES PARA CONSTRUÇÃO DE MAPAS GENÉTICOS

Efficiency of the simulated annealing, rapid chain delineation, and branch and bounds optimization methods in genetic mapping

Quintiliano Siqueira Schroden Nomelini[1], Heyder Diniz Silva[2], Tiago Costa Faria[3]

RESUMO

Um mapa genético é um diagrama onde são representados os genes com suas respectivas posições no cromossomo. Eles são essenciais para o procedimento de localização de genes envolvidos no controle genético de caracteres quantitativos ou no controle de outros caracteres de interesse econômico. No presente trabalho avalia-se, via simulação computacional de dados, a eficiência dos algoritmos simulated annealing, delineação rápida em cadeia e ramos e conexões, para a construção de mapas genéticos. Nas condições avaliadas, o algoritmo ramos e conexões foi o mais rápido, sendo que tanto este, quanto a delineação rápida em cadeia apresentaram 100% de eficiência. A eficiência do simulated annealing para ordenação de marcadores variou com o número de marcadores, para 5 e 10 foi de 100%, para 15 99,8% e com 20 marcadores a eficiência obtida foi de 99,2%.

Termos para indexação: Delineação rápida em cadeia, ramos e conexões, marcadores moleculares, ordenação.

ABSTRACT

The efficiency of Simulated Annealing (SA), Rapid Chain Delineation (RCD) and Branch and Bounds (BB) algorithms was evaluated by a Monte Carlo method. Regarding the conditions appraised the Branch and Bounds showed to be the fastest among them. Both RCD and BB were 100% efficient. The efficiency of SA depends on the length of the linkage group to be ordered. For 5 and 10 the efficiency was 100%, for 15 it was 99.8% and for 20 it was 99.2%.

Index terms: Simulated annealing, rapid chain delineation, branch and bound, molecular markers.

(Recebido em 20 de setembro de 2005 e aprovado em 12 de janeiro de 2009)

INTRODUÇÃO

Um mapa genético é um diagrama onde são representados os genes com suas respectivas posições no cromossomo, ou seja, a distribuição sequencial dos genes ao longo dos cromossomos. Jansen et al. (2001) destacam que os mapas genéticos ou mapas de ligação têm grande aplicação em muitas áreas da genética, tais como o mapeamento de QTLs (Quantitative Trait Loci), clonagem gênica e seleção assistida por marcadores.

Um dos problemas envolvidos na construção de um mapa genético refere-se à ordenação dos locos gênicos, ou seja, verificar se a ordem correta de três locos, *A, B* e *C* são: *ABC*, *BAC* ou *BCA*. Uma vez que a ordem correta não é conhecida, torna-se necessário definir um critério para identificar qual, dentre as possíveis, é a melhor. Hackett et al. (2003) listam vários critérios descritos na literatura, para identificação da "melhor" ordem, dentre eles a minimização da soma das frequências de recombinação adjacentes.

De acordo com Weir (1996), a inferência da ordem correta dos marcadores, baseada na soma de coeficientes de recombinação adjacentes (Sum of Adjacent Recombination Coefficients - *SAR*), fundamenta-se no fato de que a *SAR*, sob a ordem incorreta, nunca é menor do que sob a ordem correta. Assim, a ordem correta dos locos pode ser obtida como sendo a ordem que minimiza o valor de *SAR*.

Weir (1996) salienta, no entanto, que, quando a análise envolve um número muito grande de locos, o número de ordens possíveis e que devem ser avaliadas, pode tornar o procedimento impraticável, uma vez que, para *m* locos, esse número é igual a: $n^o\ de\ ordens = \frac{m!}{2}$. Ou seja, para 10 locos existem 1.814.400 possíveis ordens!

Como apontado por Jansen et al. (2001), o número de locos utilizados para construção de um mapa genético,

[1]Graduado em Matemática pela Universidade Federal de Uberlândia – Mestre em Estatística e Experimentação Agropecuária pela Universidade Federal de Lavras – Universidade Federal de Uberlândia – Campus Pontal – Av. José João Dib 2545 – 38302-000 – Progresso – Ituiutaba, MG – quintiliano@pontal.ufu.br
[2]Graduado em Agronomia pela Universidade Federal de Lavras – Professor Efetivo pela Universidade Federal de Uberlândia. FAMAT-UFU – Campus Santa Mônica – 37400-902 – Uberlândia, MG.
[3]Graduado em Ciência da Computação pela Universidade Federal de Uberlândia – Mestrando em Inteligência Artificial pela Universidade Federal de Uberlândia – Rua Tapuios, 838 – Saraiva – 38408-416, Uberlândia MG.

em algumas espécies pode variar de 1000 a 2000, o que torna o procedimento de busca exaustiva impossível de ser realizado, mesmo com modernos computadores. Sendo então imprescindível à utilização de métodos de otimização para tal fim.

Como salientado por Hackett et al. (2003), a ordenação de marcadores moleculares é um variante do problema clássico do "caixeiro viajante", o qual tem recebido muita atenção na área de otimização. Dentre as técnicas de otimização disponíveis na literatura, e que tem sido utilizadas na construção de mapas genéticos destaca-se o simulated annealing (Kirkpatrick et al., 1983), que é um método randômico onde os parâmetros são variados de acordo com regras de probabilidade. É uma técnica probabilística capaz de encontrar o mínimo global mesmo na presença de vários mínimos locais. O algoritmo examina todas possíveis ordens de uma grande porção amostral. Esse algoritmo baseia-se na observação de que quando metais liquefeitos são esfriados lentamente, cristalizam-se em um estado de energia mínima. Esse algoritmo encontra-se implementado em alguns programas computacionais para construção de mapas genéticos: PGRI (Lu & Liu, 1995), GMENDEL (Liu & Knapp, 1990) CARTHAGENE (Schiex & Gaspin, 1997).

Outros algoritmos comumente utilizados para construção de mapas genéticos são o "ramos e conexões", Branch and Bound (Thompson, 1987), que é semelhante a uma árvore de decisão e garante obter a melhor, ordem examinando apenas um conjunto pequeno de todas possíveis ordens, e a delineação rápida em cadeia, Rapid Chain Delineation (Doerge, 1996).

Neste trabalho, objetivou-se verificar, via simulação computacional de dados, a eficiência do algoritmo simulated annealing no problema da ordenação de marcadores moleculares.

MATERIAL E MÉTODOS

A eficiência dos algoritmos simulated annealing, delineação rápida em cadeia e ramos e conexões, no que concerne à ordenação de marcadores moleculares foram avaliadas via simulação computacional de dados. Para tanto, foram simuladas sequências com 5, 10, 15 e 20 marcadores moleculares, com ordem e distâncias entre marcadores conhecidas. As distâncias entre dois marcadores consecutivos, em Morgans, foram simuladas segundo uma distribuição uniforme (0,0001 a 0,2000).

Simulada a sequência, calcularam-se as distâncias entre todos os pares de marcadores, obtendo, assim, a matriz de distâncias em Morgans, que foi transformada em uma matriz de frequências de recombinação, por meio da função de mapeamento de Haldane (Haldane, 1919).

O critério de otimização adotado para identificação da "melhor" ordem a SAR ("Sum of Adjacent Recombination Coeficients"), a qual, conforme demonstrado por Weir (1996), nunca é menor sob a ordem incorreta do que sob a correta.

O algoritmo simulated annealing implementado consistiu em gerar uma ordem aleatória inicial qualquer (E_1). Calcular a soma das frequências das recombinações adjacentes desta ordem (SAR_1). Em seguida, provocar uma permutação aleatória na ordem inicial, obtendo uma nova ordem, E_2, e a SAR associada a essa nova ordem (SAR_2).

Se E_2 for melhor que E_1, isto é, $SAR_2 < SAR_1$, a nova ordem é adotada. A perturbação aleatória provocada consistiu na troca de posição entre dois marcadores.

Adotou-se como critério de parada do algoritmo, a "melhor" ordem que foi obtida quando, em M aconteceram sucessivas perturbações, não se conseguiu nenhuma redução na *SAR*, sendo M = 100 para sequências com 5 marcadores, M = 200 para sequências com 10 marcadores, M = 300 para sequências com 15 marcadores e M = 500 para sequências com 20 marcadores.

A Delineação Rápida em Cadeia (DRC), foi implementada de acordo com as indicações de Doerge (1996) e consistiu em que, a partir das frequências de recombinação entre todos os pares de marcadores, pegar-se o par de marcadores mais próximos. Em seguida, acrescentar à cadeia o loco mais próximo a um dos extremos e, assim, sucessivamente, até que todos os locos tenham sido colocados na cadeia. E, por último, fazer permutações de marcadores adjacentes e ver se diminui a *SAR*, adotando como ordem correta a de menor *SAR*.

Para o algoritmo ramos e conexões ("Branch and Bounds") (Thompson, 1987) adotou-se a seguinte estrutura: tomar um par de marcadores qualquer, por exemplo, *B, D*. O próximo loco a ser inserido, sendo A, que pode estar em três posições, correspondendo às ordens *ABD*, *BAD*, *BDA*. Calcular as SARs para as ordens *ABD*, *BAD*, *BDA*. Identificar dentre as três a de menor SAR. Desprezar as demais ordens e prosseguir com a inserção do próximo marcador. Repetir o processo até que todos os marcadores tenham sido inseridos.

A eficiência dos três métodos de otimização foi avaliada em termos do número de vezes em que o algoritmo chegou à ordem correta e do tempo médio de convergência. Para tanto, foram simuladas 1000 sequências aleatórias para cada um dos quatro tamanhos considerados (5, 10, 15 e 20 marcadores).

Para a execução do presente trabalho, desenvolveu-se um programa em linguagem Java, pertencente ao

paradigma orientado ao objeto. O programa consiste em um módulo central, responsável pela simulação das sequências e cálculo das matrizes de distâncias e os módulos responsáveis pelos algoritmos. O módulo central também requisita execução dos algoritmos, e faz a consequente coleta dos resultados. Esses resultados, então, são exibidos na tela. As informações da saída do programa são a porcentagem de acertos, o tempo médio de execução e o tempo total de execução para cada algoritmo, além do número de marcadores, número de execuções e tempo total de execução do programa.

RESULTADOS E DISCUSSÃO

Na TABELA 1, encontram-se apresentadas as porcentagens de acerto, isto é, o número de vezes que se obteve a ordem correta, em função do algoritmo e do número de marcadores. Verifica-se, que para sequências pequenas, 5 ou 10 marcadores, os três algoritmos apresentaram 100% de eficiência, ou seja, para as mil sequências simuladas, os três algoritmos obtiveram a ordem correta. Contudo, ao aumentar o número de marcadores para 15, a eficiência do simulated annealing caiu para 99,8% e para 99,2% com 20 marcadores.

Tabela 1 – Porcentagem de convergência para ordem correta, em função do número de marcadores considerados*.

	Número de marcadores			
Algoritmos	5	10	15	20
Simulated Annealing	100,0	100,0	99,8	99,2
DRC	100,0	100,0	100,0	100,0
Ramos e conexões	100,0	100,0	100,0	100,0

*valores obtidos a partir de1000 seqüências.

Por tratar-se de um método randômico, a possibilidade de obtenção de um resultado que não é o melhor não é descartada, tendo, já sido levantado por alguns autores. Kirkpatrick et al. (1983) ao apresentarem o simulated annealing levantam a possibilidade de o algoritmo caminhar para um ponto de mínimo local. Sugerem que, no intuito de explorar melhor o espaço das soluções, permita-se que uma ordem pior seja aceita com uma probabilidade P onde:

$$P = \min\left(e^{\frac{-(SAR_2 - SAR_1)}{k_b T}}, 1 \right)$$

Nessa função de probabilidade, T corresponde à temperatura, e k_b é uma constante física conhecida como a constante de Boltzmann e igual a $8.623 * 10^{-5}$. Inicialmente $k_b T$ é um baixo valor, significando que serão permitidas ordens piores com uma probabilidade alta. Com a evolução do processo, P é abaixado, ficando cada vez mais difícil de ser adotada uma ordem pior. Hackett et al. (2003) ao utilizar o simulated annealing para obtenção de mapas de ligação em espécies autotetraplóides, eliminaram a constante k_b do cálculo da probabilidade, sendo que a cada M perturbações aleatórias, a temperatura T era reduzida, por um fator a, para aT. Verificaram que uma temperatura inicial $T=20$, um fator de resfriamento a=0,85 e $M=100n$, onde n é o número de marcadores, eram suficientes para explorar o conjunto de possíveis ordens (espaço das soluções) convenientemente.

No intuito de melhorar a eficiência do simulated annealing para sequências maiores, optou-se por adotar como critério de parada, a não obtenção de melhora em 100n perturbações, isto é, 1500 perturbações para as sequências de 15 marcadores e 2000 para as de 20. Com isso, a eficiência para sequências de 15 marcadores foi de 100%, mas, com 20 marcadores, a eficiência foi de 99,8%, indicando que apenas esse critério não foi suficiente para garantir a exploração do todo o espaço das soluções. Outros mecanismos que permitam uma melhor exploração dos espaços das soluções, como permitir, com alguma probabilidade, a utilização de sequências piores, tornam-se necessários, sendo que sua inclusão poderá aumentar a eficiência do simulated annealing, devendo levá-la para 100% nas condições deste trabalho, como ocorreu com os algoritmos DRC e ramos e conexões.

Um fato que merece destaque no que concerne aos algoritmos DRC e ramos e conexões, refere-se à presença de dois ou mais pares de marcadores com a mesma distância (empates). Os algoritmos, da forma como descritos, não lidam com tais situações, o que levou a exclusão da possibilidade de tal fato no processo de simulação. Desse modo, a eficiência máxima (100%) independente do número de marcadores apresentados por estes métodos, deve ser vista com ressalvas. Deve-se destacar ainda que o simulated annealing não apresenta tal problema.

Os tempos médios de execução variaram de 0,041 mseg, (DRC para 5 marcadores), a 1,973 mseg (simulated annealing com 20 marcadores) (TABELA 2), em um computador equipado com processador de 1.3 Ghz e 256 Mb de memória RAM. Ao aumentar o critério de parada no simulated annealing, os tempos médios subiram para 3,64 mseg em sequências de 15 marcadores e 5,70 com 20.

Tabela 2 – Tempo médio (mseg) de execução dos algoritmos, em função do número de marcadores considerados*.

	Número de marcadores			
Algoritmos	5	10	15	20
Simulated Annealing	0,381	0,483	1,128	1,973
DRC	0,041	0,126	0,417	1,113
Ramos e conexões	0,051	0,092	0,171	0,362

* valores obtidos a partir de1000 seqüências.

CONCLUSÕES

Nas condições avaliadas, o algoritmo ramos e conexões foi o mais rápido para todos os níveis de marcadores estudados. Sendo que, com 5 marcadores, obteve o menor tempo de ordenação, 0,051 mseg. Tanto este quando a delineação rápida em cadeia apresentaram 100% de eficiência para os diferentes números de marcadores.

A eficiência do simulated annealing para ordenação de marcadores variou com o número de marcadores, para 5 e 10 foi de 100%, para 15 foi 99,8% e com 20 99,2%. E em relação ao tempo de convergência mostrou-se relativamente inferior aos outros dois algoritmos avaliados.

A eficiência dos algoritmos DRC e ramos e conexões, devem ser vistos com ressalvas, pois, os algoritmos aqui descritos, não lidam com situações em que aparecem dois ou mais pares de marcadores com a mesma distância. Vale destacar ainda que o simulated annealing não apresenta tal problema.

REFERÊNCIAS BIBLIOGRÁFICAS

DOERGE, R. Constructing genetic maps by rapid chain delineation. **Journal of Quantitative Trait Loci**, v.2, n.6, 1996.

HACKETT, C.A.; PANDE, B.; BRYAN, G.J. Constructing linkage maps in autotetraploid species using simulated annealing. **Theoretical and Applied Genetics**, Berlin, v.106, p.1107-1115, 2003.

HALDANE, J.B.S. The combination of linkage values, and the calculation of distance between the loci of linkad factors. **Journal of Genetics**, Berlin, v.8, p.299-309, 1919.

JANSEN, J.; JONG, A.G.; OOIJEN, J.W. Constructing dense genetic linkage maps. **Theoretical and Applied Genetics**, Berlin, v.102, p.1113-1122, 2001.

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

LIU, B.H.; KNAPP, S.J. Gmendel: a program for Mendelian segregation and linkage analysis of individual or multiple progenie populations using log-likelihood rations. **Journal of Heredit**, Oxford, v.81, p.407, 1990.

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

SCHIEX, T.; GASPIN, C. Cartagene: constructing and joining maximum likelihood genetic maps. In: PORC ISMB'97, 1997, Halkidiki, Greece. **Proceedings**... Halkidiki, 1997.

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

WEIR, B.S. **Genetic data analysis II**. Massachusetts: Sinauer Associates, 1996. 445p.