Theor Appl Genet (2003) 106:1107–1115
DOI 10.1007/s00122-002-1164-1

C. A. Hackett · B. Pande · G. J. Bryan

Constructing linkage maps in autotetraploid species using simulated annealing

Received: 23 May 2002 / Accepted: 26 August 2002 / Published online: 8 February 2003

Abstract In this paper we demonstrate how molecular markers segregating in a full-sib autotetraploid mapping population can be ordered to form a linkage map using simulated annealing. This approach facilitates the examination of orders close to the optimum to see which marker placings are fixed and identify the markers whose position is less certain. A simulation study investigates the effects of population size, marker spacing, ratio of dominant to codominant markers, typing errors and missing values. The method is applied to map 30 amplified fragment length polymorphism and microsatellite markers on linkage group IV of potato.

Keywords Autotetraploid · Potato · Ordering · Simulated annealing

Introduction

Linkage maps based on molecular markers are important tools in genetic analysis. It is therefore necessary to be able to find the best linear order of the markers, and also to explore the uncertainty in the ordering. Researchers on diploid species have developed several different criteria to define a 'best' order and have used a variety of search methods to find the order with the optimal criterion value. The criteria are functions of the recombination frequencies between pairs of markers and the associated LOD score [i.e. $\log_{10}$ of the likelihood ratio of the markers being linked to their being unlinked (Morton 1955)]. They include the maximum likelihood, the minimum sum of adjacent recombination frequencies (SARF), the maximum sum of adjacent LODs (SALOD), the minimum number of crossovers and the least square locus order. In some populations, for example a doubled haploid population derived from the F_1 generation of a cross between two inbred parents, the LOD score and the recombination frequency are related by a single monotonic function for all pairs of markers, and so criteria such as the SARF and SALOD give the same ordering. In crosses such as a full-sib population from two heterozygous parents, and especially if both dominant and codominant markers are scored, then some marker pairs will have a higher LOD score than others with the same recombination frequency (Maliepaard et al. 1997), and this must be taken into account in ordering the markers. Typing errors may lead to a large increase in map length using some criteria, while others, such as the least squares order, are more robust.

A linkage group with n markers has $n!/2$ possible orders, so an exhaustive search is feasible only for small n. The marker ordering problem is a variant on the classical travelling salesman problem, which has received much attention in the optimisation literature. Algorithms that have been applied for ordering markers include seriation, a stepwise search, the branch and bound algorithm, simulated annealing and genetic algorithms. Seriation was introduced by Buetow and Chakravarti (1987) to minimise SARF, a stepwise search is used in the JOINMAP programme (Stam and Van Ooijen 1995) to minimise the least squares order and the branch and bound algorithm is used by Thompson (1987) to minimise the number of obligatory crossovers. Simulated annealing has been used by several researchers to optimise different criteria, mainly for human data. Lander and Green (1987) investigated simulated annealing and a branch and bound search to find the marker order with the maximum log-likelihood for three-generation human pedigrees, using a Hidden Markov Model to reconstruct the expected number of recombinant meioses between markers. Weeks and Lange (1987) used simulated annealing with two different criteria, the sum of adjacent LOD scores and the

Communicated by J.W. Snape

C. A. Hackett (✉)
Biomathematics and Statistics Scotland, Scottish Crop Research Institute, Invergowrie, Dundee, Scotland DD2 5DE
e-mail: christine@bioss.ac.uk
Fax: +44-382-562426

B. Pande · G. J. Bryan
Scottish Crop Research Institute, Invergowrie,
Dundee Scotland DD2 5DE

least squares criterion of Lalouel (1977). Falk (1992) used simulated annealing to minimise the sum of adjacent recombination frequencies, but noted that this criterion may perform badly if the estimates of recombination frequency differed in their amount of information. The PGRI software (Lu and Liu 1995) uses simulated annealing and/or branch and bound and can minimise SARF or maximise the likelihood. GMENDEL (Liu and Knapp 1990) also uses simulated annealing with the minimum SARF. CARTHAGENE (Schiex and Gaspin 1997) has a choice of methods, including simulated annealing and genetic algorithms, to maximise the likelihood. Jansen et al. (2001) have recently proposed a two-stage approach to constructing dense marker maps in a backcross population, using simulated annealing first to order a framework of markers, then to order further markers relative to the framework. Their criterion was the total number of expected recombinations, using Gibbs sampling to handle missing data.

Linkage analysis in autopolyploid species has received much less attention due to the complexities in modelling polysomic inheritance. However, there are some important crops that are autopolyploids, including potato (tetraploid), alfalfa (tetraploid), sugarcane (octoploid) and strawberry (octoploid). The simplest model for polysomic inheritance is that of random chromosomal segregation – i.e. the random pairing of chromosomes to give bivalents and recombination within each bivalent. This model appears appropriate for modelling in alfalfa, where most cells have the full complement of 16 bivalents at metaphase I (Bingham and McCoy 1988), and in potato where bivalents predominate, although low frequencies of quadrivalents, trivalents and univalents have been observed (Swaminathan and Howard 1953). The complexities include multiple copies of alleles: if an individual, P_1, carrying an amplified fragment length polymorphism (AFLP) marker is crossed to an individual, P_2, without the marker, the offspring can have expected presence:absence ratios 1:1, 5:1 or 1:0 depending on whether P_1 has one, two or three or more copies of the allele respectively. For a multiallelic marker such as a simple sequence repeat (SSR), there can be up to four alleles for each parent, giving up to 24 different phases for each parent for a pair of markers.

Single-dose (simplex) markers in coupling phase in autopolyploid species can be analysed as if the individuals were diploid (Wu et al. 1992), and this approach has been used by Al-Janabi et al. (1993) and Da Silva et al. (1993) for mapping in sugarcane (*Saccharum spontaneum* SES 208) and by Brouwer and Osborn (1999) and Diwan et al. (2000) for alfalfa*(Medicago sativa* L.). Da Silva et al. (1995) and Brouwer and Osborn (1999) have placed double-dose (duplex) markers afterwards, near the simplex markers with the lowest recombination frequencies. Hackett et al. (1998) developed a theory for calculating recombination frequencies and LOD scores between pairs of dominant markers of any dosage in a full-sib population from a cross between two autotetraploid parents and ordered simulated markers from this pairwise information using JOINMAP's stepwise search for the least squares order. Meyer et al. (1998) and Bradshaw et al. (1998) used the same approach to analyse an experimental population of potato (*Solanum tuberosum* ssp. *tuberosum*). Luo et al. (2001) applied the EM algorithm to calculate recombination frequencies and LOD scores between pairs of dominant and/or codominant markers with up to eight alleles, in any phase. The order calculated by JOINMAP from such pairwise data was found to change markedly with small changes in the data.

In this paper, simulated annealing is used to optimise the least squares criterion for molecular markers segregating in a full-sib population from a cross between two autotetraploid parents. Orders close to the optimum can be examined to identify areas of uncertainty in the linkage map and to see which marker configurations are the most difficult to place. A small simulation study investigates the effects of population size, marker spacing, number of codominant markers, typing errors and missing values. We apply the algorithm to map a mixture of AFLPs and microsatellites on potato linkage group IV, where important resistance genes are known to be located (Gebhardt and Valkonen 2001). These include quantitative trait loci (QTLs) for resistance to the white potato cyst nematode *Globodera pallida* (Stone) (Bradshaw et al. 1998) and *Phytophthora infestans* (Leonards-Schippers et al. 1994) and the *R2* major gene for resistance to *P. infestans* (Li et al. 1998).

Methods

Simulated annealing

The idea of simulated annealing is derived from observations in thermodynamics. The slow cooling (annealing) of molten metal gives a minimum energy state, while faster cooling may leave molecules in an alternative state with higher energy. When minimising a criterion by simulated annealing, random changes are made to the state of the system and the criterion is evaluated for the new state. Changes leading to a decrease in the criterion are always accepted; changes leading to an increase in the criterion are accepted with a probability that decreases slowly according to a parameter (referred to as the 'temperature'). The system can therefore escape from local minima to find the global minimum.

In this application, a state of the system is an order of the n marker loci, i.e. some permutation of $\{1,\ldots,n\}$. Random changes are generated as suggested by Lin (1965) for the travelling salesman problem. A segment of the order is chosen randomly, then either this segment's direction is reversed, or it is transferred to another random location. The choice between these types of move is made randomly, with equal probability. The probability of accepting a change is

$$p = \min(\exp[-(S_2 - S_1)/T], 1)$$

where S_1, S_2 are the criterion values for the old and new states respectively, and T is the 'temperature'. The criterion S used here is the least squares criterion (see next section). If $S_2 \leq S_2$, then p=1, and the change is always accepted. M random changes are generated at each temperature T. The temperature is then reduced by a factor α to αT, decreasing the probability of accepting a change that increases the least squares criterion, and further changes are generated.

The simulated annealing algorithm used here was adapted from a Fortran 77 routine for simulated annealing of a continuous sample space by Goffe et al. (1994). The routines to generate random changes were modified for a non-circular order from the travelling salesman routines of Press et al. (1986). Some initial investigations showed that a starting temperature $T = 20$, a cooling factor $\alpha = 0.85$ and $M = 100n$ random changes at each temperature explored the set of possible orders sufficiently. Every accepted order was stored so that near-optimal orders could be identified and examined. It was observed on simulated data that the algorithm tended to quickly find orders that were correct except for reversal of adjacent markers. The programme was modified to include a ripple of the best order at each temperature prior to cooling, to test whether the exchange of any pair of adjacent markers lowered the criterion further. The best order after the ripple formed the starting point for random exchanges at the next temperature.

Least squares criterion

The criterion used here was the least squares criterion (Stam 1993), also used by JOINMAP. Modified versions of this were used by Jensen and Jorgensen (1975) and Lalouel (1977). Let r_{ij} be the recombination frequency between markers M_i and M_j and let W_{ij} be the corresponding LOD score. Let x_{ij} be the map distance between markers M_i and M_j, calculated as $x_{ij} = F(r_{ij})$ for some mapping function F. Here we have chosen to use Haldane's mapping function, but an alternative mapping function could be used. The pairwise map distances are combined to give a linkage map by estimating distances c_i between adjacent markers M_i and M_{i+1} so as to minimise the squared differences between the map distance x_{ij} calculated directly from the recombination frequency and that calculated as a sum of the distances between the intermediate markers:

$$\left[x_{ij} - (c_i + c_{i+1} + \cdots + c_{j-1})\right]^2.$$

Each squared difference is weighted by the LOD score W_{ij} to take into account the differences in precision of the estimates x_{ij}. This gives the least squares criterion for an order as

$$S = \sum_{i<j} W_{ij}\left[x_{ij} - (c_i + c_{i+1} + \cdots + c_{j-1})\right]^2.$$

Recombination frequencies of 0.5 between distantly linked markers are replaced with a value of 0.499 to avoid a map distance of infinity: the associated LOD score is close to zero so such pairs have negligible influence on the calculation of S. The distances $\{c_i\}$, their standard errors and the criterion S can be evaluated by weighted linear regression. This does not constrain the estimates of the distances to be positive, but a distance that is significantly less than zero (using a t-test) indicates an implausible order.

Simulation study

A simulation study was carried out to investigate how this approach ordered molecular markers in different situations. The cross was a full-sib population from two autotetraploid parents. If most of the markers are multi-allelic, then a combined map of the two parents can be estimated directly, as in the simulation study of Luo et al. (2001). However experimental marker data from such populations typically consists of a large number of dominant AFLP markers and a small number of multi-allelic SSR markers, and in this case it is necessary to construct a map for each parent separately and then align the maps using multi-allelic SSR markers and double-simplex AFLP markers, although the latter are less informative (Meyer et al. 1998). Our simulation study therefore considers the construction of a linkage map for parent 1 only. Table 1 shows the parental marker configurations used for the simulations, with three SSRs (two with four different alleles in parent 1, and one with three different alleles) and 17 AFLP markers (12 simplex, three duplex and two double-simplex, i.e. a single-dose allele in each parent). Parent 2 is assumed not to share any SSR alleles with parent 1, and its alleles at the two double-simplex markers are in repulsion phase.

Table 1 Marker configurations used in the simulation study

Locus	Type[a]	Parent 1				Parent 2			
		1	2	3	4	1	2	3	4
L1	SSR	*a*	*b*	*c*	*d*	*o*	*o*	*o*	*o*
L2	S	*a*	*o*	*o*	*o*	*o*	*o*	*o*	*o*
L3	S	*o*	*a*	*o*	*o*	*o*	*o*	*o*	*o*
L4	D	*o*	*o*	*a*	*a*	*o*	*o*	*o*	*o*
L5	S	*o*	*o*	*a*	*o*	*o*	*o*	*o*	*o*
L6	S	*o*	*o*	*o*	*a*	*o*	*o*	*o*	*o*
L7	S	*a*	*o*	*o*	*o*	*o*	*o*	*o*	*o*
L8	DS	*o*	*o*	*o*	*a*	*a*	*o*	*o*	*o*
L9	S	*o*	*a*	*o*	*o*	*o*	*o*	*o*	*o*
L10	SSR	*a*	*a*	*b*	*c*	*o*	*o*	*o*	*o*
L11	S	*o*	*o*	*a*	*o*	*o*	*o*	*o*	*o*
L12	D	*a*	*o*	*a*	*o*	*o*	*o*	*o*	*o*
L13	S	*o*	*o*	*o*	*a*	*o*	*o*	*o*	*o*
L14	S	*a*	*o*	*o*	*o*	*o*	*o*	*o*	*o*
L15	S	*o*	*a*	*o*	*o*	*o*	*o*	*o*	*o*
L16	D	*a*	*o*	*o*	*a*	*o*	*o*	*o*	*o*
L17	S	*o*	*o*	*a*	*o*	*o*	*o*	*o*	*o*
L18	S	*o*	*o*	*o*	*a*	*o*	*o*	*o*	*o*
L19	DS	*o*	*o*	*a*	*o*	*o*	*o*	*o*	*a*
L20	SSR	*a*	*b*	*c*	*d*	*o*	*o*	*o*	*o*

[a] SSR, Multi-allelic microsatellite, with *a*, *b*, *c*, *d* denoting different alleles, *o* = null allele; S, simplex marker; D, Duplex marker; DS, Double-simplex marker

The markers were simulated as equally spaced along a single set of four homologous chromosomes, with recombination frequencies of 0.05 or 0.10 between adjacent markers. Population sizes of 400 and 200 offspring were used. The effects of introducing genotyping errors (with an error rate of 2%), or missing values (5%) were investigated. The effect of excluding the two SSR markers with four alleles, L1 and L20, which are expected to contribute most to the precision of the map, was investigated. Ten replicates of each scenario were simulated.

Recombination frequencies and LOD scores were calculated for all pairs of markers in all possible phases using the EM algorithm (Luo et al. 2001). The recombination frequency and the LOD score for the phase with the highest likelihood for each pair were used to estimate the order with the minimum least squares criterion by simulated annealing.

Experimental data

Experimental data were obtained on a population derived from a cross between the potato cultivar Stirling and the advanced SCRI breeding line 12601ab1 (Bradshaw et al. 1995). The mapping of AFLP markers in 94 F_1 plants from this cross is described by Meyer et al. (1998). An extended population of 227 F_1 plants from this cross was scored with both AFLP and SSR markers by Pande (2002). The map presented there was calculated using the clustering approach of Luo et al. (2001) to separate the markers into linkage groups, and then the EM algorithm to calculate recombination frequencies and LOD scores. JOINMAP was used to order the markers using this pairwise data. Linkage groups were identified wherever possible by means of markers whose locations in diploid potato populations are known. In this paper we reconsider the ordering of markers from linkage group IV of parent 12601ab1, identified by SSRs STM3016 (Milbourne et al. 1997) and S140 (Ghislain, personal communication), and by six AFLP markers that also segregated in a diploid potato mapping experiment (Isidore 2001). This chromosome is known to be associated with a QTL for quantitative resistance to the white potato cyst nematode *Globodera pallida* (Stone) (Bradshaw et al. 1998).

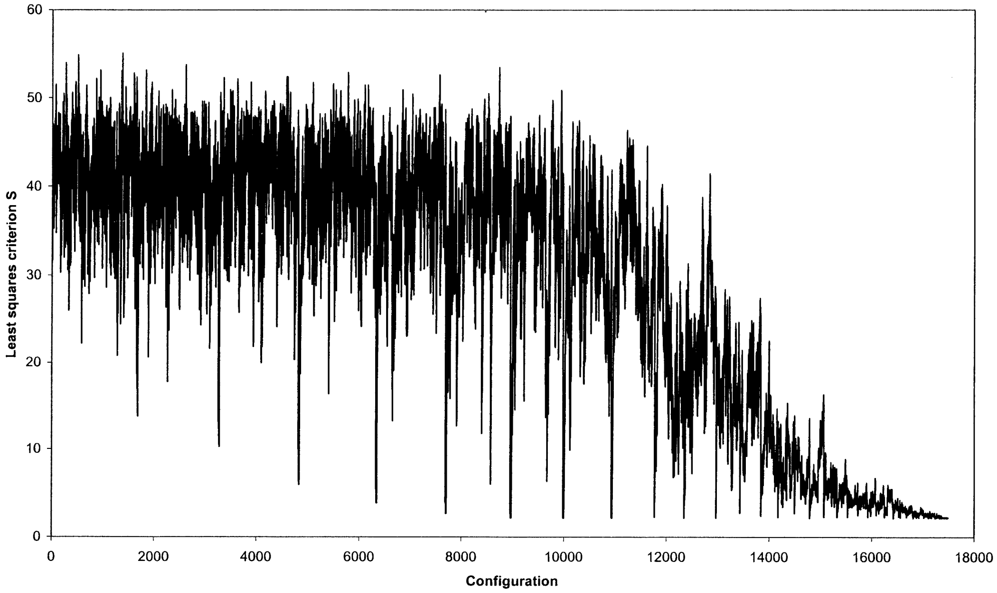

Fig. 1 Profile of the least squares criterion *S* while searching for the optimal order

Results

Simulation study

One simulation of the configuration in Table 1 (based on 400 offspring, recombination frequencies of 0.05 between adjacent markers and no genotyping errors or missing data) is discussed here in detail, and the rest are summarised. Recombination frequencies and LOD scores were calculated for all pairs of the 20 markers, giving 190 pairs. The true phase had the highest likelihood for 183 pairs. Four pairs of simplex markers in repulsion phase were incorrectly identified as being linked in coupling: all of these were distantly linked with recombination frequencies greater than 0.48 and LOD scores less than 0.02, so these pairs have little influence on the weighted least squares criterion. The three duplex-duplex pairs of markers, which are in a mixed phase, were also incorrectly identified as being in a repulsion phase. The simulated annealing algorithm was run twice to investigate the effect of using the wrong phase for these three pairs, but the final order was unchanged and the marker locations differed by at most 0.7 cM. The results below are based on using the repulsion phase.

A starting configuration (L8, L12, L19, L16, L7, L20, L17, L13, L3, L2, L4, L9, L15, L11, L1, L18, L10, L6, L14, L5) for the simulated annealing algorithm was generated randomly: this had a least squares criterion *S* of 43.7224. At the first temperature of 20, few configurations are rejected (7.5%) as the set of possible orders is explored. An order with *S* = 22.6875 is found at this temperature; this is reduced to *S* = 19.7825 by reversing adjacent loci. Loci L1-L5 are positioned at one end, although not in the correct order, and loci L16 and L17 are at the other end. After two temperature reductions, an order with *S* = 10.7047 has been found. This has L1–L7 in the correct order at one end and L18, L16 and L19 at the other end. The subsequent reversing stage moves L20 to the opposite end from L1 and decreases *S* to 4.0141. After two further steps, the criterion has decreased to 3.2757, and all loci are correctly ordered except for L8, a double-simplex marker, which is between L14 and L15. *S* continues to decrease slowly until L8 reaches its correct position, with *S* = 2.2254. No further improvements are found. Figure 1 shows the profile of *S* as the temperature decreased.

For this simulation, *S* was calculated 36,858 times, and 10,298 of the orders were accepted. This is not an exhaustive search of the possible orders (there are approximately 1.2×10^{18} orders for 20 loci), but it was unusual for different initial orders to give different final configurations. By sorting the accepted orders into increasing order of the criterion *S*, we can compare the optimal and near-optimal orders to see which areas of the map are well-established and which are less certain. The criterion was in excess of 50 for some marker orders, but there were a large number of markers with values of *S* close to the optimum. Among the best 100 orders, for example, the criterion increased from 2.2254 to 2.4251. Table 2 shows the number of occurrences of each marker in each position in the top 100 orders: L1 occurs in position 1 for 97/100 orders and otherwise in position 2; L2 occurs in position 2 for 91/100 orders and otherwise in positions 1 or 3, etc. L8 occurs in the correct position least often (48/100 orders). The last column of Table 2 shows a weighted mean location for each locus based on the top 100 orders, where order *i* is weighted with weight Z_i according to the least squares criterion:

$$Z_i = 1 - \left(\frac{S_i - S_1}{S_{100} - S_1}\right).$$

This weighted order agrees with the simulated order.

It is also relevant to consider the sign of the estimated distances $\{c_i\}$ between adjacent markers. These distances

Table 2 Number of occurrences of each marker locus in each position (Pos.) for the top 100 orders. WP is the average position for each locus, weighted according to the least squares criterion

Pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	WP
L1	97	3																			1.01
L2	3	91	6																		2.02
L3		4	66	29	1																3.32
L4		2	28	64	6																3.70
L5				7	71	22															5.15
L6					22	76	2														5.82
L7						2	71	24	3												7.25
L8							26	48	24	2											8.02
L9							1	28	59	11	1										8.83
L10									14	86	0										9.89
L11										1	97	2									11.01
L12											2	87	11								12.08
L13												11	83	6							12.93
L14													6	73	21						14.19
L15														21	77	2					14.81
L16															2	90	8				16.03
L17																8	81	11			17.05
L18																	10	59	31		18.21
L19																	1	30	68	1	18.69
L20																			1	99	20.00

Table 3 Summary of ten replicates of each simulation set

Set	Size[a]	Spacing[b]	Expected length	% Missing	% Error	Number of loci	No. correct[c]	Med. better[d]	Pairs reversed[e]	Mean length[f]	Neg. distance[g]
A	400	0.10	212.0	0	0	20	6	0	0.6	171.9	0
B	200	0.10	212.0	0	0	20	1	43.5	3.2	167.8	2
C	400	0.05	100.1	0	0	20	4	2.5	0.8	96.4	0
D	200	0.05	100.1	0	0	20	1	13.5	2.5	94.4	1
E	400	0.05	100.1	5	0	20	3	9	0.9	96.0	0
F	200	0.05	100.1	5	0	20	1	24	3.0	93.5	1
G	400	0.05	100.1	0	2	20	1	26.5	1.9	105.3	0
H	200	0.05	100.1	0	2	20	0	49.5	3.7	102.3	2
I	400	0.05	89.6	0	0	18	4	5.5	1.4	83.1	1
J	200	0.05	89.6	0	0	18	0	195.5	5.0	79.6	3

[a] Size, Number of offspring
[b] Spacing, Recombination frequency between adjacent markers
[c] No. correct, The number of simulations for which the simulated order had the minimum least squares criterion
[d] Med. better, The median number of orders with a smaller value of the least squares criterion than the true order
[e] Pairs reversed, The number of adjacent pairs exchanged in the optimal order compared to the true order
[f] Mean length, The mean length of the optimal orders (centiMorgans)
[g] Neg. distance, The number of simulations for which the true order had an estimated distance that was significantly less than zero

should all be positive, and this is the case for the optimal order. The second order has the locations of marker L7 and L8 reversed, and the corresponding distance is significantly less than zero (distance estimated as –0.0605, SE. 0.0196). Similarly, the third order has reversed L18 and L19, with the corresponding distance significantly less than zero. The fourth order has reversed L3 and L4, but the distance, while negative, is not significantly less than zero. It seems appropriate, in assessing alternative orders that also fit the marker data, to ignore those with significant negative distances.

Table 3 summarises the results for ten simulation sets and ten replicates of each set. The following summary statistics are presented: the number of simulations for which the simulated order had the minimum least squares criterion, the median number of orders better than the true order (i.e. with a smaller value of the least squares criterion), the mean number of adjacent pairs reversed in the optimal order compared with the true order, the mean length of the optimal map and the number of simulations for which the true order had an estimated distance that was significantly less than zero. An order such as L4, L1, L2, L3, L5 …L20 is regarded as reversing three adjacent pairs relative to the true order. The median number of orders is preferred to the mean here because it is affected less by extreme values. It should also be noted that as the number of orders better than the true order increases, it is increasingly likely to be an underestimate, because the simulated annealing algorithm is not guaranteed to show all the orders with a criterion between that of the true order and the optimum.

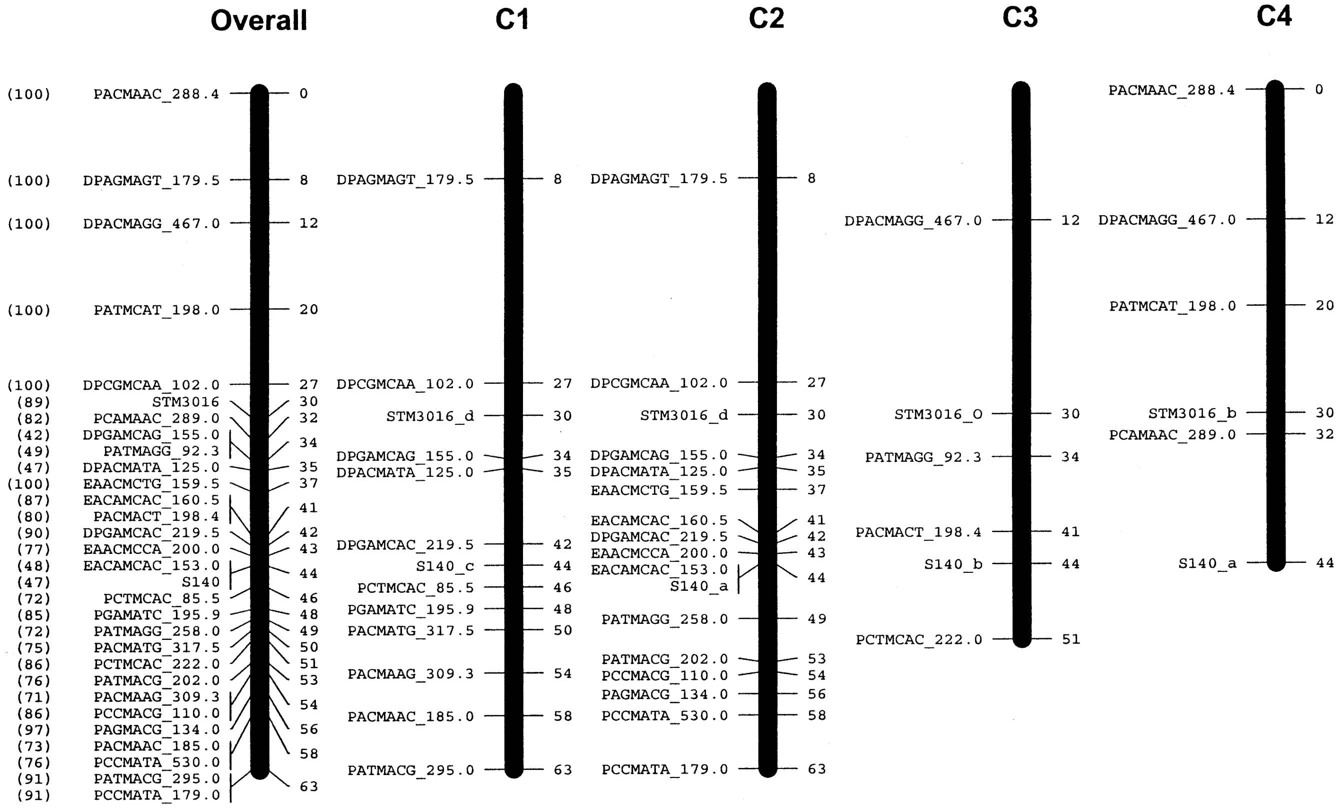

Fig. 2 Linkage map of potato chromosome IV for parent 12601ab1, showing markers and map positions (centiMorgans) for the four homologous chromosomes and for the combined order. Duplex markers are prefixed by a *D*, and suffices *a*, *b*, *c*, *d* indicates different alleles at the SSRs STM3016 and S140. The numbers in brackets show how many times that locus occupied that position for the top 100 orders

The simulated annealing algorithm ordered the markers best for set A, with 400 offspring, a recombination frequency of 0.1 between adjacent markers and no missing values or genotyping errors. The optimal order agreed with that simulated for six of the ten replicates, while the other four replicates had one, or at most two, adjacent pairs of markers reversed. However, the estimated map length was shorter than that simulated, probably due to the high proportion of marker pairs with estimated recombination frequencies between 0.4 and 0.5. For example, the recombination frequencies between L1 and L2–L8 increase monotonically from 0.1 to 0.3 as the true separation increases, while those between L1 and L9–L20 are all in the range 0.4–0.5 with no particular order relative to the true separation, which increases from 89 cM to 212 cM. The same shortening occurs for a population of 200 offspring and a recombination frequency of 0.1 between adjacent markers.

The algorithm generally performs well for a population of 400 offspring. Typically the median number of orders better than the true order is less than or equal to 9, and the order is correct apart from the reversing of one or two adjacent pairs. The true order has a significantly negative distance for only one simulation, in set I (without the informative markers L1 and L20). The estimated map length for a recombination frequency of 0.05 between adjacent markers was close to that simulated. The replacement of 5% of the marker information with missing values did not have a substantial effect. However the introduction of a small rate (2%) of genotyping errors into a population of 400 caused the median number of orders better than the true order to increase to 26.5, with up to four adjacent pairs of markers reversed. The length of the map was also increased by the inclusion of genotyping errors for both 400 and 200 offspring.

The algorithm had more difficulties in reconstructing the true order when the population size was reduced to 200. At most one of the ten replicate simulations was completely correct, while one or two replicate simulations had significantly negative distances for the true order. The best reconstruction was with a marker spacing of 0.05, and no missing values or genotyping errors. The median number of orders better than the true order was 13.5, with a mean of 2.5 pairs of adjacent markers reversed. A wider marker spacing caused more difficulties at this population size, with up to five pairs of adjacent markers reversed. The introduction of genotyping errors gave optimal orders with up to nine pairs of adjacent markers reversed. The worst simulation was when the highly informative loci L1 and L20 were dropped in the populations of 200 offspring, with a median number of orders better than the true order of 195.5.

The extent of the misplacement varied according to the type of the marker. The double simplex markers L8 and L19 were displaced furthest, up to four places from their true position. One of these markers was also involved whenever the true order had a significantly negative distance between a pair of adjacent markers. The three duplex loci L4, L12 and L16 were displaced up to three places, while the simplex loci were displaced by one or two places. Misplacements of the codominant loci L1, L10 and L20 were unusual and by at most one place. Reversals between a duplex locus and a neighbouring simplex locus were particularly common.

Experimental data

Using the approach of Luo et al. (2000), the most likely SSR genotypes for the parents Stirling × 12601ab1 were aabc × bddo for STM3016 and accd × aabc for S140, where different letters denote different alleles, and *o* denotes a null allele. Here we consider only the 12601ab1 parent. Twenty-eight AFLP bands present in 12601ab1 but absent in Stirling were identified as being on linkage group IV by cluster analysis (Luo et al. 2001). Twenty-two of these segregated in an approximate 1:1 ratio and were assumed to be simplex markers (oooo × aooo), and six segregated in an approximate 5:1 ratio and were taken as duplex markers (oooo × aaoo). These are shown by a D before the marker name. Recombination frequencies and LOD scores were calculated between all pairs of the AFLP and SSR markers for the most likely phase, as described by Luo et al. (2001).

A preliminary order was calculated using the JMMAP module of JOINMAP, and this was used as an initial order for simulated annealing. This order had a least squares criterion of 3.485. The optimal configuration by simulated annealing is shown in Fig. 2, and had a least squares criterion of 3.198. The first five markers are the same as the JOINMAP order, and the last five markers have only one pair reversed compared to the JOINMAP order, but there are considerable rearrangements in the middle of the chromosome. There were a large number of orders close to the optimal one: among the top 100 orders, the least squares criterion increased from 3.198 to 3.204. None of the top 100 orders had any estimated distances that were significantly negative. The left side of Fig. 2 shows the number of orders in the top 100 for which that marker occupied that place; for example, markers PATMACG_295.0 and PCCMATA_179.0 were next to last and last, respectively, for 91/100 orders (and reversed for the other nine). This indicates two sections where the marker order is particularly difficult to establish, {DPGAMCAG_155.0, PATMAGG_92.3, DPACMATA_125.0} at 34–35 cM and EACAMCAC_153.0, S140 at 44 cM, where markers are particularly close. The weighted mean order over the top 100 is the same as the optimal order.

Discussion

This study has shown that simulated annealing can be used to order markers in a tetraploid population. It also enables the top orders to be compared to identify whether there are several orders with very similar values of the least squares criterion and, if so, to see which markers are in the same position in the top orders. We have chosen to examine the top 100 orders here, and to take a weighted mean order based on these, but the choice of 100 is arbitrary and could be varied.

It would be possible to use an alternative criterion for marker ordering in place of the weighted least squares criterion *S*. In diploids, maps are often based on the order with the maximum likelihood, but this is extremely complicated in tetraploid species. Xie and Xu (2000) attempted to write down a likelihood for more than two markers in an autotetraploid cross, based on a hidden Markov model, but their formulation does not model the process of bivalent formation correctly. Their likelihood could be corrected by the use of a multipoint likelihood conditional on each possible bivalent pairing, and then a summation over each bivalent pairing (Hackett 2001), but the resulting model would be very complex. *S* is more complicated to evaluate than criteria based only on adjacent pairs of markers such as the sum of adjacent LODS or adjacent recombination frequencies, but it is advantageous to use information from all pairs of markers in an autotetraploid cross, where pairs of markers vary considerably in their information content (Luo et al. 2001). *S* has the further advantage that it has been shown to be less sensitive to allele typing errors than multipoint likelihoods (Shields et al. 1991).

The best orderings are obtained with a population of 400 individuals, and with this size of population markers separated by a recombination frequency of 0.1 or 0.05 were ordered well. However, the power to detect linkage decreases and the standard error of the estimated recombination frequency increases as the marker separation increases or the population decreases (Hackett et al. 1998). Markers that are very tightly linked are also hard to order reliably, as a very large population is needed for enough recombinations between them. The simulation study showed that a low level of missing values had very little effect on the estimation of the map, but that even a low level of genotyping errors made the ordering less accurate and could increase the total map length. The error rate of 2% used here is thought to be typical for such mapping populations (Waugh, personal communication) This lengthening of the map agrees with results already known in diploid analyses (Buetow 1991).

Codominant markers are particularly informative in the ordering and can be up to four times as informative as dominant markers (Luo et al. 2001). They link not only homologous chromosomes from the same parent but also provide a reliable method to link maps from the two parents. Double-simplex markers could also be used to link the parental maps, but the simulation study has shown that these are particularly difficult to order accurately due to their large standard errors in any configuration except coupling (Meyer et al. 1998). The low information content of double-simplex markers also causes difficulties in aligning and merging maps of each parent in full-sib families from diploid heterozygous parents (Grattapaglia and Sederoff 1994). Pairs of duplex markers may also cause some difficulties in mapping if they are not in a coupling phase, as repulsion and mixed phase can only be distinguished by reference to other linked markers (Hackett et al. 1998). It may be necessary to check all such pairs and rerun the ordering with the recombination frequency and LOD score for the correct phase, but the effect on the map is unlikely to be substantial unless there is a high proportion of such pairs.

A reliable linkage map is a prerequisite for QTL mapping. Hackett et al. (2001) have derived a theoretical method for interval mapping of QTLs in autotetraploid species. One step of this is a reconstruction of the possible chromosome configurations for each offspring and their probabilities, based on the minimum number of obligatory crossovers. This criterion could in principle be used to order the markers instead of the weighted least squares criterion used here, but it is too time-consuming to calculate for a large number of possible orders. However, its application to the optimal order from simulated annealing may reveal, for example, apparent double crossovers originating from an error in data entry. A companion paper using QTL interval mapping of potato cyst nematode resistance in the Stirling × 12601ab1 population is in preparation.

Acknowledgements This work was financially supported by grants from the UK Biotechnology and Biological Research Council, and by the Scottish Executive Environment and Rural Affairs Department. We thank K. McLean for technical assistance.

References

Al-Janabi SM, Honeycutt RJ, McClelland M, Sobral BWS (1993) A genetic linkage map of Saccharum spontaneum L. 'SES 208'. Genetics 134:1249–1260

Bingham ET, McCoy TJ (1988) Cytology and cytogenetics of alfalfa. In: Hanson AA (ed) Alfalfa and alfalfa improvement. Agron Monogr 29, ASA, CSSA and SSSA, Madison, Wis., pp 737–776

Bradshaw JE, Stewart HE, Wastie RL, Dale MFB, Phillips MS (1995) Use of seedling progeny tests for genetical studies as part of a potato (Solanum tuberosum subsp. tuberosum) breeding programme. Theor Appl Genet 90:899–905

Bradshaw JE, Hackett CA, Meyer RC, Milbourne D, McNicol JW, Phillips MS, Waugh R (1998) Identification of AFLP and SSR markers associated with quantitative resistance to Globodera pallida (Stone) in tetraploid potato (Solanum tuberosum subsp. tuberosum) with a view to marker-assisted selection. Theor Appl Genet 97:202–210

Brouwer DJ, Osborn TC (1999) A molecular marker linkage map of tetraploid alfalfa (Medicago sativa L.) Theor Appl Genet 99:1194–1200

Buetow KH (1991) Influence of aberrant observations on high-resolution linkage analysis outcomes. Am J Hum Genet 49:985–994

Buetow KH, Chakravarti A (1987) Multipoint gene mapping using seriation. I. General methods. Am J Hum Genet 41:180–188

Da Silva J, Honeycutt RJ, Burnquist W, Al-Janabi SM, Sorrells ME, Tanksley SD, Sobral BWS (1995) Saccharum spontaneum L. 'SES 208' genetic linkage map combining RFLP- and PCR-based markers. Mol Breed 1:165–179

Da Silva JAG, Sorrells ME, Burnquist W, Tanksley SD (1993) RFLP linkage map and genome analysis of Saccharum spontaneum. Genome 36:782–791

Diwan N, Bouton JH, Kochert G, Cregan PB (2000) Mapping of simple sequence repeat (SSR) DNA markers in diploid and tetraploid alfalfa. Theor Appl Genet 101:165–172

Falk CT (1992) Preliminary ordering of multiple linked loci using pairwise linkage data. Genet Epidemol 9:367–375

Goffe WL, Ferrier GD, Rogers J (1994) Global optimization of statistical functions with simulated annealing. J Econometrics 60:65–99

Gebhardt C, Valkonen JPT (2001) Organization of genes controlling disease resistance in the potato genome. Annu Rev Phytopathol 39:79–102

Grattapaglia D, Sederoff R (1994) Genetic linkage maps of Eucalyptus grandis and Eucalypta urophylla using a pseudo-testcross: mapping strategy and RAPD marker. Genetics 137:1121–1137

Hackett CA (2001) A comment on Xie and Xu: 'Mapping quantitative trait loci in tetraploid species'. Genet Res 78:187–189

Hackett CA, Bradshaw JE, Meyer RC, McNicol JW, Milbourne D, Waugh R (1998) Linkage analysis in tetraploid potato: a simulation study. Genet Res 71:143–154

Hackett CA, Bradshaw JE, McNicol JW (2001) Interval mapping of quantitative trait loci in autotetraploid species. Genetics 159:1819–1832

Isidore E (2001) Construction and application of a multifunctional ultra-high-density genetic map in potato. PhD thesis, University of Dundee, UK

Jansen J, de Jong AG, van Ooijen JW (2001) Constructing dense genetic linkage maps. Theor Appl Genet 102:1113–1122

Jensen J, Jorgensen JH (1975) The barley chromosome 5 linkage map. Hereditas 80:5–16

Lalouel JM (1977) Linkage mapping from pair-wise recombination data. Hereditas 38:61–77

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363–2367

Leonards-Schippers C, Gieffers W, Schafer-Pregl R, Ritter E, Knapp SJ, Salamini F, Gebhardt C (1994) Quantitative resistance to Phytophthora infestans in potato: a case study for QTL mapping in an allogamous plant species. Genetics 137:67–77

Li X, van Eck HJ, van der Voort JNAM, Huigen DJ, Stam P, Jacobsen E (1998) Autotetraploids and genetic mapping using common AFLP markers: the R2 allele conferring resistance to Phytophthora infestans mapped on potato chromosome 4. Theor Appl Genet 96:1121–1128

Lin S (1965) Computer solutions of the travelling salesman problem. Bell System Technical J 44:2245–2269

Liu BH, Knapp SJ (1990) gmendel: A program for Mendelian segregation and linkage analysis of individual or multiple progeny populations using log-likelihood ratios. J Hered 81:407

Lu YY, Liu BH (1995) pgri, a software for plant genome research. Plant Genome III, San Diego, Calif

Luo ZW, Hackett CA, Bradshaw JE, McNicol JW, Milbourne D (2000) Predicting parental genotypes and gene segregation for tetrasomic inheritance. Theor Appl Genet 100:1067–1073

Luo ZW, Hackett CA, Bradshaw JE, McNicol JW, Milbourne D (2001) Construction of a genetic linkage map in tetraploid species using molecular markers. Genetics 157:1369–1385

Maliepaard C, Jansen J, Van Ooijen JW (1997) Linkage analysis in a full-sib family of an outbreeding species: overview and consequences for applications. Genet Res 70:237–250

Meyer RC, Milbourne D, Hackett CA, Bradshaw JE, McNicol JW, Waugh R (1998) Linkage analysis in tetraploid potato and association of markers with quantitative resistance to late blight (Phytophthora infestans). Mol Gen Genet 259:150–160

Milbourne D, Meyer RC, Bradshaw JE, Baird E, Bonar N, Provan J, Powell W, Waugh R (1997) Comparison of PCR-based marker systems for the analysis of genetic relationships in cultivated potato. Mol Breed 3:127–136

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318

Pande B (2002) The genetic analysis of traits of economic importance in the principal cultivated potato, Solanum tuberosum subsp. tuberosum. PhD thesis, University of Dundee, UK

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1986) Numerical recipes: the art of scientific computing. Cambridge University Press, Cambridge

Schiex T, Gaspin C (1997) cartagene: constructing and joining maximum likelihood genetic maps. In: Proc ISMB'97. Halkidiki, Greece

Shields DC, Collins A, Buetow KH, Morton NE (1991) Error filtration, interference and the human linkage map. Proc Natl Acad Sci USA 88:6501–6505

Stam P (1993) Construction of integrated genetic linkage maps by means of a new computer package: joinmap. Plant J 3:739–744

Stam P, Van Ooijen JW (1995) joinmap version 2.0: software for the calculation of genetic linkage maps. CPRO-DLO, Wageningen

Swaminathan MS, Howard HW (1953) The cytology and genetics of the potato (Solanum tuberosum) and related species. Bibliogr Genet 16:1–192

Thompson EA (1987) Crossover counts and likelihood in multipoint linkage analysis. IMA J Math Appl Med 4:93–108

Weeks DE, Lange K (1987) Preliminary ranking procedures for multilocus ordering. Genomics 1:236–242

Wu KK, Burnquist W, Sorrells ME, Tew TL, Moore PH, Tanksley SD (1992) The detection and estimation of linkage in polyploids using single-dose restriction fragments. Theor Appl Genet 83:294–300

Xie C, Xu S (2000) Mapping quantitative trait loci in tetraploid species. Genet Res 76:105–115