Theor Appl Genet (2005) 112: 30–40
DOI 10.1007/s00122-005-0097-x

ORIGINAL PAPER

**Hans Van Os · Piet Stam · Richard G. F. Visser
Herman J. Van Eck**

RECORD: a novel method for ordering loci on a genetic linkage map

Received: 12 August 2004 / Accepted: 20 August 2005 / Published online: 14 October 2005

Abstract A new method, REcombination Counting and ORDering (RECORD) is presented for the ordering of loci on genetic linkage maps. The method minimizes the total number of recombination events. The search algorithm is a heuristic procedure, combining elements of branch-and-bound with local reshuffling. Since the criterion we propose does not require intensive calculations, the algorithm rapidly produces an optimal ordering as well as a series of near-optimal ones. The latter provides insight into the local certainty of ordering along the map. A simulation study was performed to compare the performance of RECORD and JoinMap. RECORD is much faster and less sensitive to missing observations and scoring errors, since the optimisation criterion is less dependent on the position of the erroneous markers. In particular, RECORD performs better in regions of the map with high marker density. The implications of high marker densities on linkage map construction are discussed.

Introduction

Genetic linkage maps have become an indispensable tool for locating genes or quantitative trait loci (QTL), marker assisted breeding, and map-based gene cloning. The first linkage maps were based on few loci of morphological characteristics, like the classical *Drosophila* linkage map of chromosome X (Sturtevant 1913). Sturtevant introduced the concept that the frequency of crossing-over between two genes provides an index of their distance on a linear genetic map. He proposed that 1% of crossing-over should be taken as equal to one map unit. He devised a crucial test of the principles of mapping genes by constructing crosses in which at least two or three genes were segregating simultaneously. These two- or three-point crosses provided the principles and methods of ordering and mapping genes. These principles have enabled geneticists to map genes and markers to the chromosomes of a variety of higher organisms, including man. From this historical perspective it is obvious that mapping methods embark on pair-wise distance estimates. However, when large numbers of markers segregate in a single mapping population, the analysis of recombination events from marker segregation data is more rewarding. Distance estimates of marker pairs in dense regions are blurred by errors (Buetow 1991). The segregation data are a more direct reflection of the data ambiguities. Now, with the advent of molecular markers, much larger numbers of segregating loci can be mapped within one single mapping population. As an intermediate between conventional linkage maps and sequencing the complete genome of an organism, high-density maps are currently being generated (Steen et al. 1999, 4736 SSLP-markers; Kong et al. 2002, 5136 microsatellite marker; Harushima et al. 1998, 2275 EST markers; Isidore et al. 2003, 1260 AFLP markers). These maps sometimes comprise over 500 markers per linkage group. Since the number of possible orders asymptotically increases exponentially with the number of loci to be ordered, the problem of finding the optimal or near-optimal ordering requires a search algorithm that avoids an exhaustive search. For example, with 100 loci in a linkage group the number of orders equals $(100!)/2 = 9.3{\times}10^{157}$, which clearly prohibits an exhaustive search. Another factor that may set limits to the practical application of a search algorithm is the complexity of the target function to be minimized or maximized.

Communicated by S.J. Knapp

H. Van Os · P. Stam · R. G. F. Visser · H. J. Van Eck (✉)
Laboratory of Plant Breeding,
Wageningen University, 386, 6700 Wageningen,
AJ, The Netherlands
E-mail: herman.vaneck@wur.nl
Tel.: +31-317-482837
Fax: +31-317-483457

The optimisation problem

Locus ordering on a linkage map requires a criterion that defines the 'best' map and an algorithm to find the optimal sequence of loci. The criteria that have been proposed include the maximum likelihood (Lander et al. 1987; Jansen et al. 2001), the minimum sum of adjacent recombination fractions (SARF), the maximum sum of adjacent LOD scores (SALOD) (Liu and Knapp 1990), the minimum number of cross-overs (Thompson 1987), and the 'least square locus order' (Stam 1993).

Various computer packages for linkage mapping have implemented these criteria, combined with a certain search algorithm. For example, GMENDEL (Liu and Knapp 1990) minimizes SARF using simulated annealing. The PGRI package (Lu and Liu 1995) can minimize SARF, or maximize the likelihood, using simulated annealing and/or branch-and-bound. JoinMap (Stam 1993) minimizes the least square locus order using a stepwise search, which is a combination of seriation and branch-and-bound with some additional local reshuffling. For practical purposes, the target function should not require intensive calculations, and yet be acceptable from a statistical viewpoint. Especially with incomplete data (missing observations and/or incomplete genotype information, as is the case with dominance), calculation of the complete likelihood and the least square criterion is time consuming. As a result, the methods that use these criteria are becoming too computing-intensive for constructing linkage maps of over 400 loci, for instance, on a regular basis.

In this paper, we propose a target function using the minimal number of cross-over events as the optimisation criterion, and a search algorithm that enables ordering of data sets with more than 500 loci within a reasonable time.

Materials and methods

The optimisation criterion we use is COUNT, the number of recombination events. In a backcross (BC1) with perfect data (no missing observations), this number is easily obtained by counting the number of recombinants per locus pair, and, for a given sequence of loci, by adding over adjacent loci. Although COUNT and SARF are similar, there is an essential difference: COUNT cannot decrease as more gametes (individuals) are added to the population (cf. Thompson 1987). Since the likelihood, as well as COUNT and SARF are monotonic functions of the recombination frequencies between adjacent loci, COUNT, SARF and likelihood will give the same optimal ordering for perfect data (see also Jansen et al. 2001; Hackett and Broadfoot 2003). When information is incomplete due to, for example, missing observations or dominance in an F_2 mapping population, this counting of observable cross-overs is replaced by a value x, which is the expected number of crossovers for any incomplete observation of a pair of loci. This expected number (x) in turn is based on the maximum likelihood (ML) estimate of recombination frequency (r) between the corresponding loci. Table 1 illustrates this calculation for the observation of the genotype A_1A_1B-, being A_1A_1BB or A_1A_1Bb in an F_2, where the co-dominant allele A_1 and the dominant allele B are linked in coupling phase. For other genotypes of incomplete information, the calculation runs along the same lines, using the ML-estimate of recombination frequency to calculate the conditional probabilities of the hidden genotypes.

Table 1 Calculation of the expected number of recombination events (crossovers) resulting in the genotype A_1A_2B- in an F_2 derived from the cross $A_1A_1BB \times A_2A_2bb$

Observed genotype	Hidden genotypes	Conditional probability	Number of cross-overs
A_1A_1B-	A_1A_1BB	$\frac{(1-r)^2}{1-r^2}$	0
	A_1A_1Bb	$\frac{2r(1-r)}{1-r^2}$	1

The probabilities of the hidden genotypes (A_1A_1BB and A_1A_1Bb) are expressed in terms of the recombination frequency, r

$$E(x|A_1A_1B\text{-}) = 0 \times \frac{(1-r)^2}{1-r^2} + 1 \times \frac{2r(1-r)}{1-r^2} = \frac{2r}{1+r}, \quad (1)$$

where x is the expected number of cross-overs conditional to the observation of genotype A_1A_1B- and r is the recombination frequency.

In this way a matrix, X_{ij}, representing the number of recombination events between marker pairs, is constructed. Calculation of the criterion COUNT for a given sequence of loci is done by a simple addition of those numbers of recombination events over the proper (adjacent) loci, i.e.,

$$\mathrm{COUNT} = \sum_{i=1}^{n-1} X_{\mathrm{seq}(i),\mathrm{seq}(i+1)}, \quad (1)$$

where seq(i) is the ith element of the sequence.

The computational advantage of using COUNT is that for any exchange of two positions or an inversion of a window of certain size in a given sequence, the resulting value of COUNT requires the replacement of only a few terms of the summation in Eq. 1.

In order to prevent an unnecessary computational overload, the population is tested for the presence of 'duplicate markers', that is markers with exactly the same segregation pattern, including missing observations. Groups of markers with identical segregation signature are placed in 'bins', and each bin is represented by one of its members in the subsequent analysis. The order of loci within a bin remains unresolved unless additional information, not included in the 'current' mapping experiment, is available.

The core of the search algorithm is as follows. First, a sequence is constructed stepwise, starting with a randomly chosen pair of markers, and adding one marker at a time. For each marker to be added the best position is determined (one out of $n+1$ positions if the current sequence has n elements). This is a branch-and-bound-like procedure. The order in which markers are added to the sequence is random.

Once all markers have been added to the linkage group, thus making a 'sequence', an additional search for improvement is performed in the following way. A window of a given size is moved along the sequence from head to tail and for every position of this window, the subsequence within the window is inverted, and the resulting COUNT-value calculated. If the reverse order did not offer a lower COUNT-value, the inversion was restored. If a lower COUNT-value was obtained, subsequent steps were done given the new order. This is repeated for windows of increasing size, starting with size two, until the window covers all but one of the loci in the sequence. Every improvement encountered this way is accepted. The whole procedure is repeated until no further improvements are encountered. Notice that the strictness of the branch-and-bound method is lifted by the additional final search for local improvements, with the obvious goal of avoiding getting trapped in a local minimum. However, this reshuffling by a moving window of increasing size does not guarantee finding the global minimum. Indeed, experimentation with simulated data sets containing missing observations has shown that the final solution produced by this stepwise assembling and additional search slightly depends on the order in which markers are added to the sequence. A solution for this input order dependency would be to add markers by the seriation principle (Buetow and Chakravarti 1987), i.e. at each step, add the marker that is closest to the one at the current head or tail. In the context of the traveling salesman problem this strategy is also known as a 'greedy' one: at each step, travel to the nearest city that has not been visited before. It is known that this seriation strategy is not a guarantee to arrive at the global optimum either (Thompson 1987). For that reason, we chose to simply repeat the procedure a number of times and select the best one from these replicate assemblages. With good quality data, the replicate solutions produced by RECORD are all identical. Upon experimentation with simulated data, we found that for data sets with up to 20% missing observations, increasing the number of replicate assemblages beyond ten is hardly rewarding. So we consider ten replicate build-ups of the sequence as a good compromise between speed and quality of the solution obtained.

Since the producer of a linkage map is not only interested in a single 'best' sequence of markers, but also in the certainty of that sequence, we have added the following procedure to the algorithm. Starting from the last and optimal solution, a search is performed for 'almost equivalent' solutions. An 'almost equivalent' solution is defined as one that induces a pre-set additional number of crossovers. So, a search is done for solutions that fall within this range of 'admissible' values of COUNT. The search itself is the same as described above: inversion of the sequence within a moving window, which is repeated for windows of increasing size. From the set of admissible solutions obtained this way, for each locus, its distribution of positions is recorded. Inspection of this distribution provides a quick impression of the local certainty of the sequence. Figure 1 gives a sample of RECORD output, listing the positions taken by each marker in the set of 'admissible' sequences. It shows that for approximately 50% of the loci in this example the position is fixed, whereas for 'islands' of clustered markers, the order within such a cluster is indeterminate.

REcombination Counting and ORDering can deal with the following types of mapping populations: BC1, F_2, F_3, RILs (in fact any generation obtained by repeated selfing of a hybrid between homozygous parents). Mapping populations from non-inbreds should be split into BC1 or HAP data that represent the maternal and paternal gametes, according to the two-way pseudo-testcross method (Grattapaglia and Sederoff 1994).

The algorithm described above has been implemented in a DOS-oriented, C++ written computer program, which is available from our web site (http://www.dpw.wageningen-ur.nl/pv/). We have chosen the DOS platform since it enables running large batch jobs which is convenient for the purpose of the remainder of this study, a comparison of the performance of RECORD and JoinMap using simulated data.

A comparison of JoinMap and RECORD

In JoinMap the stepwise assembling of a locus sequence is essentially the same as in RECORD, i.e. a seriation-like procedure with local reshuffling (called 'rippling' in JoinMap) in a search for improvements (Stam 1993; Stam and Van Ooijen 1995). The search method of RECORD requires $(1/2)n(n-1)$ evaluations of the target function for a sequence of length n. In JoinMap a similar number is required. However, evaluation of the JoinMap target function involves the inversion of an $n \times n$ matrix for each sequence of size $n+1$. So, asymptotically the number of operations in RECORD increases as n^2, whereas in JoinMap this increase is approximately by n^4. Moreover, calculation of COUNT, going from a given sequence to one with an inverted segment, requires the replacement of only a few terms in the summation of Eq. 1. This makes the RECORD algorithm extremely fast.

Three different experiments were performed. The first experiment was done to test whether or not the new method of minimizing recombination events implemented in RECORD can produce maps of the same quality as the approach based on pair-wise marker distances implemented in JoinMap. Both RECORD and JoinMap were tested under a number of varying conditions such as population size, missing observations, and error rate. In the second experiment, the two programs were tested for their error-sensitivity under different marker densities. In the third experiment, the speed of the software was evaluated.

Simulated data

We simulated first generation BC1 populations. The simulated data were produced as follows. A given number of loci were randomly positioned (according to a Poisson process) along a single chromosome of specified length in cM. Centimorgen values are given as if calculated from an infinite amount of genotypes. Genotypes were generated for a BC1 progeny following standard Mendelian segregation. Furthermore, we assumed no crossover interference. The number of crossover events solely depends on the distance as specified by the simulated positions of the loci on the map. Scoring results were generated by assuming that missing observations and errors were independently and randomly distributed. (Note: Throughout this paper, we imply that genotyping errors comprise human errors in the lab, scoring errors, typing errors, as well as reproducible though conflicting data points, resulting from biological phenomena such as gene conversion.)

In Experiment I, 150 independent maps of 50 loci spread along 50 cM were simulated. Next to speed, error-sensitivity is one of the most important factors while coping with high-density data sets. In this study, emphasis is placed on both error-sensitivity and speed. From each map, four populations were simulated consisting of 25, 50, 100, and 250 individuals. In all population, data noise was introduced by either 5, 10, 15, 20, and 30% errors or missing observations.

Experiment II was based on two data sets of different marker density. One data set was simulated from a map with 100 loci on a 10 cM map, and the other from 100 loci on a 100 cM map. Both data sets consisted of 100 individuals and 3% scoring errors.

Experiment III was set up to assess the calculation speed of the two algorithms. Data sets were varied in the number of loci (50, 100, 150, and 200 loci) and population size (25, 50, 100, and 250). All data sets contained 5% scoring errors, because perfect data do not provide a realistic impression of the mapping time in practice. The

```
                  000000000011111111112222222222
                  012345678901234567890123456789
 0   g3715      | 00
 1   w121       | 0000
 2   m217       | 0000
 3   g3837      |   00
 4   w174       |     00
 5   CHS        |     00
 6   w322       |       0
 7   g4560      |        0
 8   w138       |         000
 9   w433       |         000
10   m291       |         000
11   g4715-b    |            0
12   w219       |             0
13   w125       |              0
14   w291b      |               0
15   w137       |                0
16   w323       |                 0
17   m247       |                  0
18   g4028      |                   0
19   w194       |                    0
20   w423b      |                     0 0
21   w61        |                     0 0
22   w271       |                       0 0
23   w2         |                       0 0
24   m435       |                          0
25   w184       |                          0
26   w69        |                        000
27   g2368      |                             0
28   m555       |                             0 0
29   w335       |                             0 0
```

Fig. 1 Sample output of RECORD showing the rank numbers taken by markers in a series of near-optimal solutions. The *vertical numbers* (1–29) represents the expected rank number of the loci (indicated by their marker name g3715. w335), and the *horizontal numbers* (1–29) are the observed rank numbers of the loci as obtained with RECORD. The *diagonal of "0" signs* indicate the correlation between expected and observed rank numbers. *Multiple "0" signs*, as shown at positions 8–10, indicate that alternative ordering of the marker loci w138, w433, and m291 have equal or near equal likelihood. Data taken from the *Arabidopsis* genome database

Table 2 Values of simulation variables used in the three different experiments

Variables	Experiment I	Experiment II	Experiment III
Map length (cM)	50	10, 100	50
Number of loci	50	100	50, 100, 150, 200
Population size	25, 50, 100, 250	100	25, 50, 100, 250
Percentage scoring errors	0, 5, 10, 15, 20, 30	3	5
Percentage missing observations	0, 5, 10, 15, 20, 30	0	0

different settings for the simulations in the three experiments are summarized in Table 2.

A yardstick for performance

As a measure for the performance of both algorithms, we examined two different correlation coefficients between marker positions of the calculated sequence, and the true order in the map that was used to generate the data. Since we are not dealing with map positions in centimorgans, but rather with rank numbers, the first correlation coefficient is Spearman's rank correlation (r_s). The second correlation coefficient is Kendall's τ coefficient.

In order to see to what extent local rearrangements of a given sequence of rank numbers affects the correlation coefficients, we derived the following equations for local inversion of a segment. Inverting a window of size k in a sequence of length n leads to

$$r_s = 1 - 2\frac{k(k^2-1)}{n(n^2-1)} \quad \text{and} \quad \tau = 1 - 2\frac{k(k-1)}{n(n-1)}.$$

Taking k as a fraction of n and writing $k/n=p$, one obtains, as n tends to infinity:

$$r_s(p) = 1 - 2p^3 \quad \text{and} \quad \tau(p) = 1 - 2p^2. \qquad (2)$$

Figure 2 presents a graph of these relations. It shows that upon inverting 50% ($P=0.5$) of a long sequence, r_s is still 0.75, whereas τ is 0.50. Clearly, Kendall's τ is a more sensitive correlation coefficient than Spearman's r_s. Small inversions, of less than 5% of the total length, have a negligible effect on the correlation coefficients. Multiple inversions will, of course, have larger impact. For m non-overlapping inversions covering a proportion p_i of the sequence, r_s and τ become

$$r_s = 1 - \sum_{i=1}^{m} 2p_i^3, \quad \tau = 1 - \sum_{i=1}^{m} 2p_i^2 \left(\sum p_i \leq 1\right).$$

We conclude that for r_s to drop below 0.8, or for τ to drop below 0.6, for instance, a very serious distortion of the sequence is required. In fact, such a distortion would be unacceptable in a real mapping experiment. To correct possible (almost) complete map inversions, the absolute value of r_s and τ was taken for further calculations.

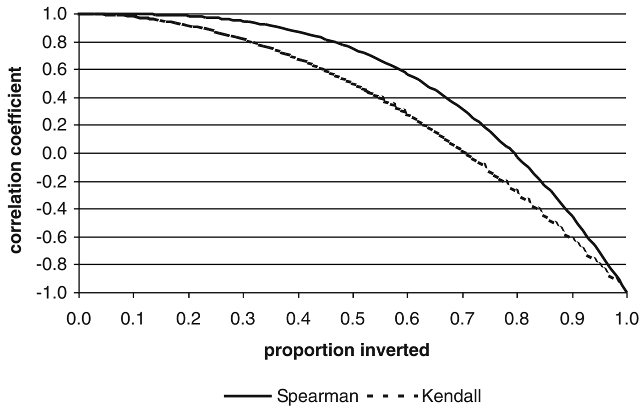

Fig. 2 Change of two different correlation coefficients, Spearman's r_s and Kendall's τ, by inverting a window of markers consisting of a proportion p of a long sequence (Eq. 2)

For testing purposes, rather general program settings were chosen for JoinMap. This means that all pair-wise data were used with a LOD score higher than 1.0 and an estimated recombination fraction smaller than 0.45. Before actual mapping starts, JoinMap calculates the likelihood of the three possible orders of every triplet. When one of these exceeds the other two by a user-defined threshold value, this order is inferred as a so-called 'fixed order'. (In the subsequent step-wise build-up and search of JoinMap, every order that is in conflict with a 'fixed order' is taboo.) In these experiments, the triplet threshold (logarithm of likelihood ratio) was set to 7.0. Finally, both JoinMap and RECORD have the option to perform a 'ripple' after adding a marker to the map. With a ripple, local marker order changes are systematically considered while improvements are maintained. In these tests, neither program performs ripples.

During this study, JoinMap 3.0 (Van Ooijen and Voorrips 2001) became available. This version of JoinMap is user-friendly because of the graphical user interface. However, for our experiments the MSDOS oriented JoinMap 2.0 was chosen because of its ability to run batch jobs. The results from this study can be extrapolated to JoinMap 3.0, since only minor changes in the algorithm have been introduced (J.W. Van Ooijen, personal communication).

Results

Experiment I

In this experiment, both JoinMap and RECORD were tested with simulated data representing 50 marker loci on a 50 cM linkage group. Irrespective of the size of the

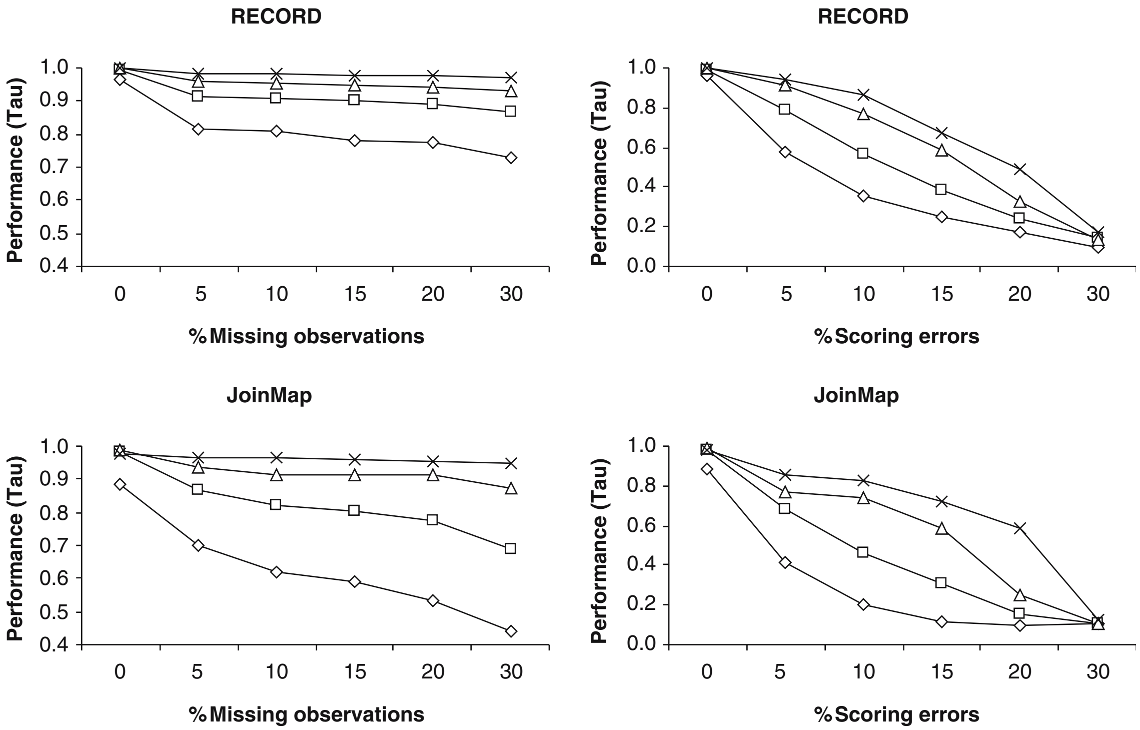

Fig. 3 Performance in Kendall's τ of RECORD and JoinMap on data sets differing in population size and noise level. The population size is indicated by: '◇' for 25, '□' for 50, 'Δ' for 100, and '×' for 250 individuals. The results are based on 150 replicate runs

mapping population ($N = 25$, 50, 100, and 250), perfect marker orders were obtained. This result demonstrates that map construction using perfect data is not really a test case. In addition, we tested two more algorithms, i.e. ComBin (Buntjer et al. 2000a, b) and JMQAD (the 'Quick-And-Dirty' module within the JoinMap 2.0 package) to recognize again that perfect maps are surely obtained with perfect data (results not shown). Apparently, the real test case for the performance of mapping algorithms is their sensitivity for ambiguities in the data caused by missing observations and/or genotyping errors. In realistic data, the proportion of missing observations and genotyping errors generally does not exceed 5%. However, to get a better view on the sensitivity of the methods for noise, both programs were tested with elevated levels of missing observations (5–30%) and scoring errors (5–30%). The performance of each of the programs, defined as the correlation coefficient between the true marker order and the order inferred by the software, was averaged over the 150 replications for every situation, and is shown in Fig. 3. It is clear that the accuracy of the marker order produced by the programs decreases with the data quality, reflecting a decrease in the ability of both programs to recover the correct order when data quality gets poor.

Missing observations do not severely harm the recovered marker order. Especially, in large mapping populations, the number of observations across descendants largely compensates the ambiguities caused by missing observations. Moreover, the vast majority of the missing observations do not induce ambiguities. Only when missing observations occur near recombinations, the placement of the markers with RECORD will be less accurate. Under these circumstances, missing observations complicate the separation of markers from neighboring loci, and make a pair of co-segregating loci of unspecified order. When more missing observations are present, the chance increases that these occur near recombinations. JoinMap, however, is more sensitive to missing observations than RECORD. Since in JoinMap not only recombination estimates between adjacent markers, but all pair-wise recombination estimates beyond a certain LOD threshold are used in the target function, and since a single missing observation slightly affects many of these pair-wise estimates, the impact of an increasing proportion of missing observations in JoinMap is greater than in RECORD.

The consequences of scoring errors are much more serious. An error may cause a separation of two co-segregating markers into two different loci. In this respect scoring errors have the same effect as recombination. While recombinations are generally confirmed by other data points, errors occur on their own, and seldom confirm each other.

In Fig. 4, an example data set is shown containing two forms of genotyping errors. Marker4 contains an error in individual 5. Individual 5 does not contain any recombination events. Therefore this particular error will not add to the cost function of RECORD, when marker4 is tested on different positions. Placing

	1	2	3	4	5	6	7	8	9
MARKER1	A	A	B	B	A	A	B	B	A
MARKER2	A	A	B	B	A	A	B	B	A
MARKER3	B	A	A	B	A	B	B	B	A
MARKER4	A	B	A	B	B	B	A	B	A
MARKER5	B	B	A	A	A	B	A	B	B
MARKER6	B	B	A	A	A	B	A	B	B

Fig. 4 Inspection of raw data (example) can expose two types of errors. Segregating alleles are indicated by 'A' and 'B'; individuals are represented in columns 1–9; marker data are shown in rows 1–6. The erroneous data point in marker 4 at individual 5 is a clear singleton that does not cause an ordering ambiguity in RECORD. The order of markers 3 and 4 is based on individuals 2 and 7, but is doubted by individual 1. Individual 1 contains an error close to a recombination event. In this case it is not clear whether markers 3 or 4 contains the error in individual 1

marker4 at the end of the linkage group will not improve the order as it causes a higher increase of the cost function in the other individuals. While RECORD is not sensitive to this kind of errors, JoinMap and other methods based on pair-wise distances consider this error as a recombination and include it in the map distance calculation.

A different situation occurs in individual 1, where the error is close to a recombination event. Initially, RECORD will invert markers 3 and 4. This change will decrease the cost function in individual 1. However, this will cause a higher increase in the total cost function due to individuals 2 and 7. This situation remains insolvable as it is not clear whether marker 3 or 4 contains the error. The best order is determined based on the other individuals in the data set. In conclusion, scoring errors provide RECORD with ordering ambiguities only when they occur near recombination events. On the other hand, pair-wise distance estimates are always affected by errors, independent of their position.

In general, larger populations have a beneficial effect on the mapping result. As population size increases, more recombination events between a pair of markers can be observed, which adds to the resolution between the markers. The positioning of the markers will be more accurate, and the relative impact of missing observations and scoring errors decreases.

T-tests (data not shown) demonstrate that RECORD produces equally good or significantly better results than JoinMap. The *T*-tests were more significant when using Kendall's τ rather than Spearman's r_s. By exception, on data sets containing 250 individuals with an exceptional high error rate of 15 or 20%, JoinMap has a small advantage over RECORD, although neither algorithm produces accurate maps in this situation. The reason for this small advantage for JoinMap is that at larger population sizes, errors have a smaller impact on the distance estimates.

Experiment II

In the second experiment, JoinMap and RECORD were tested for their ability to determine the marker order at higher densities. For this purpose, two data sets were used. The first set was based on a 100-marker map of 100 cM length ('normal' density). The second one was generated from a 'saturated' map where 100 markers were spread over a distance of only 10 cM. From both the maps, a BC1 population was simulated and a realistic amount of 3% errors was introduced. Calculated orders from both programs were compared with the true one and the results are shown by the scatter plot of Fig. 5.

The dense map was more challenging to both programs. Although the mean number of errors remains the same, the average number of true crossovers in the dense map is reduced by a factor 10 as compared to the sparse map. This causes the signal/noise ratio to decrease by a factor 10. This explains why mapping in dense regions is more error-sensitive than mapping in less dense regions. The results of experiment II show that in more dense regions RECORD performs better than JoinMap.

Experiment III

In the third experiment, RECORD and JoinMap were compared for their speed. Calculation time was measured for a number of data sets varying in the number of loci and population size on a computer with a Pentium II MMX processor of 350 MHz. Population size does not have a big effect on JoinMap. Therefore, the results were averaged over tests at different offspring sizes with the same number of loci. Figure 6 shows the increase in calculation time for both programs. We fitted power curves to these data, and as anticipated, computation times for RECORD and JoinMap nicely fit curves of powers 2 and 4, respectively. Thus, especially with data sets of over 100 loci, the speed advantage of RECORD over JoinMap is overwhelming.

Discussion

There are two major aspects to methods for efficient ordering of gene loci on a linkage map. First, the target function is important. In this paper, we propose the total number of observable recombination events between adjacent markers as the target function, with an adaptation for situations in which genotype information is incomplete or missing. From a statistical point of view the full likelihood function would be an attractive alternative. The two criteria are equivalent in case the data are perfect (no missing observations and complete genotype information). In order to investigate the behavior of COUNT and likelihood with realistic data sets, we compared the two methods using simulated data sets with incomplete information, i.e. an F_2 of size 100

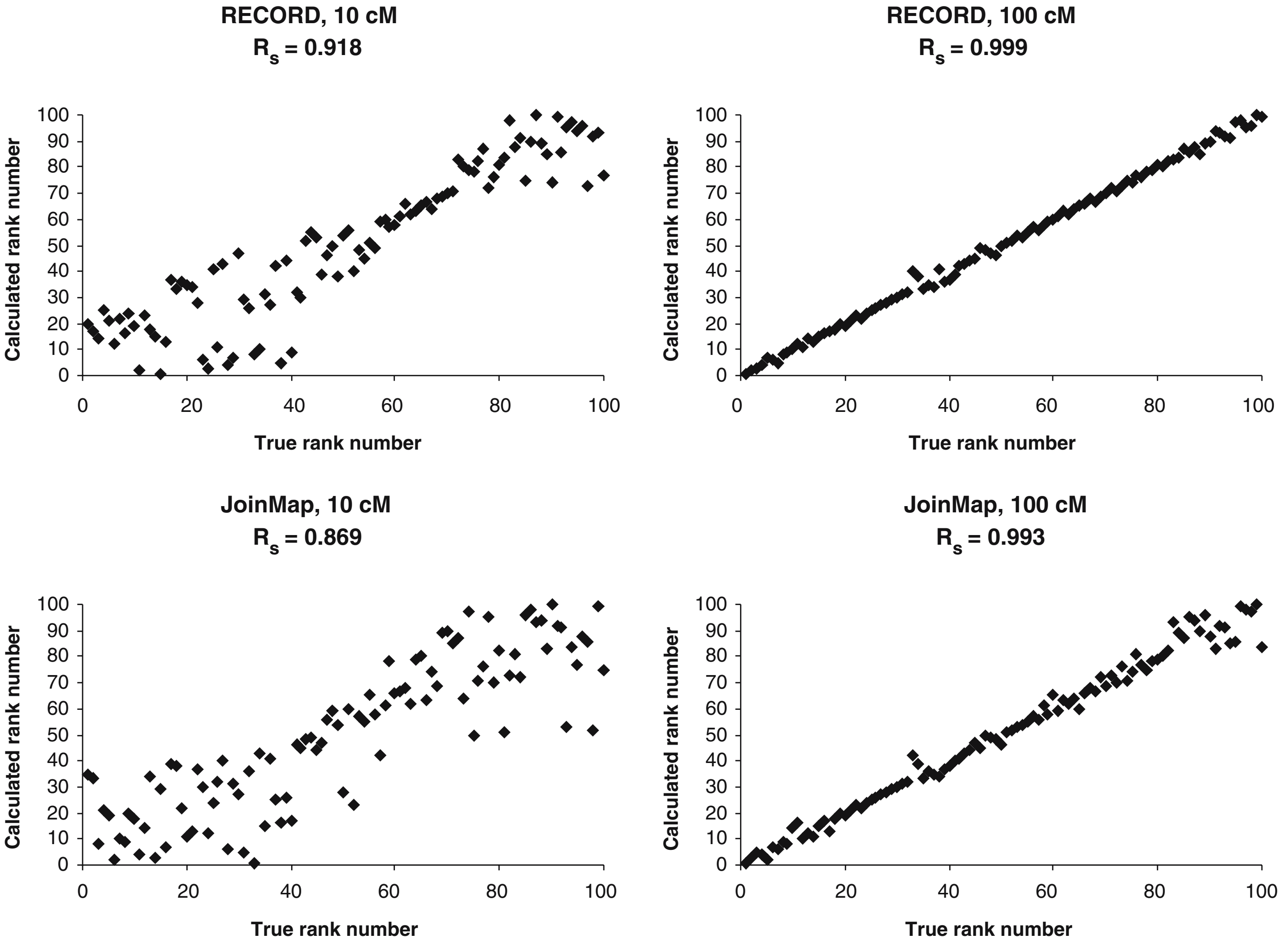

Fig. 5 Performance of RECORD and JoinMap in dense maps. The calculated rank number of markers by both RECORD and JoinMap is compared with the true rank number by Spearman's r_s. Data were obtained from experiment II

with dominant markers and 5% missing observations. The two target functions were calculated for a series of near-optimal sequences (obtained by local inversion of segments) as well as a series of random rearrangements in the correct sequence.

Specifically for the first set of sequences (which corresponds to the part of the parameter space searched by RECORD), the squared correlation between COUNT and likelihood never dropped below 0.90. An example of the results of these calculations, where the correlation is one of the poorest we encountered, is shown in the scatter diagram of Fig. 7. So, for practical purposes, our heuristic COUNT criterion appears to be quite an acceptable compromise between statistical rigour and common sense.

Several other easy-to-calculate target functions have been proposed in the past. Among these are sum of adjacent map distances (SAD), sum of adjacent recombination frequencies (SARF), and SALOD scores. For perfect data all of these are equivalent, in the sense that they have the same global optimum. However, with incomplete data both SARF and SALOD are inferior to COUNT. This is because SARF does not account for variation in the precision of pair-wise estimates, whereas SALOD may lead to erroneous results when the number of informative individuals varies between pairs of loci. Contrarily, the COUNT function comes close to the full likelihood since it uses observable recombination events (which are equivalent to likelihood), for that part of the data which has complete information, and uses maximum likelihood estimates for the data that are incomplete.

The second aspect of map construction concerns the search algorithm for the optimum. In the analogy of the travelling salesman problem, several approaches have been proposed. Among these are branch-and-bound (Thompson 1987), seriation (Buetow and Chakravarti 1987), and simulated annealing (SA; Kirkpatrick et al. 1983), or combinations thereof. Although SA generally produces optimal or near-optimal solutions, we did not choose it for the following reason. Extensive experience with linkage mapping has shown that most alternative maps that are produced by different computer packages and/or different program settings in JoinMap differ by

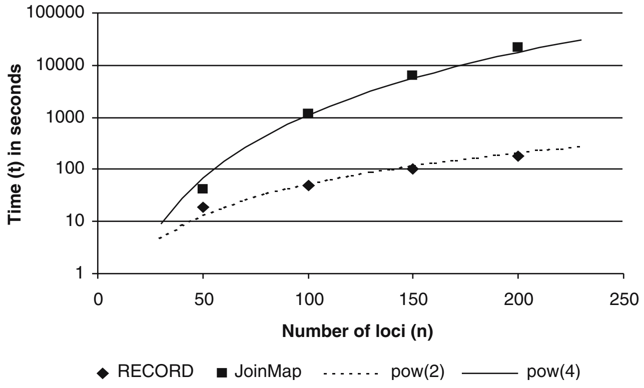

Fig. 6 Computation time for RECORD and JoinMap. Fitted power curves obtained by regression of time on number of loci. RECORD: $t = 0.00534n^2$; JoinMap: $t = 0.000011n^4$. Data were obtained from experiment III

inverted segments in the locus sequence. This is the result of ambiguities in real data and is in line with what one would expect intuitively. So, rather than the SA-search, which starts from a random sequence and subsequently randomly exchanges two loci, or randomly moves a single locus along the sequence, we decided to search that part of the parameter space which most likely represents biological reality, starting from an 'educated first guess' obtained by the branch-and-bound method.

One may, of course, think of heuristic variations to both SA and the RECORD search. For example, to first construct a 'skeleton map' of not-too-closely linked markers and, during the subsequent SA-search involving all loci, consider any exchange of position involving two skeleton markers as a taboo area of the parameter space (J. Jansen, personal communication).

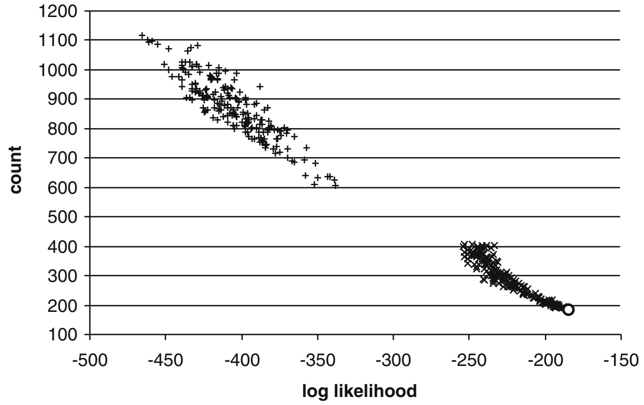

Fig. 7 Relation between COUNT and log-likelihood. Data source: a simulated F_2 population of size 100 with dominant markers and 5% missing observations. Inversions (x): result for 80 sub-optimal sequences obtained by inversion of sequence segments. Random (+): result for sequences obtained by random exchange of pairs of loci. Best (O): solution produced by RECORD. Notice the much smaller likelihoods for sequences obtained by random changes, as explored by simulated annealing, in comparison with the likelihoods obtained by local inversion of segments

An additional aspect of linkage mapping, which until recently has received little attention, concerns the (un)certainty of the map produced by a particular algorithm. We have added a feature to RECORD which provides the user with the distribution of rank numbers in a series of near-optimal solutions. Recently Jansen et al. (2001) and Hackett et al. (2003) have described a similar approach by recording the positions of loci in a series of sub-optimal solutions encountered in the SA-search.

In our comparison of the performance of RECORD and JoinMap we did not account for the fact that RECORD only produces orders, whereas JoinMap produces map positions in centimorgans. Therefore, the comparison is not a completely 'fair' one. On the other hand, correct locus ordering is of more importance than having 'exact' map distances, especially when constructing high-density maps. In such high-density maps, the resolution that can be attained is primarily dictated by the size of the mapping population, usually not surpassing 1.0–0.25 cM. Estimated 'exact' map distances in this order of magnitude do not make much sense, as their standard error readily exceeds the estimate itself.

Subsequent reasons as to why we have put emphasis on correct locus ordering and consider distance as relatively insignificant, are based on the unequal distribution of both recombination events and AFLP markers on the physical map. Highly localized hotspots or coldspots for recombination may cause manifold differences in map distance estimates between loci, depending on the sex or genetic background of the parental genotype. As a result, physical to genetic distances can vary from 25 kb/cM (Büschges et al. 1997) to 40 Mb/cM (Zhong et al. 1999). Futhermore, successful application of mapping information in map-based cloning or marker assisted selection with flanking markers also depends more on a correct marker order than accurate genetic distance estimates.

Apart from the observed difference in error-sensitivity between the programs, the results of experiment II once more demonstrate the disastrous effect typing errors will have on the ability to recover the correct locus order, especially for regions of high marker density. This confirms earlier notions by Buetow (1991) and Hackett and Broadfoot (2003) on the graveness of scoring errors on map construction. Figure 3 indicates that the penalty for a typing error is roughly fivefold the penalty for a missing observation; a similar conclusion was draw by Hackett and Broadfoot (2003). For this reason we have developed a procedure, 'SMOOTH', for the detection of 'suspect' data points in a mapping population (Hans Van Os et al., manuscript in preparation). We have successfully applied this procedure in constructing a high-density linkage map for chromosome I in diploid potato (Isidore et al. 2003).

At this moment the RECORD-approach is being used for ultra-dense map construction in potato (Isidore et al. 2003). In these situations, linkage groups may contain more than 500 markers, numbers unthinkable in being analyzed simultaneously by conventional mapping software, as it would take more than 9 days to calculate the map. Contrarily, RECORD analyses data sets of 500 markers within 20 min.

When RECORD was being developed, there were no alternative programs available that could handle these amounts of data. A new algorithm that can speed up map calculation based on pair-wise distances by using the simulated annealing approach has been tested, but is not yet available (Jansen et al. 2001; J.W. Van Ooijen, personal communication).

RECORD is capable of handling data sets of backcross populations, but to apply RECORD for the construction of the high-density map of potato, which is based on a population derived from non-inbred parents, several modifications have to be made to the raw data. First, the observations recorded in the offspring have to be split into the products of male and female meiosis. From there on, the maps from both parents have to be calculated separately. Within the parental data sets, the linkage phase of each marker has to be assessed. This can be done with the 'Quick-And-Dirty' mapping module, which is included in the JoinMap 2.0 software package. This program calculates the best marker order by minimizing the sum of adjacent distances. Although this module does not produce very accurate marker orders, it is accurate enough for linkage phase ascertainment, which can be done hand-based, on the neighboring markers. By converting all markers that are in repulsion phase into coupling phase, the data are comparable with two separate BC1 populations for each parent, also referred to as the two-way pseudo-testcross (Grattapaglia and Sederoff 1994).

The version of RECORD used in this study only produces orders of loci but no map positions in centimorgans. Currently we are preparing a version which does have this feature as well as several sophistications, like a choice of target functions, an extended search algorithm for the more ambiguous data sets, a graphical user interface, and a variety of output options.

In summary, conventional software has been sufficient in calculating linkage maps of low density. For the construction of high-density maps, there is a strong need for faster and error-tolerant methods. The method described in this paper exceeds the currently available software both in speed and accuracy.

Acknowledgments We are very grateful to Dr. Hans Jansen and Dr. Johan van Ooijen for providing valuable comments. This work was carried out under the European Union framework 5 program grant: FAIR5-PL97-3565.

References

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

Buntjer JB, Van Os H, Van Eck HJ (2000a) ComBin: software for ultra-dense mapping plant and animal genome conference VIII, San Diego, CA; http://www.intl-pag.org/pag/8/abstracts/pag8038.html

Buntjer JB, Van Os H, Van Eck HJ (2000b) Construction of ultra-dense maps using novel software plant and animal genome conference VIII, San Diego, CA; http://www.intl-pag.org/pag/8/abstracts/pag8039.html

Büschges R, Hollricher K, Panstruga R, Simons G, Wolter M, Frijters A, van Daelen R, van der Lee T, Diergaarde P, Groenendijk J, Topsch S, Vos P, Salamini F, Schulze-Lefert P (1997) The barley *Mlo* gene: a novel control element of plant pathogen resistance. Cell 88:695–705

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

Hackett CA, Broadfoot LB (2003) Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity 90:33–38

Hackett CA, Pande B, Bryan GJ (2003) Constructing linkage maps in autotetraploid species using simulated annealing. Theor Appl Genet 106:1107–115

Harushima Y, Yano M, Shomura A, Sato M, Shimano T, Kuboki Y, Yamamoto T, Lin SY, Antonio BA, Parco A, Kajiya H, Huang N, Yamamoto K, Nagamura Y, Kurata N, Khush GS, Sasaki T (1998) A high-density rice genetic linkage map with 2275 markers using a single F_2 population. Genetics 148:479–494

Isidore E, Van Os H, Andrzejewski S, Bakker J, Barrena I, Bryan G, Buntjer J, Caromel B, Van Eck HJ, Ghareeb B, Jong W de, Koert P van, Lefebvre V, Milbourne D, Ritter E, Rouppe van der Voort J, Rousselle-Bourgeois F, Vliet J van, Waugh R (2003) Towards a marker-dense meiotic map of the potato genome: lessons from linkage group I. Genetics 165:2107–2116

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

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174–181

Liu BH, Knapp SJ (1990) GMENDEL: a program for Mendelian segregation and linkage analysis of individual or multiple progeny populations using log-likelihood ratio's. J Hered 81:407

Lu YY, Liu BH (1995) A new computer package for genomic research: PGRI (Plant Genome Research Initiative), plant genome conference III, San Diego, CA; http://www.intl-pag.org/3/abstracts/201pg3.html

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

Steen RG, Kwitek-Black AE, Glenn C, Gullings-Handley J, Van Etten W, Atkinson OS, Appel D, Twigger S, Muir M, Mull T, Granados M, Kissebah M, Russo K, Crane R, Popp M, Peden M, Matise T, Brown DM, Lu J, Kingsmore S, Tonellato PJ, Rozen S, Slonim D, Young P, Knoblauch M, Provoost A, Ganten D, Colman SD, Rothberg J, Lander ES, Jacob HJ (1999) A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. Genome Res 9:AP1–AP8

Sturtevant AH (1913) The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of a association. J Exp Zool 14:43–59

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

Van Ooijen JW, Voorrips RE (2001) JoinMap® Version 3.0, software for the calculation of genetic linkage maps. Plant research international, Wageningen

Zhong XB, Bodeau J, Fransz PF, Williamson VM, Van Kammen A, De Jong HJ, Zabel P (1999) FISH to meiotic pachytene chromosomes of tomato locates the root-knot nematode resistance gene Mi-1 and the acid phosphatase gene Aps-1 near the junction of euchromatin and pericentromeric heterochromatin of chromosome arms 6S and 6L, respectively. Theor Appl Genet 98:365–370